Dr. Joseph Teguh Santoso, S.Kom, M.Kom.

REVOLUSI PENDIDIKAN di zaman AI



YAYASAN PRIMA AGUS TEKNIK

Dr. Joseph Teguh Santoso, S.Kom, M.Kom.

REVOLUSI PENDIDIKAN di zaman AI



YAYASAN PRIMA AGUS TEKNIK

PENERBIT :
YAYASAN PRIMA AGUS TEKNIK
Jl. Majapahit No. 605 Semarang
Telp. (024) 6723456. Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

Revolusi Pendidikan di zaman AI

Penulis :
Dr. Joseph Teguh Santoso, S.Kom., M.Kom.

ISBN :

Editor :
Dr. Agus Wibowo, M.Kom, M.Si, MM.

Penyunting :
Dr. Mars Caroline Wibowo. S.T., M.Mm.Tech

Desain Sampul dan Tata Letak :
Irdha Yunianto, S.Ds., M.Kom

Penebit :
Yayasan Prima Agus Teknik Bekerja sama dengan
Universitas Sains & Teknologi Komputer (Universitas STEKOM)

Anggota IKAPI No: 279 / ALB / JTE / 2023

Redaksi :
Jl. Majapahit no 605 Semarang
Telp. 08122925000
Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

Distributor Tunggal :
Universitas STEKOM
Jl. Majapahit no 605 Semarang
Telp. 08122925000
Fax. 024-6710144
Email : info@stekom.ac.id

Hak cipta dilindungi undang-undang
Dilarang memperbanyak karya tulis ini dalam bentuk dan dengan cara apapun tanpa ijin dari penulis

KATA PENGANTAR

Puji syukur ke hadirat Tuhan Yang Maha Esa atas limpahan rahmat dan karunia-Nya sehingga buku yang berjudul "*Revolusi Pendidikan di Zaman AI"* ini dapat diselesaikan. Buku ini disusun sebagai upaya untuk memberikan pemahaman yang komprehensif mengenai perubahan pendidikan di tengah perkembangan kecerdasan buatan (*artificial intelligence*/AI), khususnya dalam bidang penilaian dan asesmen pendidikan. Perkembangan AI telah membawa perubahan besar dalam berbagai aspek kehidupan, termasuk proses pembelajaran, pengembangan bahan ajar, penyusunan soal, pemberian umpan balik, pengukuran kompetensi, serta evaluasi capaian peserta didik. Teknologi seperti *machine learning*, pemrosesan bahasa alami (*natural language processing*/NLP), model bahasa besar (*large language models*/LLM), AI generatif, analitik pembelajaran, dan sistem penilaian adaptif membuka peluang untuk menghadirkan proses pendidikan yang lebih personal, efisien, responsif, dan berbasis data.

Namun demikian, pemanfaatan AI dalam pendidikan tidak semata-mata berkaitan dengan kecanggihan teknologi. Penggunaan AI juga menimbulkan pertanyaan mendasar mengenai keadilan, transparansi, privasi, keamanan data, bias algoritmik, validitas penilaian, pengawasan manusia, serta tanggung jawab moral dan hukum. Oleh karena itu, transformasi pendidikan berbasis AI perlu dilakukan secara kritis, etis, inklusif, dan tetap menempatkan manusia sebagai pusat proses pendidikan. Buku ini membahas berbagai aspek penting dalam revolusi pendidikan berbasis AI. Pembahasan diawali dengan pengantar kecerdasan buatan dalam penilaian pendidikan, kemudian dilanjutkan dengan kajian mengenai etika dan tantangan implementasinya. Buku ini juga menguraikan evolusi penilaian berbasis teknologi AI, transparansi evaluasi menulis, personalisasi penilaian, alur kerja NLP modern berbasis BERT, serta pemanfaatan AI generatif dalam pembuatan soal.

Selanjutnya, pembaca diajak memahami penerapan AI dalam pengukuran pembelajaran melalui permainan, pemanfaatan model bahasa besar dalam penilaian, penggunaan pembelajaran mesin untuk mendeteksi kecurangan asesmen, serta penerapan AI dalam penilaian global. Pembahasan mengenai penilaian dan dukungan terhadap kolaborasi AI memberikan perspektif bahwa teknologi tidak hanya digunakan untuk mengukur kemampuan individual, tetapi juga dapat membantu memahami dan mengembangkan keterampilan kolaboratif peserta didik.

Aspek validitas dan validasi tes memperoleh perhatian khusus dalam buku ini. Penggunaan AI dapat meningkatkan kemampuan asesmen, tetapi sekaligus menghadirkan kompleksitas baru dalam memastikan bahwa hasil penilaian benar-benar akurat, adil, relevan, dan dapat dipertanggungjawabkan. Oleh sebab itu, buku ini juga membahas dampak AI terhadap validitas tes serta potensi transformatif AI dalam pelaksanaan asesmen berskala besar.

Secara keseluruhan, buku ini diharapkan dapat menjadi referensi bagi mahasiswa, dosen, peneliti, pengembang teknologi pendidikan, praktisi asesmen, pengambil kebijakan,

serta pembaca umum yang ingin memahami hubungan antara pendidikan dan kecerdasan buatan. Materi yang disajikan tidak hanya berfokus pada cara kerja teknologi, tetapi juga mendorong pembaca untuk mempertimbangkan implikasi pedagogis, sosial, etis, dan kebijakan dari penerapan AI dalam pendidikan.

Penulis menyadari bahwa perkembangan AI berlangsung sangat cepat. Oleh karena itu, buku ini tidak dimaksudkan sebagai karya yang bersifat final, melainkan sebagai bagian dari upaya untuk terus mengikuti, memahami, dan mengkritisi perubahan pendidikan di era digital. Masukan, kritik, dan saran dari para pembaca sangat diharapkan untuk penyempurnaan edisi berikutnya.

Akhirnya, penulis menyampaikan terima kasih kepada semua pihak yang telah memberikan dukungan, gagasan, dan kontribusi dalam proses penyusunan buku ini. Semoga "*Revolusi Pendidikan di Zaman AI"* dapat memberikan manfaat, memperluas wawasan, dan mendorong pemanfaatan kecerdasan buatan secara bijaksana, bertanggung jawab, serta berorientasi pada peningkatan kualitas pendidikan.

Semangat & Selamat Membaca….!!!!!

Semarang, Oktober 2026
Penulis

Dr. Joseph Teguh Santoso, S.Kom., M.Kom.

DAFTAR ISI

KATA PENGANTAR ..... **i**
DAFTAR ISI ..... **iii**
BAB 1 PENGANTAR KECERDASAN BUATAN DALAM PENILAIAN PENDIDIKAN **1**
1.1 Pengantar Kecerdasan Buatan 1
1.2 Konsep Dan Penerapan AI Dalam Penilaian Pendidikan 5
1.3 Potensi Transformatif AI Dalam Penilaian Pendidikan 13
1.4 Etika, Keadilan, Dan Tantangan Implementasi AI 14
1.5 AI Dalam Transformasi Pendidikan Dan Penilaian 19
BAB 2 ETIKA DAN TANTANGAN PENILAIAN PENDIDIKAN BERBASIS AI **22**
2.1 Dasar Etika Penilaian Berbasis AI 22
2.2 Kerangka Etika Penilaian Berbasis AI 25
2.3 Bias Algoritmik Dalam Penilaian AI 28
2.4 Transparansi Dan Explainability Dalam Penilaian AI 33
2.5 Pengawasan Manusia Dalam Penilaian AI 37
2.6 Privasi Dan Perlindungan Data Dalam Penilaian AI 41
2.7 Studi Kasus Etika Platform Penilaian AI 42
2.8 Arah Pengembangan Penilaian AI Yang Etis 46
BAB 3 EVOLUSI PENILAIAN BERBASIS TEKNOLOGI AI **48**
3.1 Transformasi Penilaian Pendidikan 48
3.2 Perkembangan Penilaian Pendidikan 48
3.3 Transformasi Penilaian Pendidikan Melalui AI 49
3.4 Konsep Dasar Penilaian Berbasis AI 49
3.5 Perkembangan Penilaian Pendidikan 50
3.6 Keterbatasan Penilaian Konvensional 54
3.7 Pendekatan Kontemporer Dalam Penilaian Pendidikan 57
3.8 Pendidikan Berbasis Kompetensi 60
3.9 Instrumen Penilaian Keterampilan Non-Akademis 62
3.10 Implikasi Terhadap Kebijakan Dan Praktik Pendidikan 63
BAB 4 TRANSPARANSI EVALUASI MENULIS BERBASIS AI **66**
4.1 Perangkat Evaluasi Penulisan Otomatis (AWE) 68
4.2 Matriks Kematangan AWE 69
BAB 5 PERSONALISASI PENILAIAN UNTUK KINERJA DAN KETERLIBATAN **101**
5.1 Urgensi Personalisasi Dalam Penilaian 101
5.2 Bentuk Dan Jenis Personalisasi 103
5.3 Bukti Efektivitas Personalisasi 109
5.4 Validasi Dan Isu Terkait Untuk Asesmen Personalisasi 112
5.5 Keadilan Dalam Interpretasi Skor Dari Asesmen Personalisasi 116
5.6 Tantangan Dan Arah Pengembangan Personalisasi 118

BAB 6 ALUR KERJA NLP MODERN DALAM PENGUKURAN PENDIDIKAN 119
6.1 Alur Kerja NLP Berbasis BERT Secara Umum 121
6.2 Alur Kerja NLP Berbasis BERT Untuk Penilaian Otomatis 123
6.3 Alur Kerja NLP Berbasis BERT Untuk Penalaran Dan Pemikiran 133
6.4 Refleksi Atas Alur Kerja NLP Berbasis BERT 144
BAB 7 AI GENERATIF DAN PEMBUATAN SOAL CEPAT 147
7.1 Pengantar Pengembangan Butir Soal Tradisional 147
7.2 Keterbatasan Penulisan Butir Soal Tradisional 148
7.3 Perkembangan Awal Dalam Pembuatan Soal Otomatis (AIG) 149
7.4 Pemodelan Kognitif Sebagai Pendekatan Modern Kedua 151
7.5 Titik Balik Adopsi AI Dalam Pembuatan Butir Soal 154
7.6 Pengodean Otomatis Dan Pembuatan Penjelasan (*Rationale*) 159
7.7 Pertimbangan Hukum Dan Penggunaan Yang Bertanggung Jawab 162
7.8 Tanggung Jawab Dalam Integrasi AI Pada Pembuatan Butir Soal 166
BAB 8 MENGUKUR PEMBELAJARAN MELALUI BERMAIN 168
8.1 Pentingnya Pengukuran Dalam Permainan Edukasi 168
8.2 Desain Penilaian Berbasis Permainan 170
8.3 Psikometri Untuk Permainan Berbasis AI 174
8.4 Telemetri Permainan Untuk Analitik Pembelajaran 177
8.5 Desain Permainan Berbasis AI 178
8.6 Contoh Kasus Penilaian Berbasis GIM Dengan Dukungan AI 179
8.7 Arah Masa Depan Penilaian Berbasis Permainan 181
8.8 Ekosistem Penilaian Berbasis AI 183
BAB 9 MODEL BAHASA BESAR DALAM PENILAIAN 185
9.1 Aplikasi Dan Validitas LLM 185
9.2 Landasan Teknis Untuk LLM 185
9.3 Pendekatan Generatif Untuk Penilaian Otomatis 190
9.4 Menghasilkan Umpan Balik Yang Dipersonalisasi 193
9.5 Menghasilkan Butir Soal Dan Konten 195
9.6 Memastikan Kualitas Asesmen 198
9.7 Integrasi LLM Ke Dalam Aplikasi 204
9.8 Evolusi LLM Dalam Penilaian 208
BAB 10 PEMBELAJARAN MESIN UNTUK DETEKSI KECURANGAN ASESMEN 211
10.1 Evolusi Kecurangan Asesmen Di Era AI 211
10.2 Metode Deteksi Kecurangan 213
10.3 Augmentasi Data Dalam Deteksi Kecurangan 221
10.4 Aplikasi Empiris Augmentasi Data 226
10.5 Deteksi Kecurangan Berbasis Machine Learning Dan AI 231
BAB 11 PENGGUNAAN AI DALAM PENILAIAN GLOBAL 233
11.1 Tantangan Penilaian Di Era Transformasi Pendidikan 233
11.2 Kondisi Terkini Dan Penerapan AI Dalam Penilaian Pendidikan 235

11.3 AI Multimodal Dalam Evaluasi 245
11.4 Penilaian Pendidikan Berbasis AI 251
BAB 12 MENILAI DAN MENDUKUNG KOLABORASI AI 254
12.1 Integrasi AI Untuk Menilai Dan Mendukung Kinerja Kolaboratif 254
12.2 Mengarakterisasi Keterampilan Kolaborasi 256
12.3 Menghubungkan Kerangka Kerja Dengan Pengembangan Penilaian 259
12.4 Dimensi Penilaian Kolaboratif 261
12.5 CoBi Untuk Pembelajaran Kolaboratif 268
12.6 Mengukur Dan Mendukung Keterampilan Kolaborasi 272
BAB 13 DAMPAK AI TERHADAP VALIDITAS DAN VALIDASI TES 275
13.1 Validitas Dalam Pengujian Berbasis AI 275
13.2 Asesmen Yang Disempurnakan Oleh AI Dan Dampaknya 282
13.3 Pembelajaran Validasi Untuk Asesmen Berbasis AI 286
13.4 Metode Evaluasi Penilaian Berbasis AI 291
13.5 Kompleksitas Validasi Tes Akibat Kecerdasan Buatan 295
BAB 14 POTENSI TRANSFORMATIF AI DALAM ASESMEN SKALA BESAR 296
14.1 Masa Depan Asesmen Skala Besar Di Era AI 296
14.2 Kasus Penggunaan Penting AI Dalam Asesmen Berskala Besar 297
14.3 Dampak AI Terhadap Praktik Penilaian 310
14.4 Integrasi AI Ke Dalam Penilaian Pendidikan 312
14.5 Masa Depan Asesmen Berbasis AI 314
14.6 AI Yang Bertanggung Jawab Dalam Penilaian Pendidikan 315
DAFTAR PUSTAKA 316

BAB 1
PENGANTAR KECERDASAN BUATAN DALAM PENILAIAN PENDIDIKAN

1.1 PENGANTAR KECERDASAN BUATAN

Kecerdasan Buatan atau *Artificial Intelligence* (AI) adalah cabang ilmu komputer yang berfokus pada pengembangan algoritma yang mampu melakukan tugas-tugas seperti memahami bahasa alami, belajar dari pengalaman, memprediksi hasil, menilai karya tulis, dan mengenali ucapan. Tugas-tugas semacam itu biasanya memerlukan kecerdasan manusia untuk melaksanakannya. John McCarthy, yang sering disebut sebagai bapak AI, mendefinisikan AI sebagai "ilmu dan rekayasa pembuatan mesin cerdas, khususnya program komputer cerdas." AI telah diterapkan di berbagai industri, termasuk pendidikan, karena kemampuannya menangani tugas-tugas yang berulang, beroperasi tanpa henti, serta menghilangkan kekhawatiran terkait risiko fisik. AI merupakan bidang luas yang mencakup berbagai disiplin ilmu seperti ilmu komputer, analisis data, linguistik, dan psikologi. Inti dari AI terletak pada kapasitasnya untuk melakukan simulasi kognitif dan kemampuan beradaptasi, yang memungkinkan mesin menjalankan tugas sederhana maupun kompleks dengan tingkat otonomi yang bervariasi. Berdasarkan kemampuannya, AI dapat dikategorikan menjadi tiga jenis utama: AI sempit (*narrow AI* atau *weak AI*), AI umum (*general AI* atau *strong AI*), dan super AI.

AI sempit dikembangkan untuk melakukan tugas-tugas spesifik atau serangkaian tugas yang terbatas. Contoh tugas-tugas ini meliputi asisten virtual seperti Siri atau Alexa, algoritma rekomendasi yang digunakan oleh platform *streaming*, dan sistem penilaian otomatis konvensional. Meskipun AI sempit unggul dalam ranah tugas yang ditetapkan baginya, sistem ini tidak dapat berfungsi di luar batasan yang telah diprogram. Aplikasi tingkat lanjut, seperti model bahasa besar (*Large Language Models* atau LLM) yang digunakan dalam penilaian otomatis, menunjukkan kecanggihan yang lebih tinggi di bidang khususnya namun tetap terbatas dalam lingkup AI sempit.

AI umum adalah bentuk AI hipotetis yang mampu memahami, mempelajari, dan menerapkan pengetahuan pada berbagai tugas, mirip dengan kemampuan kognitif manusia. Hingga saat ini, AI umum masih sebatas konsep tanpa implementasi di dunia nyata. Super AI saat ini masih bersifat teoretis semata. Jika kelak terwujud, sistem ini akan melampaui manusia dalam hal pembelajaran, penalaran, dan kemampuan kognitif lainnya. Di antara ketiga jenis AI tersebut, hanya AI sempit yang telah dikembangkan dan diterapkan dalam skenario dunia nyata. Berbeda dengan AI umum (*general AI*) yang memiliki kapasitas untuk melakukan tugas intelektual apa pun yang dapat dilakukan manusia, atau AI super (*super AI*) yang melampaui kecerdasan manusia, AI sempit (*narrow AI*) bersifat khusus; ia unggul dalam tugas-tugas tertentu namun tidak memiliki kemampuan untuk melakukan generalisasi lintas domain yang berbeda.

Karakteristik dan Tujuan Utama

AI telah bertransformasi dari sistem berbasis aturan sederhana di masa lalu menjadi teknologi yang canggih dan adaptif. Beberapa atribut mendasar mencirikan sistem AI modern. Dalam bidang penilaian pendidikan, fitur-fitur penting yang menonjol meliputi pemrosesan bahasa alami (*natural language processing* atau NLP), kemampuan adaptasi, kemampuan prediktif, penalaran, dan otonomi. Fitur mendasar dari AI adalah kapasitasnya untuk belajar dari data.

Kemampuan untuk melakukan proses berulang (*iterasi*) sering kali dimungkinkan melalui pembelajaran mesin (*machine learning* atau ML), yaitu bagian dari AI di mana algoritma dilatih menggunakan kumpulan data yang luas untuk mengidentifikasi pola serta membuat prediksi atau keputusan tanpa perlu diprogram secara eksplisit untuk setiap tugas. Sebagai contoh, dalam penilaian esai otomatis, AI belajar dengan menganalisis data esai aktual beserta penilaian manusia yang terkait, sembari menerapkan kriteria penilaian yang telah diprogram sebelumnya. Selain ML, AI juga dapat belajar melalui pembelajaran penguatan (*reinforcement learning*) yang melibatkan pembelajaran dari interaksi dengan lingkungan serta penalaran simbolik yang bergantung pada aturan-aturan yang telah ditetapkan sebelumnya.

NLP memungkinkan AI untuk memahami, menafsirkan, dan menganalisis bahasa manusia. NLP sangat penting untuk berbagai aplikasi, seperti *chatbot*, penyediaan layanan dukungan pelanggan, layanan penerjemahan bahasa yang menjembatani komunikasi lintas bahasa, serta alat pembuatan konten otomatis yang mampu menghasilkan materi tertulis secara efisien. Sebagai contoh, dalam sistem pembuatan butir soal otomatis, generator soal berbasis NLP dapat menganalisis soal tes yang sudah ada, memahami struktur dan kontennya, serta menghasilkan soal baru yang sesuai dengan gaya dan tingkat kesulitan soal tersebut.

Sistem AI dapat menyesuaikan operasinya berdasarkan data baru atau kondisi yang berubah, sehingga kinerjanya meningkat seiring waktu dan sistem mampu menangani situasi tak terduga dengan lebih efektif. Sebagai contoh, tes adaptif menyesuaikan tingkat kesulitan soal berikutnya berdasarkan jawaban mahasiswa sebelumnya. Jika mahasiswa menjawab soal dengan benar, sistem akan menyajikan soal yang lebih menantang. Sebaliknya, jika mahasiswa menjawab salah, sistem akan memberikan soal dengan tingkat kesulitan yang lebih rendah.

Salah satu kemampuan mendasar sistem AI adalah kemampuannya memanfaatkan analisis data untuk memprediksi hasil di masa depan dengan mengevaluasi data terkini maupun historis. Melalui algoritma canggih dan teknik pembelajaran mesin (ML), sistem ini mampu mengungkap pola, tren, dan korelasi yang mungkin tidak terlihat secara langsung. Sebagai contoh, dalam konteks pendidikan, sistem AI dapat menganalisis pekerjaan rumah, hasil tes, dan partisipasi mahasiswa di kelas. Dengan mengevaluasi berbagai data tersebut, AI dapat memprediksi kinerja penilaian mahasiswa di masa mendatang secara akurat. Selain itu, sistem ini dapat memberikan rekomendasi yang dipersonalisasi, seperti menyarankan latihan soal yang terarah, menyusun jadwal pengulangan materi yang disesuaikan, atau menyediakan sumber daya tambahan untuk mengatasi aspek-aspek yang perlu ditingkatkan. Kemampuan prediksi ini memungkinkan pendidik untuk menerapkan intervensi yang tepat waktu dan

personal, sehingga meningkatkan pengalaman belajar setiap mahasiswa serta potensi keberhasilan akademis mereka.

Sistem AI menganalisis data dan mengambil keputusan melalui inferensi logis yang meniru cara berpikir manusia. Sebagai contoh, dalam dunia pendidikan, sistem AI yang digunakan untuk pembelajaran personal dapat mengevaluasi data kinerja mahasiswa seperti hasil tes dan pengumpulan tugas untuk mengidentifikasi kekuatan dan kelemahan mereka. Berdasarkan analisis tersebut, AI menyusun materi dan strategi pembelajaran yang disesuaikan dengan kebutuhan spesifik mahasiswa; hal ini serupa dengan cara seorang pengajar menyesuaikan metode pengajaran berdasarkan kemajuan dan tantangan yang dihadapi mahasiswa.

Banyak sistem AI beroperasi dengan tingkat otonomi tertentu, yang berarti sistem tersebut mampu menjalankan tugas secara mandiri setelah dilatih atau diprogram dengan tepat. Karakteristik ini sangat penting bagi berbagai aplikasi. Misalnya, AI dapat secara otonom memantau kemajuan mahasiswa dalam platform pembelajaran personal, menyesuaikan materi pembelajaran dengan kebutuhan individu, serta memberikan umpan balik tanpa perlu pengawasan terus-menerus dari manusia. Sifat kerja yang mandiri ini memungkinkan pendidik untuk lebih fokus pada tugas pengajaran yang lebih kompleks, sementara AI menangani penilaian rutin dan penyampaian materi yang dipersonalisasi. Meskipun fitur-fitur seperti pengenalan pola dan pengenalan gambar merupakan hal yang penting, fitur-fitur tersebut tidak sepenuhnya mencerminkan sifat AI yang lebih luas, fleksibel, dan adaptif.

Tujuan menyeluruh dari pembahasan dan penerapan AI mencakup berbagai sasaran. Hal ini meliputi peningkatan kemampuan manusia melalui peningkatan produktivitas, penyederhanaan tugas rutin dan membosankan demi efisiensi operasional, pemecahan masalah dunia nyata yang kompleks yang melampaui batasan metode komputasi tradisional, serta mendorong inovasi di berbagai bidang. Pada akhirnya, AI berupaya mengembangkan sistem yang mampu beradaptasi, belajar, dan berkembang, serta memberdayakan individu maupun organisasi.

AI dirancang untuk meningkatkan kemampuan manusia dengan menyederhanakan tugas-tugas yang kompleks dan mendorong efisiensi. Melalui algoritma canggih dan *machine learning* (ML), AI merevolusi cara kita menghadapi berbagai tantangan. Sebagai contoh, perangkat diagnostik berbasis AI memungkinkan pendidik mengidentifikasi kelemahan mahasiswa dengan akurasi dan kecepatan yang lebih baik, sehingga intervensi yang lebih tepat sasaran dapat dilakukan untuk memenuhi kebutuhan individu secara efektif. Dengan demikian, pendidik dapat lebih fokus pada aspek strategis pengajaran, yang pada gilirannya menghasilkan capaian yang lebih baik dan produktivitas yang lebih tinggi.

Dengan mengotomatisasi tugas-tugas rutin dan berulang, AI dapat meningkatkan efisiensi organisasi secara signifikan. Otomatisasi ini mencakup berbagai aktivitas, seperti proses administratif, pembersihan data, dan interaksi antara manusia dengan perangkat lunak. Akibatnya, AI membebaskan karyawan dari tugas-tugas yang monoton dan melelahkan, sehingga mereka dapat mencurahkan tenaga untuk fokus pada pekerjaan yang lebih kreatif,

strategis, dan bernilai tinggi. Transisi ini tidak hanya memperlancar alur kerja tetapi juga mendorong pola pikir strategis.

AI memiliki potensi besar untuk mengatasi tantangan kompleks yang sering kali melampaui kemampuan manusia. Dengan memanfaatkan algoritma canggih dan daya komputasi yang besar, sistem AI dapat memproses data dalam jumlah masif untuk menemukan pola dan wawasan yang mungkin terlewatkan oleh metode tradisional. Peran transformatif AI dalam pendidikan terlihat jelas pada kemampuannya untuk mempersonalisasi pengalaman belajar sesuai dengan kebutuhan dan preferensi unik setiap mahasiswa. Selain itu, AI meningkatkan strategi pengajaran dengan memberikan umpan balik secara *real-time* dan mengidentifikasi metode instruksional yang tepat.

AI merupakan pendorong inovasi yang kuat, yang memacu kemajuan di berbagai sektor serta melahirkan produk, layanan, dan teknologi yang inovatif. Dampak transformatifnya terlihat nyata dalam pendidikan dan penilaian yang dipersonalisasi, yang mengubah paradigma tradisional dan menciptakan peluang baru. AI menyempurnakan metode evaluasi dengan menjadikannya lebih personal, adaptif, dan efisien. Kemampuan adaptasi ini menghasilkan penilaian yang lebih akurat dan komprehensif, memberikan wawasan lebih mendalam bagi pendidik mengenai kemajuan mahasiswa serta memungkinkan pemberian dukungan yang disesuaikan dengan kebutuhan individu. Seiring berkembangnya teknologi AI, perannya dalam penilaian pendidikan diprediksi akan semakin besar, menawarkan peluang lebih lanjut untuk meningkatkan keberhasilan akademik dan menciptakan suasana belajar yang lebih inklusif.

Sejarah Singkat

AI telah berkembang melalui beberapa tahapan kunci. Setiap fase telah membentuk perkembangannya secara signifikan, mulai dari masa awal AI simbolik hingga kemajuan yang terlihat pada *Machine Learning* (ML), *deep learning*, dan *Large Language Models* (LLM). Selain itu, tokoh-tokoh terkemuka dan inovasi terkini turut berkontribusi pada kemajuannya. Asal-usul AI dapat ditelusuri kembali ke pertengahan abad ke-20, saat AI simbolik yang berfokus pada sistem berbasis aturan dan penalaran logis diperkenalkan. Konferensi Dartmouth tahun 1956, yang diselenggarakan oleh John McCarthy dan rekan-rekannya, secara luas dianggap sebagai titik awal formal kelahiran AI. Karya perintis ini meletakkan dasar bagi perangkat penilaian awal yang menggunakan logika berbasis aturan untuk mengevaluasi kinerja mahasiswa. Sebagai contoh, pada awal tahun 1960-an, sistem seperti "*General Problem Solver*" yang dikembangkan oleh Allen Newell dan Herbert A. Simon merupakan salah satu yang pertama menerapkan logika berbasis aturan dalam pemecahan masalah; hal ini membuka jalan bagi penilaian pendidikan di masa depan yang memanfaatkan metodologi serupa.

Pada tahun 1980-an dan 1990-an, arah pembahasan AI beralih ke *machine learning* (ML), di mana sistem belajar dari data alih-alih hanya mengandalkan metode berbasis aturan. Pergeseran ini mendorong pengembangan algoritma canggih seperti pohon keputusan (*decision trees*) dan jaringan saraf (*neural networks*), yang secara signifikan meningkatkan analisis data pendidikan serta otomatisasi penilaian, termasuk tugas seperti pemberian nilai

pada soal pilihan ganda. Sebagai contoh, dalam lingkungan kelas, sistem berbasis ML yang menggunakan jaringan saraf dapat menganalisis pola respons mahasiswa untuk mengidentifikasi kesalahpahaman umum dan memberikan umpan balik yang tepat sasaran, sehingga mengefisienkan proses penilaian.

Kemajuan terkini dalam AI mencakup pengembangan teknologi *deep learning*, yang memanfaatkan jaringan saraf dengan banyak lapisan yang mampu mempelajari representasi data pada tingkat abstraksi yang semakin tinggi. Kemampuan ini memungkinkan klasifikasi yang lebih akurat dalam kumpulan data yang kompleks. Dalam dunia pendidikan, *deep learning* telah digunakan untuk meningkatkan akurasi penilaian esai otomatis serta menyempurnakan sistem pembelajaran adaptif dengan menganalisis dan menafsirkan data mahasiswa guna menyesuaikan pengalaman pendidikan. LLM (*Large Language Models*), seperti *Generative Pre-trained Transformer* (GPT)-3 dan GPT-4, merupakan kemajuan signifikan lainnya dalam bidang AI. Model-model ini unggul dalam memproses dan menghasilkan teks yang menyerupai tulisan manusia, sehingga mengubah cara analisis dan respons terhadap masukan mahasiswa. Dengan memanfaatkan LLM, pendidik dapat menyediakan perangkat yang lebih canggih untuk menganalisis respons mahasiswa yang kompleks dan memberikan umpan balik yang dipersonalisasi. Model-model ini telah diintegrasikan ke dalam sistem pembelajaran adaptif dan penilaian esai otomatis, yang menawarkan pendekatan lebih bernuansa dan responsif terhadap penilaian pendidikan.

Tokoh dan Peristiwa Penting dalam Evolusi AI

Selain tokoh-tokoh perintis seperti Alan Turing dan Marvin Minsky yang meletakkan dasar bagi AI, para peneliti kontemporer termasuk Yann LeCun, Geoffrey Hinton, dan Yoshua Bengio telah mendorong kemajuan terkini di bidang ini. Karya inovatif mereka mengenai jaringan saraf (*neural networks*) dan pembelajaran mendalam (*deep learning*) telah memicu kemajuan signifikan dalam AI. Pembahasan yang dilakukan oleh LeCun, Hinton, dan Bengio telah menghasilkan pengembangan algoritma canggih yang menjadi landasan bagi sistem AI modern, yang kini semakin banyak dimanfaatkan dalam penilaian pendidikan.

Munculnya *Large Language Model* (LLM) yang canggih serta integrasinya ke dalam teknologi pendidikan menandai tonggak sejarah penting dalam evolusi AI. Model-model ini—sebagaimana dikaji dalam pembahasan Collins dan rekan-rekan mewakili lompatan besar dalam kemampuan sistem AI untuk memahami dan menghasilkan teks yang menyerupai tulisan manusia, sehingga memperluas penerapannya dalam konteks pendidikan. Penerapan teknologi ini telah membuka peluang baru untuk meningkatkan penilaian pendidikan, serta menawarkan solusi inovatif yang sebelumnya tidak dapat dicapai.

1.2 KONSEP DAN PENERAPAN AI DALAM PENILAIAN PENDIDIKAN

Berbeda dengan pemrograman tradisional, di mana instruksi eksplisit diberikan untuk setiap tugas, algoritma pembelajaran mesin (ML) meningkatkan kinerjanya dengan mengenali wawasan dari kumpulan data yang besar. Berdasarkan pendekatan yang digunakan dalam melatih model serta tujuannya, terdapat tiga jenis utama ML yang umum digunakan:

pembelajaran terawasi (*supervised learning*), pembelajaran tak terawasi (*unsupervised learning*), dan pembelajaran penguatan (*reinforcement learning*).

Dalam pembelajaran terawasi, tujuan utamanya adalah memprediksi hasil untuk data baru yang belum pernah dilihat sebelumnya dengan memanfaatkan pola yang dipelajari dari set pelatihan berlabel. Algoritma umum yang digunakan dalam pembelajaran terawasi meliputi pohon keputusan (*decision trees*) dan *support vector machines* (SVM). Sebagai contoh, dalam penilaian pendidikan, pembelajaran terawasi dapat diterapkan untuk mengklasifikasikan mahasiswa ke dalam tingkat kinerja yang berbeda berdasarkan contoh klasifikasi benar yang digunakan selama pelatihan. Contoh tambahan dapat ditemukan dalam Alachiotis dan rekan-rekan.

Sebaliknya, pembelajaran tak terawasi berfokus pada pengungkapan pola tersembunyi di dalam data. Teknik standar yang digunakan meliputi pengelompokan (*clustering*) dan reduksi dimensi. Metode pengelompokan, seperti *k-means* dan *hierarchical clustering*, mengelompokkan titik-titik data berdasarkan kemiripan, sehingga mengungkap pengelompokan alami dalam kumpulan data yang dapat membantu mengidentifikasi pola, membuat prediksi, atau menjadi dasar pengambilan keputusan. Teknik untuk mereduksi dimensi, seperti analisis komponen utama (*principal component analysis*) dan *t-distributed stochastic neighbor embedding*, menyederhanakan data dengan mengurangi jumlah variabel namun tetap mempertahankan hubungan-hubungan penting. Sebagai ilustrasi, pembelajaran tak terawasi dapat mengelompokkan mahasiswa berdasarkan data kinerja atau mengidentifikasi tren dalam perilaku belajar.

Di sisi lain, pembelajaran penguatan bertujuan untuk mengembangkan strategi yang memaksimalkan imbalan kumulatif dari waktu ke waktu dengan belajar dari kesalahan. Pembelajaran penguatan sangat relevan untuk pengembangan sistem pembelajaran adaptif yang menyesuaikan strategi instruksional berdasarkan respons dan kemajuan mahasiswa. Setiap jenis ML memiliki penerapan unik dan cocok untuk masalah yang berbeda, yang mencerminkan fleksibilitas dan potensi teknologi ML di berbagai bidang.

Algoritma dan Model Utama

Beberapa algoritma dan model penting memegang peranan kunci dalam penerapan ML untuk penilaian pendidikan. Algoritma-algoritma ini menjadi tulang punggung bagi sistem canggih yang menganalisis, memprediksi, dan meningkatkan hasil pendidikan. Setiap model memiliki keunggulan dan kegunaan yang unik, sehingga memahami peran serta fungsinya menjadi hal yang sangat penting. Pohon keputusan (*decision trees*) adalah model berbentuk pohon yang digunakan untuk tugas klasifikasi dan regresi. Model ini bekerja dengan cara memecah kumpulan data secara rekursif menjadi beberapa subset berdasarkan nilai fitur, sehingga menghasilkan struktur menyerupai pohon di mana setiap simpul (*node*) mewakili kriteria keputusan. Sifat pohon keputusan yang transparan dan mudah diinterpretasikan membuatnya sangat berguna untuk tugas-tugas yang menuntut hasil yang jelas dan dapat dipahami, seperti mendiagnosis kinerja mahasiswa dan mengidentifikasi faktor-faktor yang memengaruhi hasil pendidikan. Sebagai contoh, model ini dapat digunakan untuk

memprediksi kinerja mahasiswa serta menentukan faktor-faktor yang berkontribusi terhadap keberhasilan atau kegagalan akademik.

Jaringan saraf (*neural networks*) adalah model komputasi yang terinspirasi oleh otak manusia, yang terdiri dari lapisan-lapisan simpul saling terhubung yang mampu memproses dan mempelajari data secara fleksibel. Kemampuannya dalam memodelkan hubungan non-linear yang kompleks memungkinkan model ini mengungkap pola-pola rumit di dalam data. Kerangka kerja *deep learning* mutakhir, seperti *convolutional neural networks* (CNN) dan *recurrent neural networks* (RNN), memungkinkan jaringan saraf melakukan tugas-tugas canggih seperti memberikan rekomendasi pembelajaran yang dipersonalisasi dan melakukan analisis prediktif. Model-model ini memperluas batasan dalam penilaian pendidikan serta menunjukkan efektivitasnya dalam aplikasi yang kompleks, seperti penilaian esai dan pemrosesan bahasa alami atau NLP.

SVM banyak diterapkan dalam tugas klasifikasi dan regresi karena kemampuannya menentukan *hyperplane* ideal guna memaksimalkan jarak pemisah antar-kelompok. Dalam masalah klasifikasi, khususnya klasifikasi biner, SVM sangat efektif membedakan dua kelompok dengan mengidentifikasi batas pemisah terbaik di antara keduanya. Sebagai contoh, SVM dapat memprediksi hasil seperti kemungkinan keberhasilan atau kegagalan mahasiswa berdasarkan fitur-fitur tertentu, misalnya nilai ujian atau metrik keterlibatan.

Dengan memanfaatkan fungsi kernel untuk memetakan data input ke dalam ruang berdimensi lebih tinggi, SVM sangat andal dalam menangkap hubungan *non-linear* yang kompleks. Selain itu, kemampuannya untuk bekerja dengan baik pada dataset yang lebih kecil sembari meminimalkan *overfitting* semakin menegaskan fleksibilitasnya. Setiap algoritma menawarkan kemampuan unik untuk penilaian pendidikan, sehingga meningkatkan presisi, skalabilitas, dan adaptabilitas aplikasi *machine learning* (ML). Dengan memanfaatkan beragam pendekatan ini, para pendidik dan peneliti dapat mengembangkan metode penilaian yang lebih inovatif, personal, dan efektif untuk mendukung spektrum kebutuhan serta konteks pembelajaran dan penilaian yang luas.

Pemrosesan Bahasa Alami (*Natural Language Processing*)

NLP memungkinkan komputer berinteraksi dengan manusia melalui bahasa alami secara bermakna dan efisien. Proses ini dimulai dengan pengumpulan data, di mana data tekstual dalam jumlah besar diperoleh dari situs web, platform media sosial, dokumen, dan repositori lainnya. Selanjutnya, dilakukan pra-pemrosesan data; ini merupakan tahap krusial yang melibatkan pembersihan dan standardisasi teks guna memastikan akurasi dan konsistensi untuk analisis tahap berikutnya.

Teknik-teknik seperti penghapusan *stop-word* (kata umum yang tidak membawa informasi penting), tokenisasi (memecah teks menjadi kata atau frasa), lematisasi (mengubah kata ke bentuk dasarnya), dan *stemming* (mereduksi kata ke bentuk akarnya) diterapkan untuk menjamin konsistensi dan keseragaman di seluruh dataset. Setelah pra-pemrosesan, teks diubah menjadi representasi numerik yang sesuai untuk model *Machine Learning* (ML). Teknik umum yang digunakan meliputi *bag-of-words* (merepresentasikan teks berdasarkan frekuensi kata) dan *N-grams* (menangkap urutan kata). Metode yang lebih canggih, seperti *embedding*

Word2Vec dan GloVe, mengubah teks menjadi vektor yang menangkap fitur sintaksis (urutan dan struktur kata) maupun fitur semantik (makna dan konteks).

Pada tahap ini, algoritma ML atau model *deep learning* digunakan untuk menganalisis data dan menjalankan berbagai tugas NLP, termasuk klasifikasi teks, klasifikasi emosi, penandaan entitas, dan penerjemahan bahasa. Bergantung pada kompleksitas tugas dan data, algoritma seperti Naïve Bayes, SVM, atau jaringan saraf (*neural networks*) dan *transformer* yang canggih dapat diterapkan. Setelah pelatihan selesai, model dievaluasi menggunakan metrik seperti akurasi, presisi, dan *recall* untuk mengukur efektivitasnya. Berdasarkan hasil tersebut, model dapat disesuaikan (*fine-tuning*) dan dioptimalkan. Begitu hasil yang diinginkan tercapai, model NLP diterapkan untuk penggunaan di dunia nyata. Pemantauan berkelanjutan dan perbaikan secara iteratif memastikan model tetap efektif saat memproses data baru.

Mekanisme NLP yang rumit memperlihatkan interaksi kompleks antara berbagai komponen yang bekerja sama untuk mencapai pemahaman bahasa yang menyeluruh. Bagian sentral dari NLP adalah analisis sintaksis, yang menelaah struktur kata dengan memecahnya menjadi bagian-bagian penyusunnya guna memahami bagaimana kata-kata tersebut digabungkan. Lapisan dasar ini meletakkan landasan bagi analisis sintaksis, di mana hubungan gramatikal antar-kata dievaluasi secara cermat untuk memastikan kalimat mematuhi kaidah bahasa. Saat pemahaman kita semakin mendalam, muncul analisis semantik yang berfokus pada makna kata dan frasa dalam berbagai konteks. Aspek ini sangat penting karena mengubah rangkaian teks sederhana menjadi konsep yang bermakna. Tahap ini dilengkapi dengan analisis wacana, yang mengkaji bagaimana kalimat saling berkaitan dan bagaimana makna dibangun dalam rangkaian teks yang lebih panjang, sehingga memungkinkan sistem menjaga koherensi dan alur komunikasi.

Terakhir, analisis pragmatik meningkatkan pemahaman dengan menggali wawasan dari teks. Proses ini melibatkan penafsiran makna yang dimaksudkan di balik kata-kata dengan mempertimbangkan faktor-faktor seperti niat pembicara, konteks sosial, dan nuansa budaya. Meskipun berbeda satu sama lain, setiap komponen ini bekerja sama untuk meningkatkan kinerja sistem NLP, mengubah data mentah menjadi informasi bermakna, serta memfasilitasi interaksi yang lancar antara manusia dan mesin. Upaya kolektif ini menyoroti kecanggihan teknologi NLP dan menunjukkan potensinya dalam memahami serta memproses bahasa manusia secara akurat dan mendalam.

Penerapan

NLP telah mentransformasi berbagai industri seperti layanan kesehatan, hukum, dan pendidikan. Integrasinya ke dalam *chatbot* telah menjadi hal penting dalam menyediakan layanan pelanggan 24 jam sehari, 7 hari seminggu, yang memungkinkan bisnis menangani berbagai pertanyaan secara *real-time*. Kemampuan ini tidak hanya meningkatkan pengalaman pengguna tetapi juga memperbaiki efisiensi operasional. NLP sangat bermanfaat bagi pemula yang menggunakan alat pembelajaran atau penilaian daring, karena menyederhanakan interaksi dan membuat platform tersebut lebih mudah diakses.

Di bidang penilaian pendidikan, NLP mendorong inovasi yang signifikan. Sebagai contoh, teknologi pengenalan suara (*speech recognition*) yang didukung oleh NLP

memungkinkan penyertaan lebih banyak respons lisan dalam penilaian sesuatu yang sebelumnya terbatas karena besarnya tenaga manusia yang dibutuhkan untuk mendengarkan dan menilai respons semacam itu. Dengan NLP, penilaian dapat mencakup beragam format pertanyaan, seperti jawaban lisan, sehingga memberikan evaluasi yang lebih komprehensif terhadap kemampuan mahasiswa.

Teknik NLP dapat meningkatkan pembelajaran yang dipersonalisasi dengan memberikan umpan balik instan berbasis AI terhadap tulisan mahasiswa. Kemampuan ini memungkinkan pengajar untuk memberikan bimbingan yang lebih terarah sekaligus memfasilitasi mahasiswa dalam melakukan pembelajaran mandiri. Kapasitas NLP untuk menganalisis teks dan memberikan saran yang disesuaikan menciptakan pengalaman belajar personal yang mengakomodasi kebutuhan serta perkembangan setiap mahasiswa, yang pada akhirnya meningkatkan hasil pendidikan.

Jaringan Saraf dan Pembelajaran Mendalam (*Deep Learning*)

Inti dari sebuah jaringan saraf terdiri atas neuron, yaitu unit dasar yang saling terhubung melalui berbagai lapisan. Lapisan masukan (*input layer*) menerima data, sedangkan lapisan keluaran (*output layer*) menghasilkan hasil akhir. Di antara keduanya terdapat lapisan tersembunyi (*hidden layers*), tempat jaringan melakukan perhitungan dan transformasi yang memungkinkannya mengenali pola serta hubungan yang kompleks di dalam data. Struktur dasar jaringan saraf menghubungkan lapisan-lapisan secara berurutan, meneruskan informasi dari lapisan masukan melalui lapisan tersembunyi menuju lapisan keluaran. Untuk menyempurnakan proses pembelajaran, bobot (*weights*) dan bias diperkenalkan. Bobot mengontrol kekuatan sinyal yang diteruskan antar-neuron dengan melakukan penskalaan terhadap nilai masukan. Di sisi lain, bias memungkinkan jaringan untuk menyesuaikan fungsi aktivasi, sehingga memberikan fleksibilitas.

Dengan menambahkan suku konstanta pada jumlah masukan yang telah diberi bobot, bias memungkinkan jaringan membuat prediksi yang lebih akurat, bahkan ketika nilai masukan tertentu adalah nol. Penyesuaian ini memungkinkan model untuk lebih sesuai dengan data dan membantu mencegah *underfitting*. Data masukan awal bergerak maju melalui jaringan untuk menghasilkan keluaran (*forward propagation*). Pada jaringan saraf yang lebih canggih, *backpropagation* menyesuaikan bobot dan bias dengan menghitung selisih kesalahan (*error*) antara keluaran prediksi dan keluaran aktual. *Gradient descent*, sebuah teknik optimasi, membantu mengurangi kesalahan ini dengan menghitung gradien (atau kemiringan) fungsi kesalahan terhadap parameter model. Teknik ini kemudian memperbarui parameter ke arah yang menurunkan tingkat kesalahan, dengan mengambil langkah-langkah kecil menuju minimalisasi kesalahan tersebut. Seiring waktu, proses iteratif ini menyempurnakan model serta meningkatkan akurasinya dalam membuat prediksi.

Jaringan saraf menunjukkan kinerja luar biasa dalam berbagai aplikasi, khususnya terkait pembuatan prediksi dan identifikasi pola di dalam kumpulan data yang besar. Sebagai contoh, dalam konteks pendidikan, jaringan saraf dapat memprediksi kinerja mahasiswa di akhir tahun dengan menganalisis tren dari penilaian berkala, penyerahan pekerjaan rumah, dan catatan partisipasi. Selain itu, jaringan saraf sangat efektif dalam mengenali pola yang

mungkin luput dari perhatian. Kemampuan ini dapat diterapkan untuk mendeteksi perilaku yang tidak wajar atau mencurigakan selama ujian, seperti potensi tindakan kecurangan. Dengan menganalisis penyimpangan dalam pola respons atau mengidentifikasi anomali pada waktu yang dihabiskan untuk pertanyaan tertentu, jaringan saraf menawarkan lapisan analisis canggih yang mendukung integritas proses penilaian. Selain itu, jaringan saraf dapat membantu pengembangan kurikulum dengan menganalisis data kinerja historis untuk merekomendasikan jalur pembelajaran yang dipersonalisasi bagi mahasiswa, memastikan bahwa materi pembelajaran dirancang agar sesuai dengan kebutuhan dan gaya belajar mereka yang khas.

Kemajuan dalam *Deep Learning* dan Penerapannya

Jaringan saraf merupakan fondasi dari *deep learning* 7, sebuah bidang khusus dalam AI yang menggunakan jaringan saraf berlapis-lapis untuk menangkap dan memodelkan pola-pola kompleks dalam kumpulan data berskala besar. Perkembangan terkini di bidang ini menyoroti evolusi arsitektur jaringan saraf, khususnya *deep neural networks* (DNN). DNN memungkinkan pelatihan model dengan jumlah parameter yang sangat besar, sehingga secara signifikan meningkatkan akurasi dan efisiensi tugas-tugas seperti pengenalan gambar dan NLP. Dalam sistem penilaian otomatis, DNN dapat mengevaluasi esai dengan menganalisis kompleksitas, koherensi, dan tata bahasa. Kemampuan ini memungkinkan penilaian yang lebih bernuansa dan sangat mendekati standar penilaian manusia.

CNN merupakan kemajuan lain dalam *deep learning* yang dirancang khusus untuk memproses data visual. CNN unggul dalam deteksi objek dan klasifikasi gambar dengan memanfaatkan hierarki spasial. Dalam dunia pendidikan, CNN dapat dimanfaatkan dalam perangkat yang menganalisis tulisan tangan mahasiswa pada dokumen hasil pemindaian atau gambar, sehingga memungkinkan penilaian otomatis terhadap tugas tertulis serta mengidentifikasi aspek-aspek di mana mahasiswa mungkin mengalami kesulitan dalam keterampilan menulis mereka.

RNN, beserta varian lanjutannya seperti jaringan *long short-term memory* (LSTM), dirancang khusus untuk menangani data sekuensial. Kemampuan ini menjadikannya sangat efektif untuk tugas-tugas seperti prediksi deret waktu (*time-series*), pengenalan ucapan, dan pemodelan bahasa. Dalam konteks pendidikan, LSTM dapat diintegrasikan ke dalam sistem bimbingan belajar cerdas (*intelligent tutoring systems*), yang memungkinkan sistem tersebut beradaptasi dengan kebutuhan unik setiap pembelajar melalui analisis interaksi mereka dari waktu ke waktu.

Beragam arsitektur jaringan saraf ini mencerminkan semakin canggihnya model *deep learning*, yang memungkinkannya menangani berbagai tantangan dunia nyata yang kompleks secara lebih luas, khususnya di bidang pendidikan. Seiring dengan kemajuan teknologi ini, tercipta landasan bagi inovasi dalam penilaian otomatis, platform pembelajaran yang dipersonalisasi, serta berbagai sistem lainnya. Fleksibilitas yang kian berkembang ini menyoroti potensi transformatif *deep learning* dalam lingkungan pendidikan, yang pada akhirnya menghasilkan peningkatan capaian bagi mahasiswa maupun pendidik.

Gambaran Umum Kerangka Kerja

Dibangun di atas arsitektur *deep learning* yang canggih, khususnya *transformer*, LLM unggul dalam memproses dan memprediksi teks. Dilatih menggunakan kumpulan data yang luas dan beragam yang mencakup berbagai bahasa dan topik, LLM mampu menangkap serta memahami nuansa bahasa manusia. Pelatihan ekstensif ini memungkinkan model-model tersebut berkinerja sangat baik dalam berbagai tugas terkait bahasa, termasuk *prompt engineering*. Kemampuan mereka untuk menghasilkan teks yang koheren dan relevan secara kontekstual menjadikan mereka komponen tak tergantikan dalam pembuatan konten dan aplikasi penulisan otomatis. Berikut adalah beberapa contoh utama LLM:

- GPT-4, yang dikembangkan oleh OpenAI, merupakan kemajuan terbaru dalam seri GPT. Model ini menunjukkan kemampuan luar biasa dalam menghasilkan teks yang koheren dan memahami instruksi (*prompt*) yang kompleks, serta memperlihatkan keahlian tinggi dalam pembuatan konten dan pelaksanaan aplikasi NLP tingkat lanjut.
- Claude, yang dikembangkan oleh Anthropic, dinamai berdasarkan Claude Shannon, "bapak teori informasi." Model ini mengedepankan interaksi yang bermanfaat, etis, dan kolaboratif. Dengan fokus pada keamanan dan praktik AI yang bertanggung jawab, Claude menawarkan kemampuan pembuatan teks yang canggih, setara dengan GPT-4.
- *Turing-NLG*, yang dikembangkan oleh Microsoft, adalah model bahasa canggih yang dioptimalkan untuk berbagai tugas terkait bahasa. Dirancang untuk menghasilkan teks berkualitas tinggi, model ini mampu beradaptasi secara mulus dengan berbagai konteks linguistik dan kebutuhan pengguna. Fleksibilitasnya menjadikan Turing-NLG sebagai alat yang ampuh dalam bidang NLP.

Bidirectional Encoder Representations from Transformers (BERT) adalah model berbasis *transformer* yang dikembangkan oleh Google dan unggul dalam memahami konteks kata berdasarkan teks di sekitarnya. Berbeda dengan model GPT yang berfokus pada pembuatan teks, BERT dirancang terutama untuk pemahaman dan interpretasi teks. Model ini berkinerja sangat baik dalam tugas menjawab pertanyaan, analisis sentimen, dan klasifikasi kalimat. Kotak Abu-abu 1 dan 2 menjelaskan cara kerja struktur BERT serta bagaimana BERT memprediksi kata-kata berikutnya.

KOTAK ABU-ABU 1

Arsitektur BERT

1. Embedding masukan
    - Mengonversi token menjadi vektor berdimensi tinggi.
2. Perhatian multi-kepala
    - Mekanisme perhatian untuk menangkap ketergantungan kata dan konteks.
3. Lapisan umpan maju
    - Memproses embedding dan meningkatkan ekstraksi fitur menggunakan DNN.
4. Lapisan keluaran
    - Menghasilkan representasi kata yang dikontekstualisasikan untuk aplikasi selanjutnya.

KOTAK ABU-ABU 2

Bagaimana BERT memprediksi kata-kata berikutnya

- ***Masked language modeling* (MLM):** Selama pelatihan, BERT mengganti beberapa kata dalam kalimat dengan token khusus [MASK]. Model kemudian memprediksi kata aslinya berdasarkan konteks di sekitarnya. Sebagai contoh, dalam kalimat "Hasil nasional selaras dengan [MASK]," BERT mungkin menyembunyikan kata "polls" (jajak pendapat), sehingga menjadi "Hasil nasional selaras dengan [MASK]." Tugas model adalah memprediksi kata yang disembunyikan tersebut berdasarkan konteks di sekitarnya.
- **Pemahaman kontekstual:** Mekanisme atensi dua arah (*bidirectional attention*) milik BERT memungkinkannya mempertimbangkan konteks dari kedua arah (sisi kiri dan kanan token [MASK]). Pendekatan dua arah ini membantu BERT memprediksi kata yang hilang secara akurat berdasarkan konteks kalimat secara utuh. Dalam contoh tersebut, BERT menggunakan konteks dari frasa "Hasil nasional selaras dengan" untuk memprediksi bahwa kata yang disembunyikan adalah "polls."
- **Penyesuaian halus (*fine-tuning*) untuk tugas spesifik:** BERT dilatih awal menggunakan MLM dan kemudian disesuaikan secara halus untuk tugas-tugas spesifik menggunakan data berlabel. Misalnya, setelah belajar dari contoh seperti "Hasil nasional selaras dengan [MASK]" dan memprediksi kata "polls," BERT dapat disesuaikan untuk tugas-tugas seperti analisis sentimen atau pengenalan entitas bernama (*named entity recognition*), dengan menerapkan pemahaman kontekstualnya pada aplikasi-aplikasi tersebut.

Kemampuan dan Pemanfaatan dalam Pendidikan

LLM menandai kemajuan signifikan dalam bidang AI dan NLP. Model-model ini memanfaatkan jaringan saraf (*neural networks*) untuk memproses kumpulan data teks yang sangat besar, serta mempelajari hubungan yang rumit antara kata, frasa, sintaksis, dan nada bahasa. Melalui pelatihan pada korpus berskala besar, model-model ini mengembangkan kemampuan untuk memahami dan menghasilkan bahasa manusia dengan tingkat akurasi yang tinggi. Seiring dengan paparan terhadap lebih banyak data, kemampuan mereka dalam menangkap konteks dan menghasilkan respons yang koheren terus meningkat, sehingga memperkuat kemahiran mereka dalam memahami dan mereplikasi bahasa alami.

LLM memiliki fleksibilitas tinggi dengan penerapan di berbagai bidang. Model ini mampu menghasilkan teks yang menyerupai tulisan manusia untuk berbagai tugas seperti penerjemahan, penjawaban pertanyaan, peringkasan teks, dan analisis sentimen. Kekuatan utamanya terletak pada kemampuan mengenali nuansa linguistik, yang memungkinkan model ini menangani pemahaman tingkat permukaan maupun interpretasi semantik yang lebih mendalam. Di bidang pendidikan, LLM memiliki potensi yang sangat besar. Model ini dapat menghasilkan contoh tulisan bagi mahasiswa, yang berfungsi sebagai acuan untuk teks yang terstruktur dengan baik dan penggunaan kosakata yang tepat dalam berbagai konteks. Selain itu, LLM dapat meningkatkan pembelajaran yang dipersonalisasi dengan menyediakan umpan balik otomatis, bimbingan belajar adaptif, dan materi pembelajaran yang disesuaikan, yang pada akhirnya meningkatkan pengalaman mengajar maupun belajar.

1.3 POTENSI TRANSFORMATIF AI DALAM PENILAIAN PENDIDIKAN

Kemajuan teknologi sedang mengubah masa depan penilaian pendidikan, dengan solusi berbasis AI yang memelopori transformasi ini. Sistem-sistem ini berpotensi merevolusi pengajaran dan penilaian dengan memungkinkan pengalaman pengajaran yang dipersonalisasi sesuai dengan kebutuhan masing-masing mahasiswa. Kemampuan ini sangat penting dalam ruang kelas yang beragam saat ini, di mana mahasiswa memiliki gaya belajar, kemampuan, dan latar belakang budaya yang berbeda-beda. AI memberdayakan pendidik untuk menerapkan pendekatan yang lebih adaptif dan inklusif, guna memastikan setiap mahasiswa mendapatkan dukungan yang mereka butuhkan untuk meraih keberhasilan.

Dengan menganalisis data secara *real-time*, penilaian berbasis AI dapat mengidentifikasi kekuatan mahasiswa serta aspek-aspek yang perlu ditingkatkan, sehingga memungkinkan pendidik memberikan dukungan yang tepat sasaran dan tepat waktu. Pendekatan ini meningkatkan keterlibatan mahasiswa serta mendorong pembelajaran dan retensi pengetahuan yang lebih mendalam. Selain itu, AI dapat mengatasi bias yang sering ditemukan dalam penilaian tradisional, sehingga mendorong praktik pendidikan yang lebih adil dan setara. Dengan mempersonalisasi penilaian agar sesuai dengan jalur unik setiap pembelajar, teknologi berbasis AI membantu meningkatkan hasil pembelajaran dan menciptakan lingkungan di mana semua mahasiswa dapat berkembang optimal. Dalam lanskap yang terus berkembang ini, para pendidik memiliki kesiapan lebih baik untuk mengambil keputusan yang tepat, serta membentuk masa depan di mana pendidikan menjadi semakin dinamis, responsif, dan inklusif.

Peningkatan Akurasi dan Efisiensi Penilaian

Penilaian berbasis AI memiliki potensi transformatif untuk meningkatkan akurasi dan efisiensi proses evaluasi secara signifikan. Dengan memanfaatkan algoritma canggih dan model pembelajaran mesin (*machine learning*), AI mampu menganalisis data dalam jumlah besar secara *real-time*, serta menghasilkan wawasan akurat mengenai kinerja mahasiswa, pola belajar, dan keterlibatan kognitif. Kemampuan teknologi ini meminimalkan risiko kesalahan manusia (*human error*), sehingga memastikan evaluasi menjadi lebih objektif dan konsisten.

Selain itu, AI mempercepat proses pemberian nilai dan siklus umpan balik, yang memungkinkan pendidik memberikan umpan balik yang tepat waktu, berbasis data, dan dipersonalisasi. Kemampuan ini mendukung pembelajaran individual, mendorong keterlibatan mahasiswa yang lebih mendalam, serta memfasilitasi pertumbuhan berkelanjutan dengan cara memenuhi kebutuhan unik setiap pembelajar secara lebih efektif. Hasilnya, pendidik dapat mengalihkan fokus mereka dari tugas administratif ke strategi instruksional yang meningkatkan keberhasilan mahasiswa.

Di samping itu, sistem AI dirancang agar adaptif, sehingga mampu mempersonalisasi penilaian sesuai dengan kebutuhan unik masing-masing pembelajar. Adaptabilitas ini memungkinkan evaluasi yang lebih disesuaikan, yang melampaui sekadar gambaran sesaat mengenai kemampuan mahasiswa saat ini. Dengan terus memantau dan menganalisis kemajuan pembelajar dari waktu ke waktu, sistem ini memberikan pemahaman yang lebih kaya dan mendalam mengenai kekuatan, aspek yang perlu ditingkatkan, serta pola belajar

mereka. Penilaian berbasis AI tidak hanya berhenti pada pemberian nilai; sistem ini dapat mengungkap tren perilaku belajar mahasiswa, memprediksi kinerja di masa depan, dan bahkan menyarankan intervensi yang tepat sasaran. Intervensi yang disesuaikan ini sangat penting untuk mengidentifikasi dan mendukung mahasiswa yang berisiko tertinggal, guna memastikan tidak ada mahasiswa yang terabaikan. Pada akhirnya, sistem AI berpotensi mendorong hasil pendidikan yang lebih adil dengan menawarkan jalur keberhasilan yang dipersonalisasi bagi setiap pembelajar.

Penguatan Dukungan bagi Mahasiswa dan Pendidik

Penilaian berbasis AI memiliki potensi untuk merevolusi lanskap pendidikan dengan meningkatkan dukungan bagi mahasiswa maupun pendidik. Melalui pemberian umpan balik yang dipersonalisasi, identifikasi kesenjangan pembelajaran, dan penyesuaian instruksi untuk memenuhi kebutuhan individu, AI dapat menciptakan pengalaman pendidikan yang lebih responsif dan efektif. Penilaian berbasis AI menawarkan jalur pembelajaran yang dipersonalisasi bagi mahasiswa, yang menyesuaikan diri secara *real-time* dengan kekuatan dan kelemahan individu mereka. Pendekatan yang disesuaikan ini meningkatkan keterlibatan mahasiswa terhadap materi pelajaran dan menumbuhkan rasa kepemilikan atas perjalanan belajar mereka, yang pada akhirnya menghasilkan capaian akademik yang lebih baik. Selain itu, AI memberikan umpan balik secara langsung, yang memungkinkan mahasiswa untuk segera memahami perkembangan mereka dan mengidentifikasi aspek-aspek yang perlu diperbaiki, melampaui kecepatan metode penilaian tradisional.

Para pendidik juga dapat memperoleh manfaat besar dari kemampuan AI. Dengan mengotomatisasi tugas-tugas administratif rutin seperti penilaian dan pemantauan kemajuan mahasiswa, AI membebaskan waktu berharga bagi pendidik untuk lebih fokus pada kegiatan pengajaran dan pembimbingan. Lebih jauh lagi, AI mampu menganalisis kumpulan data yang luas untuk mengungkap tren dan wawasan yang mungkin tidak terlihat secara kasatmata. Wawasan yang dapat ditindaklanjuti ini dapat memandu strategi instruksional, sehingga memungkinkan pendidik untuk menyempurnakan metode pengajaran dan memberikan dukungan yang lebih baik kepada mahasiswa mereka. Dengan demikian, penilaian berbasis AI meningkatkan pengalaman belajar individu sekaligus menciptakan lingkungan belajar yang lebih efisien dan efektif. Melalui penyelarasan antara kebutuhan mahasiswa dan kemampuan pendidik, AI mendorong ekosistem pendidikan yang lebih inklusif dan adaptif, serta memberdayakan seluruh peserta didik untuk berkembang secara optimal.

1.4 ETIKA, KEADILAN, DAN TANTANGAN IMPLEMENTASI AI

Permasalahan terkait etika dan keadilan mencakup aspek bias, keamanan data, dan privasi. Isu keadilan muncul ketika algoritma dan sistem AI secara tidak sengaja lebih menguntungkan kelompok tertentu, yang berpotensi melanggengkan ketimpangan yang sudah ada atau menciptakan bentuk diskriminasi baru. Bias dapat muncul melalui berbagai cara, seperti penggunaan data pelatihan yang bias atau adanya asumsi yang keliru dalam perancangan model; hal ini dapat menghasilkan keluaran yang tidak seimbang dan berdampak secara tidak proporsional terhadap kelompok masyarakat yang terpinggirkan.

Keamanan dan privasi data juga merupakan komponen dalam diskusi etis seputar AI. Mengingat teknologi ini sering kali bergantung pada data pribadi dalam jumlah besar, melindungi informasi tersebut dari akses tidak sah dan kebocoran data menjadi hal yang sangat penting. Selain itu, individu harus memiliki transparansi yang jelas mengenai bagaimana data mereka dikumpulkan, digunakan, dan disimpan. Menangani isu-isu etika dan keadilan ini sangatlah mendasar untuk membangun kepercayaan terhadap sistem AI serta menjamin bahwa sistem tersebut beroperasi dengan cara yang menghormati hak dan martabat seluruh pengguna.

Kekhawatiran terkait Keadilan dan Bias

Keadilan dan bias dalam pembelajaran serta penilaian berbasis AI merupakan isu krusial yang memerlukan perhatian serius karena dapat memengaruhi hasil dan efektivitas sistem tersebut. Bias dapat muncul dari berbagai sumber termasuk data, algoritma, dan keterlibatan manusia yang masing-masing berkontribusi secara berbeda terhadap perilaku dan pengambilan keputusan model AI. Bias data merupakan salah satu tantangan dalam AI yang sering kali terabaikan, yang berujung pada keluaran (*output*) yang tidak dapat diandalkan kondisi yang umum digambarkan dengan istilah "*garbage in, garbage out*" (sampah yang masuk menghasilkan sampah yang keluar). Karena sistem AI bergantung pada data untuk belajar dan membuat prediksi, bias yang terdapat dalam data pelatihan dapat terbawa ke dalam sistem. Sebagai contoh, jika data historis mencerminkan bias sosial terkait gender, ras, atau status sosial-ekonomi, sistem AI tidak hanya dapat mencerminkan bias tersebut tetapi juga memperkuatnya. Dalam lingkungan pendidikan, hal ini dapat menyebabkan penilaian yang tidak adil, di mana kelompok mahasiswa tertentu secara sistematis dirugikan akibat data yang bias.

Bias algoritma adalah kekhawatiran krusial lainnya yang bersumber dari desain algoritma AI. Meskipun data yang digunakan relatif bebas bias, algoritma tetap dapat memproses informasi dengan cara yang secara tidak proporsional menguntungkan atau merugikan kelompok tertentu. Bias ini dapat terjadi ketika fitur atau variabel tertentu dalam proses pengambilan keputusan algoritma merepresentasikan satu kelompok dengan lebih baik dibandingkan kelompok lain sehingga menghasilkan keluaran yang timpang. Dalam penilaian pendidikan, hal ini dapat mengakibatkan pemberian skor yang tidak adil, di mana sebagian mahasiswa dinilai lebih tinggi atau lebih rendah berdasarkan faktor-faktor yang tidak berkaitan dengan kinerja mereka yang sebenarnya. Pada ujian yang memiliki konsekuensi besar (*high-stakes tests*), seperti ujian lisensi profesi, hal ini bahkan dapat menyebabkan individu kehilangan kredensial mereka.

Bias keputusan manusia juga memainkan peran penting. Individu yang mengembangkan, menerapkan, dan mengelola sistem AI dapat secara tidak sengaja memasukkan bias mereka sendiri, baik secara sadar maupun tidak sadar. Bias ini dapat muncul pada berbagai tahapan pengembangan AI, seperti pelabelan data, pemilihan model, atau penafsiran hasil. Keputusan manusia tersebut memengaruhi kinerja sistem, dan jika terdapat bias, hal itu akan tercermin dalam penilaian AI. Sebagai contoh, seorang pengembang mungkin

secara tidak sengaja memilih fitur yang menguntungkan kelompok mahasiswa tertentu, sehingga memperparah ketimpangan yang sudah ada.

Menangani aspek keadilan dan bias dalam AI menuntut perhatian berkelanjutan terhadap sumber-sumber bias tersebut di sepanjang siklus hidup AI, mulai dari pengumpulan data dan pengembangan algoritma hingga pengawasan serta penafsiran oleh manusia. Bias dalam penilaian berbasis AI memiliki konsekuensi mendalam, yang mengarah pada hasil yang tidak adil dan memperparah ketimpangan pendidikan yang sudah ada. Ketika sistem AI menunjukkan bias, hasil penilaian dapat secara tidak proporsional merugikan kelompok mahasiswa tertentu. Sebagai contoh, sistem AI yang dirancang untuk mengevaluasi esai mahasiswa mungkin secara konsisten memberikan nilai lebih rendah kepada mahasiswa dari latar belakang tertentu, bahkan ketika kualitas karya mereka setara dengan mahasiswa lain. Bias menciptakan evaluasi akademik yang timpang, di mana sebagian mahasiswa dirugikan secara tidak adil bukan karena kualitas karya mereka, melainkan akibat bias yang tertanam dalam sistem tersebut.

Di luar hasil penilaian individu, bias ini memiliki implikasi yang lebih luas. Jika tidak ditangani, bias tersebut dapat memperkuat ketimpangan pendidikan yang ada, sehingga menyulitkan kelompok marginal untuk meraih keberhasilan. Ketika sistem AI secara konsisten merugikan mahasiswa tertentu, sistem ini turut memperparah hambatan sistemik yang sudah ada, serta memperlebar kesenjangan antara kelompok yang memiliki privilese dan kelompok yang kurang terwakili. Siklus bias ini, begitu tertanam dalam AI, cenderung melanggengkan dirinya sendiri dan menghambat upaya untuk mendorong kesetaraan serta inklusivitas pendidikan bagi populasi mahasiswa yang beragam.

Transparansi, kemampuan untuk direproduksi (*reproducibility*), dan kesadaran akan bias merupakan strategi penting untuk memitigasi risiko yang terkait dengan sistem AI, khususnya dalam bidang-bidang sensitif seperti penilaian dan pengambilan keputusan. Transparansi menuntut sistem AI untuk mengungkapkan proses pengambilan keputusannya serta memberikan penjelasan yang jelas dan mudah dipahami mengenai hasil yang dihasilkannya. Pendekatan ini mencakup penjelasan mengenai logika di balik pemberian nilai, penilaian, atau rekomendasi, yang pada gilirannya membangun kepercayaan pengguna dan menciptakan landasan bagi akuntabilitas.

Kemampuan untuk direproduksi sangatlah vital bagi sistem AI, karena menekankan perlunya hasil yang konsisten dalam kondisi yang sama. Ketika sistem AI menunjukkan kemampuan untuk direproduksi, hal ini meningkatkan kepercayaan terhadap keandalan dan keadilannya, memperkuat kredibilitasnya, serta memungkinkan pengujian dan validasi yang efektif. Hal yang tak kalah pentingnya adalah kesadaran akan bias, yang menuntut pengembang untuk secara cermat mendeteksi dan mengatasi bias dalam sistem AI. Pendekatan ini melibatkan penggunaan data pelatihan yang beragam dan representatif, serta pelaksanaan audit secara berkala untuk mendeteksi dan memitigasi bias yang mungkin muncul seiring berjalannya waktu. Dengan menyadari adanya bias dan berupaya mengatasinya, pengembang dapat membantu memastikan bahwa sistem AI mengambil keputusan yang adil dan setara, sehingga terhindar dari diskriminasi atau dampak merugikan yang tidak disengaja.

Kekhawatiran terkait Privasi dan Keamanan Data

Privasi dan keamanan data merupakan hal yang sangat penting dalam sistem AI, terutama karena besarnya volume data yang ditangani serta risiko yang menyertainya. Mulai dari tahap awal pengumpulan data hingga pelatihan model AI dan pada akhirnya penggunaan model tersebut oleh pengguna akhir, kekhawatiran mengenai privasi melingkupi setiap langkah prosesnya.

Pengumpulan data, yang sering kali memuat informasi pribadi yang sensitif, menghadirkan tantangan besar, khususnya terkait perolehan persetujuan dan pemastian penggunaan data yang etis. Organisasi harus menghadapi kompleksitas dalam memperoleh kumpulan data berskala besar sembari tetap menjunjung tinggi standar etika. Seiring meningkatnya volume data yang diproses oleh sistem AI, ancaman kebocoran data pun semakin besar. Akses tidak sah terhadap informasi pribadi dapat menimbulkan konsekuensi serius, termasuk pencurian identitas dan penipuan yang menyasar mahasiswa, sehingga menyoroti betapa krusialnya prioritas keamanan data bagi pengembang dan organisasi. Selain itu, kemampuan AI untuk menganalisis dan melakukan referensi silang informasi dari berbagai sumber memunculkan masalah lain: risiko identifikasi ulang individu dari kumpulan data yang seharusnya telah dianonimkan.

Sistem AI juga meningkatkan kemampuan pengawasan, yang berpotensi melanggar hak privasi individu. Lebih jauh lagi, bias yang tertanam dalam algoritma AI dapat menghasilkan dampak diskriminatif, yang merugikan individu secara tidak adil berdasarkan karakteristik seperti ras atau jenis kelamin. Selama tahap pelatihan model AI, terdapat risiko kebocoran data, di mana informasi sensitif dapat terekspos secara tidak sengaja, sehingga memperumit masalah keamanan data. Seiring terus berkembangnya teknologi AI, kerangka regulasi yang mengatur penerapannya pun turut berkembang. Organisasi harus menghadapi tantangan berkelanjutan dalam memastikan kepatuhan terhadap hukum dan peraturan privasi, yang senantiasa menyesuaikan diri dengan perkembangan baru di bidang ini. Untuk mengatasi masalah-masalah kompleks ini secara efektif, organisasi harus menerapkan kerangka tata kelola data yang tangguh, mengadopsi praktik AI yang transparan, serta melakukan pemantauan berkelanjutan demi melindungi privasi individu dan menjamin keamanan data.

Tantangan Implementasi

Tantangan implementasi mencakup beberapa faktor penting. Kualitas data sangatlah krusial, karena ketidakakuratan dapat berujung pada pengambilan keputusan yang buruk dan sistem yang tidak efektif. Selain itu, infrastruktur dan teknologi yang tangguh sangat diperlukan untuk mendukung sistem baru, termasuk ketersediaan perangkat keras, perangkat lunak, dan kemampuan jaringan yang memadai.

Tantangan besar lainnya adalah kurangnya tenaga profesional yang terampil. Kesenjangan ini sering kali menyebabkan penundaan dalam implementasi dan pemeliharaan sistem, sehingga menyoroti kebutuhan mendesak akan strategi pelatihan dan perekrutan yang tepat sasaran. Terakhir, biaya awal yang tinggi serta ketidakpastian mengenai pengembalian investasi (ROI) dapat menghambat upaya untuk membenarkan pengeluaran kepada para

pemangku kepentingan. Dengan mengatasi masalah-masalah ini, organisasi dapat meningkatkan peluang keberhasilan implementasi dan memanfaatkan sepenuhnya manfaat dari inisiatif mereka.

Persyaratan Teknis dan Infrastruktur

Membangun dan memelihara sistem AI memerlukan infrastruktur tangguh yang mampu mengintegrasikan berbagai komponen perangkat keras dan perangkat lunak secara mulus. Komponen utama dalam sistem ini adalah penyimpanan data berskala besar, mengingat AI bergantung pada volume data yang masif agar dapat beroperasi secara optimal. Untuk mengakomodasi hal tersebut, solusi penyimpanan yang dapat ditingkatkan skalanya (*scalable*), seperti penyimpanan berbasis cloud dan basis data terdistribusi, menjadi sangat penting. Setelah data tersimpan, data tersebut harus diproses secara efisien. Perangkat lunak seperti Apache Hadoop dan Apache Spark berperan penting dalam menangani kumpulan data (*dataset*) yang besar, sehingga memungkinkan informasi diproses dan dipembelajaran secara efektif.

Pelatihan model AI membutuhkan sumber daya komputasi yang besar, sehingga penggunaan perangkat keras khusus menjadi suatu keharusan. Prosesor khusus seperti GPU dan TPU sangat krusial dalam menangani komputasi intensif yang terlibat dalam proses ini. Setelah sumber daya komputasi yang memadai tersedia, pengembang menggunakan kerangka kerja (*framework*) ML untuk membangun dan melatih model AI. Kerangka kerja seperti *TensorFlow* dan *PyTorch* menyediakan perangkat dan pustaka (*library*) yang diperlukan untuk memfasilitasi serta memperlancar proses pembuatan model.

Platform operasi ML (MLOps) sangat penting untuk memastikan kelancaran operasional model ML di sepanjang siklus hidupnya. Perangkat seperti MLflow dan Kubeflow memfasilitasi pengelolaan seluruh tahapan mulai dari pengembangan hingga penerapan (*deployment*) dan pemantauan serta menyediakan kerangka kerja komprehensif untuk mengawasi proyek AI. Jaringan merupakan elemen fundamental dalam infrastruktur ini. Jaringan yang cepat dan andal sangat diperlukan untuk mentransfer kumpulan data besar serta parameter model di antara berbagai komponen sistem AI. Platform cloud, termasuk *Amazon Web Services* (AWS), Google Cloud, dan Azure, menawarkan solusi jaringan yang dapat disesuaikan skalanya (*scalable*) untuk beban kerja AI, guna memastikan aliran data dan konektivitas yang efisien.

Terakhir, perangkat pengembangan dan penerapan (*deployment*) sangatlah penting untuk membuat dan mengimplementasikan model AI. Lingkungan pengembangan terintegrasi (*Integrated Development Environments*) seperti *Jupyter Notebooks* dan *PyCharm* umum digunakan untuk pengembangan model. Sementara itu, perangkat penerapan seperti Docker dan Kubernetes memfasilitasi peluncuran model-model tersebut secara efisien dan dapat disesuaikan skalanya. Secara keseluruhan, komponen-komponen ini membentuk infrastruktur AI yang tangguh, yang mendukung kelancaran pengembangan, penerapan, dan pemeliharaan sistem AI tingkat lanjut.

Pelatihan dan Dukungan

Pelatihan dan dukungan membantu mengatasi berbagai tantangan yang sering muncul saat mengimplementasikan sistem AI. Salah satu pendekatan praktisnya adalah dengan mengadakan lokakarya dan seminar secara berkala. Sesi-sesi ini menjaga tim agar tetap mengikuti perkembangan tren, perangkat, dan teknologi AI terkini, sekaligus menumbuhkan budaya pembelajaran berkelanjutan dan inovasi. Kegiatan ini juga menciptakan lingkungan kolaboratif tempat anggota tim dapat bertukar gagasan, mendiskusikan tantangan, dan mengembangkan keterampilan baru yang relevan dengan peran mereka.

Selain lokakarya, anggota tim didorong untuk mengikuti kursus daring dan mengambil sertifikasi di bidang AI dan ML guna memperdalam pemahaman serta keahlian mereka. Pembelajaran terstruktur yang dapat dilakukan sesuai kecepatan masing-masing ini memberikan fleksibilitas dan memungkinkan karyawan mengembangkan pengetahuan khusus, sehingga mereka dapat menjalankan peran mereka secara lebih efektif. Sertifikasi juga dapat memperkuat kredensial profesional mereka, sekaligus meningkatkan kepercayaan diri dan nilai tambah mereka di dalam organisasi.

Program bimbingan adalah strategi ampuh lainnya untuk mengatasi tantangan implementasi. Dengan memasangkan anggota tim yang kurang berpengalaman dengan pakar AI, organisasi dapat mendorong pembelajaran langsung dan panduan yang dipersonalisasi. Pendekatan satu lawan satu ini memungkinkan peserta didik untuk mengajukan pertanyaan yang ditargetkan, menerima saran praktis, dan menerapkan pengetahuan mereka secara langsung ke proyek dunia nyata. Pada saat yang sama, mentor memiliki kesempatan untuk menyempurnakan keterampilan mereka melalui pengajaran, memupuk budaya pertumbuhan dan perkembangan bersama di seluruh tim.

Selain itu, dukungan teknis yang berkelanjutan sangat penting untuk kelancaran dan efisiensi pengoperasian sistem AI. Membangun infrastruktur pendukung yang solid memungkinkan tim memecahkan masalah teknis dengan cepat dan mengoptimalkan kinerja aplikasi AI. Baik untuk mengatasi bug sistem, menyempurnakan algoritme, atau meningkatkan integrasi dengan sistem yang ada, dukungan teknis yang andal meminimalkan gangguan dan membuat inisiatif AI terus bergerak maju. Dengan menerapkan strategi-strategi ini, pendidikan berkelanjutan melalui lokakarya dan kursus, bimbingan untuk pengembangan keterampilan, dan dukungan teknis khusus, organisasi tidak hanya dapat mengatasi tantangan penerapan AI tetapi juga menciptakan landasan bagi kesuksesan dan inovasi jangka panjang dalam penerapan AI.

1.5 AI DALAM TRANSFORMASI PENDIDIKAN DAN PENILAIAN

AI telah mengubah berbagai bidang dengan cepat, termasuk pendidikan dan penilaian, melampaui kemampuan teknologi tradisional. Dalam pendidikan, AI telah merevolusi metode konvensional serta meningkatkan proses pengajaran dan pembelajaran dengan memungkinkan instruksi yang dipersonalisasi, yang disesuaikan dengan gaya dan kecepatan belajar masing-masing individu. Hasilnya, pendidikan menjadi lebih mudah diakses dan praktis. Di ranah penilaian, inovasi seperti pembuatan butir soal otomatis (*automatic item*

generation) dan rekayasa *prompt* (*prompt engineering*) telah menyederhanakan proses pembuatan soal tes. Pada saat yang sama, penilaian otomatis dan umpan balik instan telah meningkatkan efisiensi, objektivitas, dan ketepatan waktu dalam penilaian. Kemajuan ini telah meningkatkan akurasi penilaian dan memungkinkan pendidik untuk lebih fokus pada pengajaran yang dipersonalisasi, sehingga memperbaiki pengalaman belajar.

Meskipun terdapat peningkatan signifikan ini, potensi *transformatif* AI dalam pendidikan belum sepenuhnya terwujud. Masih terdapat peluang besar untuk memperkenalkan metode penilaian yang lebih inovatif guna memenuhi beragam kebutuhan belajar dan meningkatkan hasil pendidikan. Sebagai contoh, penilaian adaptif yang menyesuaikan diri secara *real-time* berdasarkan respons mahasiswa serta analitik prediktif berbasis AI yang menawarkan wawasan lebih mendalam mengenai kinerja mahasiswa merupakan bidang yang sangat potensial untuk dikembangkan.

Namun, potensi ini juga membawa sejumlah tantangan etis dan praktis. Aspek keadilan dan bias dalam algoritma AI menjadi hal yang krusial, terutama mengingat beragamnya populasi mahasiswa yang dilayani oleh sistem pendidikan. Kekhawatiran mengenai transparansi keputusan berbasis AI sangatlah beralasan, khususnya ketika sistem tersebut berisiko memperparah ketimpangan yang sudah ada secara tidak sengaja. Masalah privasi dan keamanan data menjadi prioritas utama, mengingat AI sering kali bergantung pada sejumlah besar data mahasiswa yang bersifat sensitif. Tanpa langkah perlindungan yang memadai, mahasiswa dan pendidik berisiko menghadapi bahaya yang sebenarnya dapat dihindari.

Penerapan AI dalam pendidikan juga menghadapi beberapa tantangan logistik. Pemeliharaan data berkualitas tinggi sangat penting untuk kinerja AI yang optimal; namun, banyak institusi pendidikan menghadapi kesulitan dalam mempertahankan standar tersebut. Membangun infrastruktur yang diperlukan untuk mendukung teknologi AI bisa memakan biaya besar dan waktu yang lama, serta membutuhkan sistem tangguh yang mampu menangani kompleksitas algoritma AI. Selain itu, terbatasnya keahlian di bidang AI dalam institusi pendidikan menjadi hambatan signifikan, di mana banyak pendidik dan administrator merasa ragu mengenai cara mengintegrasikan teknologi ini secara efektif ke dalam kerangka kerja yang sudah ada. Lebih jauh lagi, menentukan *Return on Investment* (ROI) atau pengembalian investasi untuk AI dalam pendidikan tetap menjadi tantangan tersendiri, mengingat manfaatnya sering kali tidak dapat dirasakan secara langsung maupun diukur dengan mudah.

Meskipun tantangan-tantangan ini cukup besar, hal tersebut bukanlah sesuatu yang mustahil untuk diatasi. Dengan pelatihan dan dukungan yang memadai, serta komitmen untuk mendorong pembelajaran berkelanjutan, organisasi dapat berhasil mengelola kompleksitas integrasi AI. Memberdayakan pendidik dengan keahlian untuk memanfaatkan perangkat AI secara efektif merupakan hal yang krusial. Selain itu, seiring berkembangnya teknologi AI, organisasi harus terus mengembangkan dan menyempurnakan kebijakan serta pedoman yang menangani masalah etika, menjamin privasi data, dan memitigasi risiko terkait keadilan serta bias. Dengan demikian, dunia pendidikan dapat lebih optimal dalam memanfaatkan potensi

AI, sehingga membuka jalan bagi pendekatan pengajaran dan penilaian yang lebih personal, adil, dan inovatif.

BAB 2
ETIKA DAN TANTANGAN PENILAIAN PENDIDIKAN BERBASIS AI

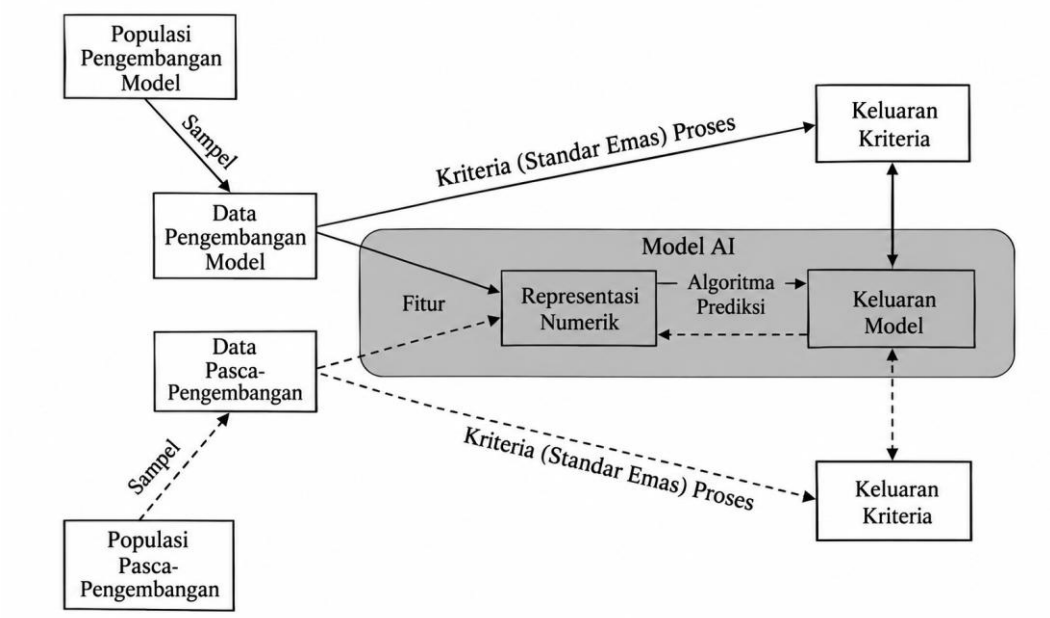
2.1 DASAR ETIKA PENILAIAN BERBASIS AI

Perkembangan AI dalam Penilaian Pendidikan

Perkembangan pesat teknologi kecerdasan buatan (AI) dalam konteks pendidikan sedang mengubah lanskap penilaian dengan kecepatan yang belum pernah terjadi sebelumnya. Menurut Kelompok Ahli Tingkat Tinggi (*High-Level Expert Group*) Komisi Eropa, sistem AI adalah "sistem yang menunjukkan perilaku cerdas dengan pembahasan lingkungannya dan mengambil Tindakan dengan tingkat otonomi tertentu untuk mencapai tujuan-tujuan spesifik". Dalam penilaian pendidikan, AI mencakup berbagai aplikasi yang luas: mulai dari penilaian otomatis dan tes adaptif hingga sistem tutor cerdas dan pembuatan konten penilaian yang dipersonalisasi.

Inovasi-inovasi ini meningkatkan efisiensi, memungkinkan umpan balik waktu nyata (*real-time*), serta memberikan skalabilitas dan fleksibilitas dalam mengukur keterampilan dan pengetahuan peserta didik. Sebagai contoh, AI generatif kini mendukung pembuatan butir soal yang baru dan adaptif, sementara algoritma canggih memungkinkan pembelajaran data multimodal yang menggabungkan audio, video, dan teks untuk menilai konstruk kompleks seperti komunikasi atau kolaborasi. Pada saat yang sama, analitik berbasis AI dapat memberikan wawasan kepada pendidik, peserta didik, dan pembuat kebijakan, yang mendorong pembelajaran berbasis data serta mendukung mobilitas pendidikan.

Seiring sistem-sistem ini semakin terintegrasi secara mendalam ke dalam infrastruktur penilaian, penting untuk menyadari bahwa solusi penilaian berbasis AI bukan sekadar model tunggal, melainkan sebuah proses kompleks yang terdiri dari berbagai komponen. Gambar 2.1 mengilustrasikan arsitektur tipikal dari sistem penilaian berbasis AI. Gambar tersebut menyoroti bagaimana pengembangan dimulai dengan pengambilan sampel populasi dan data yang *representatif*, diikuti oleh transformasi data tersebut menjadi fitur numerik yang sesuai untuk pemodelan dan prediksi. Yang terpenting, setiap prediksi yang dihasilkan oleh sistem berbasis AI ditafsirkan dengan mengacu pada "kriteria" atau proses standar emas seperti skor penilai manusia atau tolok ukur ahli yang menjadi landasan validitas sistem. Setelah diterapkan, sistem terus memproses data baru dari beragam populasi, sehingga memerlukan evaluasi berkelanjutan terhadap prediksi berdasarkan kriteria yang diperbarui guna memantau akurasi maupun keadilan.



Gambar 2.1 Representasi skematis dari sistem penilaian berbasis AI yang umum, yang mencakup tahap pengembangan model maupun pasca-penerapan, serta melibatkan aspek data, populasi, representasi numerik, dan kriteria evaluasi standar emas. Setiap tahapan mulai dari pengambilan sampel dan pembentukan fitur hingga keluaran model dan proses kriteria dapat menimbulkan kesalahan atau bias dalam proses penilaian.

Perspektif tingkat sistem ini sangatlah krusial: kerentanan etis termasuk bias, ancaman terhadap reliabilitas, dan ketidakadilan dapat muncul tidak hanya dari algoritma pemodelan itu sendiri, tetapi juga dari pengaruh hulu (seperti pengambilan sampel data dan pemilihan fitur) atau kriteria hilir dan praktik interpretasi. Dengan demikian, inferensi berisiko tinggi yang dihasilkan oleh AI dapat memengaruhi akses terhadap peluang pendidikan, kemajuan karier, dan lintasan pembelajaran sepanjang hayat. Oleh karena itu, penerapan AI dalam penilaian pendidikan harus dilakukan dengan cara yang tidak hanya tangguh secara teknis, tetapi juga bertanggung jawab secara etis dan adil secara sosial.

Pentingnya Etika dalam Penilaian AI

Tantangan etis menyertai setiap tahapan siklus hidup penilaian berbasis AI, mulai dari pengumpulan data dan pengembangan model hingga penerapan serta pemantauan sistem. Bias algoritmik, kurangnya transparansi, ketidakjelasan dalam pengambilan keputusan, dan risiko penyalahgunaan dapat melemahkan tujuan mendasar dari penilaian pendidikan: mendukung pembelajaran, menjamin kesetaraan, dan melindungi hak-hak individu. Bias dalam sistem AI dapat muncul akibat data yang tidak representatif, asumsi pemodelan yang keliru, atau interaksi dengan konteks sosioteknis yang kompleks. Bias semacam itu dapat melanggengkan atau bahkan memperparah ketimpangan historis, sehingga secara tidak sengaja merugikan kelompok demografis tertentu atau individu dengan kebutuhan belajar yang khas. Hal ini memunculkan pertanyaan krusial mengenai keadilan bukan sekadar sebagai properti statistik, melainkan sebagai realitas nyata yang dialami oleh peserta didik saat menjalani sistem pendidikan yang dibentuk oleh AI.

Selain itu, sifat tidak transparan dari banyak sistem berbasis AI yang mutakhir yang sering disebut sebagai masalah *black box* (kotak hitam) mengurangi kemampuan untuk menjelaskan dan menafsirkan sistem, padahal hal ini diperlukan agar pengawasan oleh manusia dapat berjalan secara bermakna. Para pemangku kepentingan, termasuk peserta didik dan pendidik, mungkin tidak dapat menyanggah, memahami, atau memercayai hasil keluaran sistem tanpa adanya penjelasan yang jelas serta mekanisme akuntabilitas yang kokoh. Privasi dan keamanan data tetap menjadi perhatian utama, terutama karena sistem penilaian kini mengumpulkan dan memproses data peserta didik yang semakin rinci dan sensitif.

Berbagai lembaga internasional dan nasional termasuk UNESCO, OECD, NIST, Departemen Pendidikan AS, serta organisasi profesi di bidang pengukuran seperti *American Educational Research Association, American Psychological Association,* dan *National Council on Measurement in Education* telah merumuskan serangkaian prinsip yang selaras mengenai tata kelola etis AI dalam pendidikan. Prinsip-prinsip tersebut menekankan pada keadilan dan mitigasi bias, transparansi, kemampuan untuk memberikan penjelasan (*explainability*), pengawasan manusia, otonomi pribadi, akuntabilitas, privasi dan keamanan, serta pemantauan berkelanjutan demi perbaikan yang terus-menerus. Oleh karena itu, menciptakan keseimbangan antara potensi inovatif yang bersifat disruptif dari AI dan tanggung jawab etis yang melekat dalam pengukuran pendidikan merupakan tantangan utama bagi para peneliti, pengembang, maupun praktisi.

Ruang Lingkup Pembahasan Bab

Bab ini mengkaji secara kritis tantangan dan tanggung jawab etis yang terkait dengan integrasi AI dalam penilaian pendidikan. Dengan mengacu pada standar internasional dan penelitian terkini, tujuan kami adalah menyediakan kerangka kerja yang jelas serta strategi praktis bagi para praktisi, peneliti, dan pembuat kebijakan untuk penggunaan AI secara bertanggung jawab. Kami mengawalinya dengan memperjelas konsep-konsep dasar termasuk etika, keadilan, dan kepatuhan serta memperkenalkan pendekatan filosofis yang mendasari pengambilan keputusan etis dalam penilaian. Bagian inti bab ini membahas isu-isu etis utama:

- Bias algoritmik dan dampaknya terhadap keadilan, disertai pembahasan mengenai mekanisme serta strategi mitigasinya.
- Transparansi dan kemampuan untuk memberikan penjelasan (*explainability*) sebagai syarat mutlak bagi kepercayaan pemangku kepentingan dan akuntabilitas.
- Pengawasan manusia dan risiko bias otomatisasi, dengan menekankan pentingnya pertimbangan profesional serta mekanisme perlindungan kelembagaan.
- Isu privasi dan keamanan, yang berfokus pada perlindungan data sensitif peserta didik.

Untuk mengilustrasikan tema-tema tersebut, kami menyajikan studi kasus mendalam mengenai platform penilaian berbasis AI, serta pembahasan kekuatan dan keterbatasannya dari sudut pandang etika. Kami menutupnya dengan rekomendasi praktis terkait pemantauan berkelanjutan, pelibatan pemangku kepentingan, dan pengembangan kebijakan, sembari menegaskan bahwa inovasi etis harus menyeimbangkan kemajuan teknis dengan komitmen terhadap kesetaraan dan kemaslahatan umum.

2.2 KERANGKA ETIKA PENILAIAN BERBASIS AI

Istilah "etis" secara luas mengacu pada penerapan prinsip dan standar moral, pembedaan antara yang benar dan yang salah, serta tindakan yang selaras dengan kode etik dan nilai-nilai yang berlaku. Dalam praktik profesional, etika menjadi pedoman pengambilan keputusan, yang mengarahkan individu dan organisasi untuk bertindak dengan integritas, keadilan, dan rasa hormat terhadap seluruh pemangku kepentingan. Sistem penilaian yang diperkuat oleh AI mendorong perhatian baru terhadap prinsip-prinsip etis ini.

Etika, Keadilan, dan Kepatuhan

Dalam diskusi mengenai penilaian pendidikan berbasis AI, istilah etika, keadilan, dan kepatuhan terkadang dianggap dapat saling menggantikan. Namun, meskipun konsep-konsep ini saling berkaitan erat, masing-masing memiliki makna dan implikasi yang berbeda terhadap penelitian, pengembangan, dan penerapan sistem penilaian. Etika mencakup prinsip-prinsip luas mengenai apa yang benar, adil, dan bermanfaat bagi individu maupun masyarakat. Pertimbangan etis menuntut refleksi atas nilai-nilai moral, kemungkinan konsekuensi dari suatu keputusan, serta kewajiban terhadap berbagai pemangku kepentingan termasuk peserta didik, pendidik, keluarga, dan masyarakat luas. Dalam penilaian pendidikan, etika tidak hanya menyangkut akurasi teknis, tetapi juga kewajiban moral untuk menghormati dan memberdayakan peserta didik, mendorong kesetaraan, serta memajukan hasil pendidikan dan sosial yang positif.

Keadilan merupakan komponen sentral namun lebih spesifik dari etika. Dalam konteks pengukuran pendidikan, keadilan secara khusus mengacu pada upaya memastikan bahwa penilaian tidak menciptakan atau memperkuat keuntungan maupun kerugian yang tidak adil bagi kelompok atau individu mana pun. Hal ini mencakup:

- Menghindari bias berdasarkan ras, gender, status sosial-ekonomi, disabilitas, bahasa, atau atribut lainnya;
- Menyediakan akses dan kesempatan yang setara;
- Memperlakukan kasus yang serupa dengan cara yang sama dan kasus yang berbeda dengan cara yang berbeda secara tepat;
- Mempertimbangkan secara cermat bagaimana sistem penilaian dapat memberikan dampak yang berbeda terhadap populasi yang beragam.

Keadilan sering kali dioperasionalkan melalui kriteria yang dapat diukur dalam pengujian dan pembelajaran mesin (*machine learning*), misalnya dengan mengevaluasi hasil yang berbeda (*differential outcomes*), dampak yang timpang (*disparate impact*), atau fungsi butir soal yang berbeda (*differential item functioning*). Meskipun demikian, metrik teknis semata tidak dapat menggantikan refleksi etis yang berkelanjutan, karena definisi dan harapan mengenai keadilan dapat bervariasi di berbagai masyarakat dan konteks.

Sebaliknya, kepatuhan mengacu pada ketaatan terhadap hukum, peraturan, dan kebijakan institusional yang telah ditetapkan. Hal ini dapat mencakup undang-undang privasi data (seperti FERPA atau GDPR), hukum anti-diskriminasi, standar pengujian profesional, atau prosedur organisasi khusus terkait keamanan dan integritas tes. Meskipun kepatuhan

memberikan batasan pengaman yang penting serta standar minimum yang jelas bagi praktik, hal tersebut tidak dengan sendirinya menjamin hasil yang etis atau adil. Bisa saja terjadi situasi di mana sistem yang telah memenuhi ketentuan tetap menghasilkan dampak yang secara etis dipertanyakan atau tidak adil dalam praktiknya.

Walaupun perbedaan antara etika, keadilan, dan kepatuhan memberikan landasan praktis bagi penggunaan AI yang bertanggung jawab dalam penilaian pendidikan, memahami cara menghadapi dilema etis yang kompleks sering kali memerlukan refleksi yang lebih mendalam. Dalam hal ini, tradisi filosofis menawarkan sumber daya yang sangat berharga. Dengan mengkaji berbagai aliran pemikiran etis, kita dapat lebih memahami nilai-nilai, prioritas, dan pertukaran kepentingan (*trade-offs*) yang membentuk penilaian kita, terutama ketika metrik teknis atau standar regulasi tidak memadai untuk menyelesaikan kasus-kasus yang kompleks. Dalam konteks penilaian berbasis AI, perspektif filosofis ini tidak hanya menjadi acuan dalam mendefinisikan dan mengupayakan keadilan, tetapi juga membantu kita menyadari kewajiban kita yang lebih luas terhadap peserta didik dan masyarakat. Bagian berikut ini memperkenalkan pendekatan-pendekatan filosofis utama yang dapat memperkaya pengambilan keputusan etis dalam lanskap yang terus berkembang ini.

Pendekatan Filosofis dalam Etika

Berbagai tradisi filosofis telah membentuk penalaran etis, di mana masing-masing menawarkan perspektif berbeda mengenai makna bertindak secara moral atau mengambil keputusan etis. Memahami teori-teori ini dapat membantu memperjelas cara kita menyikapi persoalan etis dalam penilaian pendidikan, khususnya ketika muncul dilema baru akibat penggunaan AI dan perangkat algoritmik.

- **Konsekuensialisme (*utilitarianisme*)**: Tradisi ini menilai moralitas keputusan berdasarkan hasil akhirnya, dengan mempertanyakan apakah suatu sistem AI mampu meningkatkan capaian pendidikan bagi jumlah peserta didik terbanyak. Meskipun seorang penganut konsekuensialisme mungkin mendukung algoritma yang mampu meningkatkan prestasi secara keseluruhan, pendekatan ini berisiko mengabaikan dampak buruk terhadap kelompok minoritas atau kelompok yang terpinggirkan, sehingga menuntut pertimbangan cermat mengenai siapa yang diuntungkan dan siapa yang mungkin dirugikan.
- ***Deontologi* (etika berbasis kewajiban)**: Etika deontologis berfokus pada prinsip dan hak yang tidak boleh dilanggar, terlepas dari apa pun konsekuensinya. Dalam pengukuran pendidikan, hal ini berarti menjunjung tinggi kewajiban seperti sikap nondiskriminatif dan prinsip persetujuan berdasarkan informasi yang memadai (*informed consent*). Sebagai contoh, model penilaian yang secara sistematis merugikan kelompok tertentu dianggap tidak etis dari perspektif deontologis, meskipun model tersebut mampu meningkatkan kinerja rata-rata.
- **Etika kebajikan (*virtue ethics*)**: Alih-alih hanya berfokus pada aturan atau hasil akhir, etika kebajikan menekankan pada karakter dan niat para praktisi. Bagi pihak yang mengembangkan dan menerapkan sistem penilaian berbasis AI, hal ini berarti menumbuhkan kualitas diri seperti kejujuran, kerendahan hati, serta kesediaan untuk

menerima dan menindaklanjuti masukan, terutama saat menghadapi keterbatasan sistem atau ketidakpastian etis.

- **Etika kepedulian (*care ethics*)**: Perspektif ini berpusat pada konteks, hubungan antarindividu, dan komitmen untuk merespons kerentanan. Dalam penerapannya pada penilaian berbasis AI, etika kepedulian (*care ethics*) mendorong kita untuk mempertimbangkan dampak nyata terhadap individu, khususnya mereka yang paling sulit memperjuangkan kepentingan diri sendiri, seperti penutur non-asli atau Mahasiswa penyandang disabilitas.
- **Pendekatan pluralistik berbasis prinsip**: Sebagian besar kerangka kerja regulasi terkini seperti yang disusun oleh UNESCO, OECD, dan NIST mengacu pada berbagai tradisi serta menekankan nilai-nilai inti seperti keadilan, transparansi, akuntabilitas, privasi, dan otonomi. Pendekatan pluralistik ini memandu pengambilan keputusan etis dalam konteks yang kompleks, serta membantu para praktisi menyeimbangkan prioritas yang saling bersaing, misalnya antara akurasi dan kemampuan untuk memberikan penjelasan (*explainability*), atau antara efisiensi dan kesetaraan.

Isu Etis Utama dalam Penilaian AI

Dalam praktiknya, perspektif filosofis ini membekali kita untuk menghadapi tantangan etis yang ditimbulkan oleh AI dalam penilaian pendidikan; kita melangkah melampaui sekadar kepatuhan teknis demi memastikan bahwa keputusan yang diambil mencerminkan pengukuran yang tepat sekaligus komitmen nyata terhadap keadilan bagi seluruh peserta didik. Pemantauan terhadap penelitian terkini, pedoman internasional, dan standar profesional mengungkap serangkaian kekhawatiran etis yang kerap muncul dan sangat relevan bagi sistem penilaian berbasis AI:

- **Bias dan keadilan**: Sistem penilaian berbasis AI dapat mereplikasi dan bahkan memperparah pola ketimpangan historis jika sistem tersebut menggunakan data atau asumsi yang mencerminkan bias sosial yang ada. Memastikan sistem penilaian yang adil yaitu memberikan akses dan kesempatan yang setara bagi semua peserta didik memerlukan kewaspadaan berkelanjutan, strategi deteksi dan mitigasi bias yang tangguh, serta kesiapan untuk terus mengevaluasi apa yang dianggap "adil" dalam konteks yang beragam dan terus berkembang.
- **Transparansi dan kemampuan untuk dijelaskan (*explainability*):** Kompleksitas dan sifat "kotak hitam" (*opacity*) dari banyak sistem AI menyulitkan pemangku kepentingan untuk memahami bagaimana sistem penilaian berbasis AI mengambil keputusan atau untuk menentang hasil yang mereka anggap tidak adil. AI yang dapat dijelaskan (*explainable AI*) diperlukan tidak hanya untuk membangun kepercayaan dan penerimaan pemangku kepentingan, tetapi juga untuk menjunjung tanggung jawab etis terkait akuntabilitas dan pengambilan keputusan partisipatif.
- **Pengawasan manusia**: Seiring dengan semakin otomatisnya sistem penilaian, terdapat peningkatan risiko "bias otomatisasi" (*automation bias*), di mana pengambil keputusan dan pengguna terlalu mudah mengandalkan hasil keluaran algoritma. Mempertahankan peran aktif manusia (*human agency*) dan penilaian profesional

sangatlah penting untuk pengawasan yang efektif, koreksi kesalahan, dan keselarasan dengan nilai-nilai pendidikan inti

- **Privasi dan keamanan**: Penilaian yang diperkuat oleh AI melibatkan pengumpulan dan pembelajaran data pendidikan sensitif dalam jumlah besar. Melindungi privasi, martabat, dan keamanan peserta didik merupakan kewajiban regulasi sekaligus keharusan etis, yang menuntut tata kelola data yang ketat, transparansi dalam penggunaan data, dan kewaspadaan terhadap ancaman baru.

Dalam praktiknya, isu-isu ini saling tumpang tindih dan sering kali saling bergantung. Sebagai contoh, upaya untuk membuat model lebih mudah dijelaskan dapat memengaruhi kinerja prediktif, dan prinsip minimalisasi data demi privasi dapat menimbulkan ketegangan dengan kebutuhan akan pembelajaran keadilan yang komprehensif. Menangani pertukaran kepentingan (*trade-offs*) semacam itu memerlukan praktik reflektif, konsultasi yang luas, dan upaya untuk terus kembali pada tujuan utama, yaitu melayani peserta didik dan kepentingan publik. Bagian-bagian selanjutnya dalam bab ini akan mengkaji setiap ranah tersebut secara mendalam, pembahasan mekanisme munculnya risiko etis dalam penilaian berbasis AI, menyajikan strategi mitigasi praktis, serta menyoroti pelajaran nyata dari berbagai kemajuan dan tantangan terkini di bidang ini.

2.3 BIAS ALGORITMIK DALAM PENILAIAN AI

Seiring dengan terus berlanjutnya integrasi AI dan pembelajaran mesin (*machine learning*) dalam penilaian pendidikan, para pemangku kepentingan harus memberikan perhatian cermat terhadap keberadaan dan dampak bias algoritmik. Sebagaimana dalam penilaian tradisional, aspek keadilan tetap menjadi faktor kunci bagi validitas dan kegunaan skor yang dihasilkan AI bahkan mungkin lebih krusial lagi mengingat skala dan sifat "kotak hitam" (kurangnya transparansi) dalam cara kerja model pembelajaran mesin. Tanpa perancangan dan pemantauan yang cermat, sistem penilaian otomatis berisiko melanggengkan atau bahkan memperparah ketimpangan yang justru ingin mereka atasi.

Sumber Bias dalam Penilaian AI

Model AI dan pembelajaran mesin mewarisi dan terkadang memperbesar bias yang terdapat dalam desain, data, maupun penerapannya. Sebagai contoh, dalam penilaian otomatis, baik algoritma NLP maupun algoritma statistik tradisional yang mendasari sistem penilaian berbasis AI telah terbukti menimbulkan dampak yang berbeda-beda dan terkadang merugikan bagi kelompok demografis tertentu. Hal ini dapat terjadi dalam berbagai jenis aplikasi, mulai dari penilaian otomatis sederhana untuk jawaban singkat hingga pembelajaran multimodal terhadap tugas terbuka yang kompleks.

Konsep "Tujuh Sumber Dampak Merugikan dalam Pembelajaran Mesin" (*Seven Sources of Harm in Machine Learning*) dari Suresh dan Guttag menyediakan kerangka kerja yang bermanfaat untuk memahami munculnya bias dalam sistem AI. Kerangka ini menawarkan sudut pandang komprehensif bagi para praktisi untuk mengidentifikasi, mengantisipasi, dan menangani risiko secara sistematis. Dengan mengembangkan dan mengontekstualisasikan konsep tersebut untuk ranah penilaian, kita dapat melihat hal-hal berikut:

- **Bias historis**: Ketika data pelatihan mencerminkan pola diskriminasi masa lalu baik akibat kebijakan terdahulu, desain kurikulum, maupun tren sosial sistem penilaian berbasis AI dapat melanggengkan pola-pola tersebut ke dalam prediksi masa kini. Sebagai contoh, sistem penilaian otomatis yang dilatih menggunakan respons dari mahasiswa yang umumnya berasal dari kalangan mampu atau kelompok budaya mayoritas mungkin mengalami kesulitan dalam menangani gaya bahasa, topik, atau pengalaman yang lazim di kalangan kelompok yang kurang terwakili atau berstatus sosial-ekonomi rendah. Seiring dengan perubahan komposisi peserta tes dari waktu ke waktu, ketergantungan pada data historis dapat menghasilkan skor yang secara sistematis memberikan gambaran keliru mengenai kemampuan peserta didik saat ini maupun mereka yang secara historis terpinggirkan.
- **Bias representasi**: Bias ini muncul ketika data pengembangan untuk sistem penilaian berbasis AI kurang merepresentasikan populasi tertentu. Sebagai contoh, jika data pelatihan tidak memiliki cukup contoh dari pembelajar bahasa Inggris, mahasiswa di wilayah pedesaan, atau mahasiswa dari latar belakang ras minoritas, algoritma yang dihasilkan berisiko memberikan skor yang kurang akurat atau kurang andal bagi kelompok-kelompok ini, yang berpotensi memperlebar kesenjangan pendidikan.
- **Bias pengukuran**: Kekurangan dalam cara fitur penilaian dikonseptualisasikan atau dioperasionalkan dapat menghasilkan keluaran yang bias. Hal ini dapat terjadi akibat variabel yang terabaikan (fitur penting untuk penilaian yang memadai tidak disertakan), penggunaan proksi (menggunakan ukuran yang tidak sempurna sebagai pengganti aspek yang sebenarnya dinilai oleh penilai), atau ketergantungan pada variabel yang saling terkait (misalnya, fitur yang berkorelasi dengan keanggotaan kelompok sekaligus skor, namun tidak relevan dengan rubrik penilaian). Masing-masing faktor tersebut dapat menimbulkan kesalahan sistematis yang berbeda-beda antar-kelompok. Sebagai contoh, jika sistem penilaian esai otomatis hanya berfokus pada tata bahasa permukaan tanpa mempertimbangkan relevansi konten, dan jika pengetahuan konten yang relevan bervariasi secara sistematis berdasarkan latar belakang, maka kesenjangan antar-kelompok dapat muncul.
- **Bias agregasi**: Bias muncul ketika model diterapkan pada data gabungan dari berbagai subpopulasi, namun terdapat perbedaan dalam hubungan mendasar antar-kelompok tersebut. Sebagai contoh, jika mahasiswa dari lingkungan perkotaan dan pedesaan atau dari latar belakang budaya yang berbeda menggunakan norma retoris yang khas dalam tulisan maupun lisan, satu model tunggal mungkin gagal memberikan penilaian yang akurat bagi semua subkelompok. Penerapan sistem penilaian AI berskala besar yang mencakup berbagai negara bagian, wilayah, atau negara sering kali mengungkap masalah ini.
- **Bias pembelajaran/algoritmik**: Aspek teknis dalam pembuatan model termasuk pemilihan fungsi kerugian (*loss function*), regularisasi, atau paradigma pelatihan dapat secara tidak sengaja memengaruhi hasil bagi kelompok tertentu. Sebagai contoh, jika model disesuaikan untuk meminimalkan rata-rata kesalahan kuadrat (*mean squared*

error) secara keseluruhan pada kumpulan data di mana satu kelompok mendominasi secara jumlah, model tersebut akan mengoptimalkan hasil bagi kelompok tersebut dengan mengorbankan kelompok minoritas yang kurang terwakili.

- **Bias evaluasi**: Proses pengujian dan validasi sistem penilaian berbasis AI itu sendiri dapat menimbulkan bias jika tolok ukur evaluasi atau metrik kinerja mengabaikan subkelompok tertentu yang relevan. Jika pembelajaran keadilan hanya mempertimbangkan kinerja populasi secara agregat atau serangkaian indikator demografis yang terbatas, kesenjangan yang signifikan pada populasi yang lebih kecil atau kurang terlihat mungkin tidak terdeteksi.
- **Bias penerapan**: Model yang adil dan tangguh sekalipun dapat menimbulkan dampak negatif jika digunakan di luar konteks yang dimaksudkan; misalnya, menyajikan skor model yang bersifat kontinu sebagai keputusan kategoris, atau melakukan pemrosesan lanjutan (*post-processing*) terhadap skor tanpa mempertimbangkan kinerja atau validitas di tingkat kelompok. Selain itu, jika skor otomatis digunakan untuk pengambilan keputusan yang berdampak besar (seperti penempatan kelas atau kelulusan) tanpa pemantauan manusia yang memadai, bias kecil di tingkat kelompok dapat mengakibatkan konsekuensi nyata yang signifikan.

Memahami dan mengatasi sumber-sumber bias ini sangatlah penting bagi siapa pun yang mengembangkan atau menerapkan AI dalam bidang penilaian, khususnya dalam konteks yang ditandai oleh keberagaman dan ketimpangan yang terus berlangsung.

Strategi Mitigasi Bias

Mitigasi bias algoritmik dalam penilaian pendidikan memerlukan pendekatan berkelanjutan yang berlapis. Alih-alih mengandalkan solusi sesaat, pengembangan AI yang bertanggung jawab mengintegrasikan prinsip keadilan dan pengurangan bias di seluruh tahapan: pengumpulan data, desain sistem, evaluasi, dan penerapan. Studi membahas dua aspek utama dalam mempertimbangkan strategi mitigasi bias untuk sistem penilaian berbasis AI: representativitas data dan pembelajaran bias.

1) Representativitas Data

Memastikan bahwa data yang digunakan untuk melatih dan mengevaluasi sistem penilaian berbasis AI bersifat representatif terhadap populasi yang dilayani merupakan strategi mendasar. Hal ini berarti mendefinisikan secara eksplisit populasi yang menjadi sasaran dan mencari data yang mencerminkan keragamannya dalam hal karakteristik demografis, latar belakang pendidikan, dan faktor kontekstual. Sebagai contoh, penting untuk menyertakan sampel dari kelompok seperti pembelajar bahasa Inggris, mahasiswa di wilayah pedesaan, dan mahasiswa dari latar belakang kelompok minoritas. Tanpa representasi semacam itu, sistem penilaian berbasis AI dapat memberikan kinerja yang buruk atau tidak adil bagi kelompok yang kurang terwakili, sehingga memperparah kesenjangan yang sudah ada.

Pengembang juga harus memperhatikan bagaimana variabel kontekstual, seperti status disabilitas atau akses terhadap sumber daya pembelajaran, dapat memengaruhi hasil penilaian. Seiring dengan berkembangnya populasi mahasiswa dan praktik pendidikan, data harus diperbarui secara berkala untuk mencerminkan perubahan tersebut serta menjaga

relevansi dan keadilan penilaian berbasis AI. Terdapat beragam interpretasi mengenai apa yang dimaksud dengan data yang representatif; oleh karena itu, penting untuk mencermati bagaimana data yang dipilih untuk melatih dan mengevaluasi sistem penilaian berbasis AI dapat memengaruhi kinerja dan keluarannya.

2) Pembelajaran dan Evaluasi Bias

Pembelajaran bias yang menyeluruh sangat penting pada setiap tahap pengembangan sistem AI. Proses ini dimulai dengan melibatkan para pemangku kepentingan termasuk peserta didik, pendidik, administrator, dan perwakilan masyarakat untuk mendefinisikan makna keadilan dan bias dalam konteks pendidikan tertentu. Riset menyajikan berbagai definisi mengenai keadilan. Merangkum beragam definisi keadilan tersebut ke dalam tiga konsep berbeda yang didasarkan pada asumsi independensi yang berlainan di antara tiga variabel: keluaran sistem AI, variabel target (atau keluaran standar emas yang ingin dihasilkan oleh sistem), dan atribut sensitif yang sedang dipertimbangkan (misalnya, ras, karakteristik wilayah perkotaan/pedesaan, atau jenis kelamin). Ketiga jenis keadilan tersebut adalah:

- **Independensi**: atribut sensitif bersifat independen secara statistik terhadap prediksi;
- **Pemisahan (*separation*)**: prediksi atau keluaran AI bersifat independen secara kondisional terhadap atribut sensitif, dengan mempertimbangkan variabel target;
- **Kecukupan (*sufficiency*)**: prediksi tersebut memadai untuk menggambarkan hubungan antara atribut sensitif dan variabel target; artinya, variabel target bersifat independen secara kondisional terhadap atribut sensitif, dengan mempertimbangkan prediksi tersebut.

Differential item functioning, *differential algorithmic functioning*, serta metrik keadilan penilaian otomatis dari semuanya selaras dengan definisi keadilan berdasarkan prinsip pemisahan.

Pengembang dan pihak yang menerapkan sistem penilaian berbasis AI sebaiknya mengevaluasi sistem tersebut menggunakan kombinasi berbagai pendekatan, termasuk perspektif keadilan kelompok maupun individu. Sebagai contoh, keadilan kelompok berfokus pada hasil yang setara di berbagai kelompok demografis, sementara keadilan individu mempertimbangkan apakah individu yang serupa menerima perlakuan yang serupa. Keadilan kontrafaktual (*counterfactual fairness*) menawarkan sudut pandang lain, dengan mempertanyakan apakah keputusan akan tetap sama jika atribut sensitif diubah, sementara faktor-faktor lainnya dibuat tetap konstan. Yang tak kalah penting, laporan evaluasi tidak boleh hanya menyajikan hasil agregat; laporan tersebut harus menyajikan hasil yang dirinci berdasarkan subkelompok yang relevan guna memastikan bahwa kesenjangan dapat dikenali dan ditangani. Keterlibatan pemangku kepentingan secara berkelanjutan membantu memastikan bahwa keadilan tetap selaras dengan nilai-nilai masyarakat dan tujuan pendidikan.

3) Remediasi dan Pemantauan Berkelanjutan

Menangani bias dalam penilaian pendidikan berbasis AI memerlukan kombinasi antara intervensi teknis, refleksi etis, dan kewaspadaan yang berkelanjutan. Pada tahap pemodelan, beberapa strategi algoritmik dapat membantu mengurangi bias dan mendorong keadilan.

Sebagai contoh, algoritma yang mempertimbangkan aspek keadilan (*fairness-aware algorithms*) menerapkan batasan atau penalti selama proses pelatihan untuk memastikan bahwa tujuan keadilan diseimbangkan dengan akurasi prediksi.

Penelitian terkini mengusulkan pemberian penalti atau pembatasan pada model penilaian otomatis yang menghasilkan estimasi bias bukan-nol, yang memberikan cara konkret untuk membatasi kesenjangan skor antar-kelompok. Selain itu, metode pembelajaran yang peka terhadap biaya (*cost-sensitive learning*) dan optimasi dengan batasan (*constrained optimization*) dapat digunakan untuk memprioritaskan akurasi atau meminimalkan kesalahan bagi kelompok yang secara historis terpinggirkan, sehingga semakin mengurangi potensi kesenjangan hasil. Audit dan validasi berkala menggunakan kumpulan data dunia nyata yang beragam sangat penting untuk mendeteksi sumber bias baru atau yang muncul seiring berkembangnya sistem penilaian.

Ketika evaluasi mengidentifikasi adanya bias, langkah remediasi dapat mencakup penyeimbangan ulang kumpulan data, penyempurnaan pemilihan fitur, atau revisi tujuan model untuk mengatasi kesenjangan. Pengembang harus bersikap transparan mengenai pertukaran kepentingan (*trade-off*) yang terlibat dalam intervensi ini, karena upaya meningkatkan keadilan bagi satu kelompok terkadang dapat mengurangi akurasi prediksi secara keseluruhan atau menimbulkan dampak yang tidak diinginkan bagi kelompok lain. Dokumentasi terbuka mengenai pembelajaran bias dan upaya mitigasi, disertai dengan pemantauan berkala, akan mendukung akuntabilitas dan perbaikan berkelanjutan.

Langkah-langkah teknis saja tidak dapat menjamin keadilan. Kerangka kerja hukum dan etika seperti undang-undang anti-diskriminasi, *Joint Standards*, dan panduan dari organisasi seperti UNESCO juga harus menjadi acuan dalam strategi mitigasi bias. Meskipun kepatuhan terhadap regulasi diperlukan, sama pentingnya untuk mempertimbangkan implikasi etis yang lebih luas dari keluaran sistem, terutama seiring dengan berkembangnya nilai-nilai dan harapan masyarakat. Pada akhirnya, mitigasi bias dalam AI pendidikan merupakan proses berkelanjutan yang memerlukan integrasi cermat antara praktik data, kolaborasi pemangku kepentingan, inovasi teknis, dan pertimbangan etis. Dengan menerapkan prinsip-prinsip ini di sepanjang siklus hidup sistem, pengembang dan pendidik dapat lebih menjamin bahwa teknologi penilaian berbasis AI mampu mendorong peluang dan hasil yang adil bagi semua peserta didik.

Keterbatasan dan Tantangan Bias

Meskipun telah ada kemajuan yang signifikan, keterbatasan yang terus ada menjadi tantangan bagi kemampuan kita untuk sepenuhnya menghilangkan bias dalam sistem penilaian berbasis AI. Sebagian bias mungkin bersifat struktural, berakar pada pola kemasyarakatan yang lebih luas yang tidak dapat dengan mudah diperbaiki oleh teknik berbasis data.

Pertukaran kepentingan (*trade-off*) dapat muncul antara memaksimalkan akurasi prediksi dan meminimalkan kesalahan bagi kelompok yang paling rentan, atau antara kemampuan interpretasi model dan kinerja teknis. Selain itu, lingkungan penilaian di dunia nyata bersifat dinamis; populasi mahasiswa maupun praktik berbahasa terus berubah seiring

waktu, sehingga menuntut kewaspadaan dan fleksibilitas yang berkelanjutan dalam tata kelola AI. Terakhir, tidak ada solusi teknis yang dapat menggantikan penilaian manusia yang bermakna: mendukung pendidik dan institusi dalam mengenali, mengkaji secara kritis, dan menindaklanjuti bias algoritmik tetap menjadi tugas utama yang berkelanjutan.

Bias algoritmik merupakan salah satu tantangan etis paling kritis dan persisten dalam penggunaan AI untuk pengukuran dan penilaian pendidikan. Meskipun pengembang penilaian telah lama mengakui bias sebagai ancaman terhadap keadilan dan validitas dalam tes konvensional, penerapan sistem AI dan pembelajaran mesin (*machine learning*) yang kompleks memunculkan sumber, bentuk, dan konsekuensi bias yang baru, yang sering kali beroperasi dalam skala lebih besar dan dengan transparansi lebih rendah dibandingkan alat penilaian tradisional.

2.4 TRANSPARANSI DAN EXPLAINABILITY DALAM PENILAIAN AI

Transparansi menempati posisi mendasar dalam penerapan AI yang bertanggung jawab di bidang penilaian pendidikan. Dalam psikometri, transparansi telah lama diakui sebagai hal penting untuk membangun kepercayaan terhadap tes, mendukung interpretasi skor yang valid, dan memastikan para pemangku kepentingan dapat mengambil keputusan yang tepat berdasarkan hasil penilaian. Namun, dengan integrasi AI, tuntutan konseptual terkait transparansi menjadi jauh lebih kompleks dan mendesak. Berbeda dengan penilaian tradisional, sistem berbasis AI beroperasi melalui algoritma berlapis, sering kali menggunakan kumpulan data besar, ekstraksi fitur yang canggih, dan dalam beberapa kasus model "kotak hitam" (*black box*) seperti jaringan saraf dalam (*deep neural networks*).

Sistem-sistem ini dapat menghasilkan inferensi yang tampak sulit dipahami bahkan oleh praktisi ahli sekalipun. Oleh karena itu, komitmen terhadap transparansi menuntut organisasi untuk melakukan lebih dari sekadar memberikan pengungkapan teknis dasar. Mereka harus menyediakan informasi yang jelas, bermakna, dan tepat waktu bagi para pemangku kepentingan mengenai bagaimana sistem mengambil keputusan, bagaimana pengembang membangun dan memvalidasi model, bagaimana data dikumpulkan dan digunakan, serta asumsi dan keterbatasan apa yang membentuk keluaran sistem tersebut. Algoritma tidak cukup hanya sekadar akurat; algoritma juga harus dapat dipahami, dapat diakses, dan terbuka untuk ditelaah oleh komunitas pemangku kepentingan pendidikan yang luas, termasuk peserta didik, pendidik, administrator, pembuat kebijakan, dan bahkan masyarakat umum.

Peran Transparansi dalam Penilaian AI

Pada dasarnya, transparansi dalam penilaian berbasis AI mengacu pada pengungkapan yang jelas dan terbuka mengenai bagaimana sistem dibangun, diterapkan, dan dipelihara, serta bagaimana alur prosesnya berlangsung mulai dari pengumpulan data hingga pengambilan keputusan algoritmik. Dalam praktiknya, hal ini mencakup lebih dari sekadar merilis *pseudocode* atau ringkasan metrik kinerja model. Hal ini menuntut adanya:

- Dokumentasi yang jelas mengenai asal-usul (*provenance*) dataset, kriteria pemilihan, dan potensi bias yang terdapat dalam data.

- Keterbukaan mengenai jenis fitur dan variabel yang disertakan dalam desain model, serta alasan di balik penyertaan atau pengecualiannya.
- Pengungkapan asumsi model, kriteria optimisasi, dan rincian langkah-langkah pasca-pemrosesan, seperti penyesuaian skala (*rescaling*) atau pembulatan skor otomatis.
- Penjelasan mengenai kasus penggunaan yang dituju, keterbatasan yang diketahui, serta langkah-langkah yang diambil untuk melindungi privasi dan keamanan.

Transparansi meletakkan dasar bagi kepercayaan. Hal ini menciptakan kondisi yang diperlukan untuk audit eksternal, memungkinkan pemangku kepentingan menafsirkan keluaran sistem, serta menunjukkan komitmen institusional terhadap perilaku etis. Terutama dalam lingkungan pendidikan yang memiliki konsekuensi besar di mana skor yang dihasilkan AI dapat memengaruhi penempatan kelas, kelulusan, atau akses terhadap berbagai peluang transparansi berfungsi sebagai pelindung terhadap risiko "kotak hitam" (*black box*), yaitu situasi ketika sistem AI mengambil keputusan krusial dengan cara yang tidak dapat dipahami bahkan oleh pihak yang paling terdampak. Secara praktis, transparansi memainkan beberapa peran penting:

Mendorong keterlibatan pemangku kepentingan yang berlandaskan informasi: Transparansi memberdayakan seluruh pelaku pendidikan termasuk mahasiswa, keluarga, pendidik, administrator, dan pembuat kebijakan untuk menilai secara kritis cara kerja dan implikasi dari penilaian berbasis AI. Sebagai contoh, ketika proses penerimaan yang krusial menggunakan sistem penilaian berbasis AI, dokumentasi yang jelas mengenai struktur, kinerja, dan potensi sumber bias sistem tersebut memungkinkan pengawas serta individu yang terdampak untuk mengajukan pertanyaan relevan: apakah model tersebut responsif terhadap perubahan dalam populasi pendaftar? Apakah keterbatasan teknis menimbulkan beban yang tidak proporsional bagi mahasiswa dari sekolah dengan sumber daya terbatas? Apakah data pelatihan mencerminkan beragam lintasan pembelajaran?

Mendukung keadilan, akuntabilitas, dan prosedur yang wajar (*due process*): Hanya jika pemangku kepentingan memahami kriteria yang digunakan sistem penilaian berbasis AI untuk mengevaluasi hasil kerja atau potensi mahasiswa, mereka dapat secara bermakna menentang, mengajukan banding, atau menuntut perbaikan atas keputusan yang merugikan atau keliru. Hal ini sangat penting dalam konteks di mana skor yang dihasilkan AI sangat memengaruhi kemajuan pendidikan, peluang beasiswa, atau akses ke program bagi mahasiswa berbakat (*gifted programs*). Sebagai contoh, distrik sekolah yang menerapkan sistem penilaian esai berbasis AI tidak hanya harus mengomunikasikan tingkat keandalan penilaian rata-rata, tetapi juga menjelaskan kepada pendidik dan mahasiswa mengenai pembobotan fitur respons tertentu (misalnya, tata bahasa, koherensi, relevansi topik) serta apakah algoritma yang digunakan secara sistematis merugikan kelompok tertentu (misalnya, penutur non-pribumi atau mahasiswa dari latar belakang yang kurang terwakili).

Memungkinkan pemeriksaan teknis dan perbaikan berkelanjutan: Transparansi teknis diperlukan agar ahli psikometri, ilmuwan data, dan peneliti independen dapat mengaudit, memvalidasi, dan menyempurnakan sistem penilaian AI. Hal ini mencakup publikasi arsitektur model, prosedur pelatihan, kriteria evaluasi, serta distribusi kesalahan (*error*) di berbagai

kelompok demografis atau jenis butir soal. Sistem yang transparan memungkinkan identifikasi kesenjangan kinerja secara sistematis, mendukung upaya mitigasi bias, dan mendorong budaya keterbukaan dalam ilmu pengukuran. Namun, penting untuk diakui bahwa transparansi penuh terkadang berbenturan dengan prioritas sah lainnya. Sebagai contoh, peraturan untuk mencegah manipulasi tes dapat membatasi pengungkapan rinci mengenai algoritma. Demikian pula, perlindungan informasi kepemilikan atau privasi data pelatihan mungkin mengharuskan pengungkapan yang diseimbangkan secara cermat. Mencapai keseimbangan yang tepat, seperti biasa, bergantung pada konteks dan didasarkan pada dialog berkelanjutan antara pemangku kepentingan teknis dan non-teknis.

Contoh nyata dari ketegangan ini muncul dalam konteks sistem penilaian esai otomatis, khususnya pada *e-rater* milik *Educational Testing Service* (ETS). Seiring meluasnya penggunaan *e-rater* untuk penilaian yang memiliki konsekuensi besar (*high-stakes assessments*), kekhawatiran pun muncul mengenai potensi peserta tes memanfaatkan kelemahan dalam algoritma penilaian sistem tersebut. Generator BABEL sebuah program yang dikembangkan oleh Les Perelman untuk menghasilkan esai yang pada dasarnya tidak bermakna namun disusun sedemikian rupa sehingga dapat mengecoh mesin penilaian otomatis agar memberikan skor tinggi mengilustrasikan masalah ini dengan jelas. Keberadaan BABEL menyoroti risiko dari pengungkapan detail algoritma yang terlalu terbuka: pengetahuan publik mengenai kerentanan sistem memungkinkan terjadinya serangan yang disengaja (*adversarial attacks*), yang pada akhirnya melemahkan validitas hasil tes.

Sebagai tanggapan, para peneliti ETS mengembangkan metode untuk mendeteksi teks buatan BABEL dan bentuk masukan adversarial lainnya. Pekerjaan mereka menunjukkan perlunya langkah-langkah perlindungan yang berkelanjutan dalam desain sistem, sekaligus memperkuat pentingnya transparansi yang adaptif. Pengungkapan algoritma secara penuh mungkin tidak selalu dapat dilakukan atau diinginkan; sebaliknya, organisasi harus mencapai keseimbangan yang cermat antara keterbukaan dan keamanan, guna memastikan bahwa transparansi mendukung kepentingan publik sekaligus menjaga integritas sistem penilaian.

***Explainability* dalam Penilaian AI**

Kemampuan untuk dijelaskan mengacu pada sejauh mana mekanisme internal dan keluaran (*output*) suatu sistem AI dapat dipahami oleh pemangku kepentingan manusia. Dalam konteks penilaian, hal ini mencakup lebih dari sekadar menafsirkan bobot fitur atau membuka "kotak hitam algoritmik" (*algorithmic black box*). Kemampuan ini menuntut kita untuk bertanya: dapatkah peserta didik, pengajar, dan administrator memahami alasan di balik pemberian skor, rekomendasi, atau klasifikasi tertentu oleh AI? Studi mengidentifikasi empat prinsip dasar sistem AI yang dapat dijelaskan (explainable AI):

1. **Penjelasan**: Sistem harus menghasilkan penjelasan atau dasar pemikiran (*rationale*) atas prediksi atau keputusannya.
2. **Bermakna**: Penjelasan harus dapat dipahami dan sesuai dengan konteks bagi audiens yang dituju. Sebagai contoh, informasi yang bermakna bagi seorang psikometrisian mungkin sulit dipahami oleh mahasiswa, dan sebaliknya.

3. **Akurasi penjelasan**: Penjelasan harus mencerminkan secara akurat data, fitur, dan langkah-langkah nyata yang menghasilkan keluaran tersebut, bukan sekadar narasi yang disederhanakan atau narasi "pemasaran".
4. **Batasan pengetahuan**: Sistem harus mengidentifikasi kasus-kasus di mana sistem tersebut tidak dirancang untuk beroperasi, atau situasi di mana keluarannya mungkin bermasalah (misalnya, reliabilitas rendah atau masalah bias).

Upaya mewujudkan kemampuan untuk dijelaskan dalam AI untuk pengukuran pendidikan menghadapi beberapa tantangan yang terus ada:

- **Kompleksitas model dasar**: Sistem mutakhir, seperti *deep neural networks*, metode ensemble, atau model bahasa berskala besar (*large language models*), sering kali secara inheren kurang dapat ditafsirkan dibandingkan model statistik yang lebih sederhana, sehingga sulit untuk memberikan dasar pemikiran yang jelas bagi prediksi individual.
- **Pertukaran (*trade-off*) dengan kinerja sistem**: Terdapat ketegangan yang terdokumentasi dengan bai antara transparansi atau kemampuan penafsiran model dengan akurasi prediksi secara keseluruhan. Model yang lebih sederhana dan transparan mungkin mengorbankan nuansa dan detail halus, sedangkan model yang lebih kompleks mungkin memberikan akurasi lebih tinggi namun sulit dijelaskan secara sederhana. Institusi pendidikan harus mempertimbangkan kapan kesederhanaan lebih diperlukan, terutama ketika kepercayaan atau keadilan dalam situasi berisiko tinggi (*high-stakes*) menjadi perhatian utama.
- **Keragaman audiens**: Pengguna yang berbeda membutuhkan jenis penjelasan yang berbeda pula.

Sistem yang dapat dijelaskan kepada psikometrisian mungkin tidak mudah dipahami oleh peserta didik atau orang tua. Oleh karena itu, aspek keterjelasan harus dirancang secara khusus dengan menyediakan sumber daya yang berlapis atau disesuaikan dengan audiens tertentu mulai dari dokumentasi teknis dan panduan yang disederhanakan hingga visualisasi interaktif yang memperlihatkan logika sistem.

Terlepas dari berbagai tantangan tersebut, keterjelasan tetap menjadi sarana utama untuk mendorong keadilan, kepercayaan, dan penerapan AI yang bertanggung jawab dalam penilaian pendidikan. Yang tak kalah penting, keterjelasan merupakan aspek yang bersifat teknis sekaligus organisasional; pengembangannya tidak hanya bergantung pada pemilihan algoritma, tetapi juga pada kebijakan, dokumentasi, dan pengembangan profesional.

Kepercayaan dan Penerimaan Pemangku Kepentingan

Penelitian empiris menunjukkan bahwa kepercayaan pemangku kepentingan terhadap penilaian berbasis AI sangat bergantung pada transparansi dan kemampuan penjelasan (*explainability*). Ketika pengguna tidak memahami bagaimana atau mengapa skor ditentukan, tingkat penerimaan menurun, penolakan meningkat, dan tuntutan akan intervensi manusia atau upaya banding pun bertambah. Sebaliknya, penjelasan yang dikomunikasikan dengan baik seperti panduan bagi instruktur, *tooltip* pada antarmuka, atau panduan langkah demi

langkah (*walkthrough*) dapat secara signifikan meningkatkan persepsi keadilan serta mendukung proses pembelajaran dan perbaikan berkelanjutan.

Contoh nyata dapat ditemukan pada *platform* penilaian esai otomatis. Sistem yang menampilkan fitur linguistik atau elemen struktural yang memengaruhi skor serta mampu mengaitkannya dengan rubrik penilaian yang diketahui pengguna cenderung lebih dipercaya dan diadopsi baik oleh mahasiswa maupun pengajar. Sebaliknya, algoritma tertutup (*black box*) yang bersifat eksklusif di mana penyedia tes sekalipun tidak dapat menjelaskan alasan di balik pemberian skor tertentu sering kali menghadapi penolakan dan skeptisisme, terutama ketika skor yang dihasilkan menyimpang dari ekspektasi manusia atau tampak merugikan kelompok minoritas.

Demikian pula dalam penerapan teknologi pembelajaran adaptif berskala besar, institusi yang menjelaskan secara gamblang bagaimana rekomendasi dibuat melalui penyediaan dokumentasi yang mudah diakses serta didukung oleh pengujian pengguna dan siklus umpan balik mendapati kesiapan yang lebih tinggi dari para pendidik dan peserta didik untuk mengintegrasikan teknologi tersebut ke dalam praktik sehari-hari.

Tantangan Transparansi dan *Explainability*

Terlepas dari berbagai keunggulannya, penerapan transparansi dalam penilaian berbasis AI menghadirkan tantangan teknis dan praktis yang signifikan. Model yang sangat kompleks mungkin sulit dijelaskan secara intuitif; batasan kepemilikan atau pertimbangan hak kekayaan intelektual dapat membatasi keterbukaan informasi; dan, sebagaimana telah disebutkan sebelumnya, kecanggihan model sering kali mengorbankan kejelasan, sehingga mekanisme internal sistem menjadi lebih sulit dijelaskan kepada pemangku kepentingan awam.

Selain itu, jika transparansi diterapkan secara buruk, hal tersebut dapat berubah menjadi sekadar formalitas belaka membebani pemangku kepentingan dengan detail teknis yang penuh jargon atau secara selektif menyembunyikan sumber kesalahan maupun bias model yang dianggap tidak nyaman. Oleh karena itu, transparansi harus berlandaskan komunikasi yang efektif, disesuaikan dengan kebutuhan beragam audiens, dan senantiasa memperhatikan prinsip etika mendasar: menjadikan cara kerja dan konsekuensi penilaian AI sebagai sesuatu yang terlihat jelas, dapat dipahami, dan dapat diperdebatkan atau disanggah.

Singkatnya, transparansi bukanlah solusi ajaib untuk segala masalah, namun merupakan fondasi mutlak bagi penerapan AI yang etis dalam penilaian pendidikan. Hal ini memperjelas "bagaimana" dan "mengapa" di balik penilaian otomatis, mendukung akuntabilitas, serta menopang upaya kolaboratif yang berkelanjutan dalam ilmu pengukuran demi mewujudkan pemerataan dan peningkatan mutu pendidikan.

2.5 PENGAWASAN MANUSIA DALAM PENILAIAN AI

Meskipun kemajuan dalam penilaian berbasis AI menjanjikan manfaat transformatif bagi pendidikan, sistem-sistem ini bukanlah dan tidak seharusnya menjadi penentu otonom atas potensi atau pencapaian mahasiswa. Integrasi AI yang bertanggung jawab ke dalam pengukuran pendidikan menuntut komitmen teguh terhadap pengawasan manusia yang

bermakna di setiap tahapan siklus hidup penilaian. Penilaian manusia, kesadaran kontekstual, dan pertimbangan etis tetap tak tergantikan, terutama dalam skenario pengambilan keputusan yang krusial atau sensitif.

Pengawasan manusia mencakup spektrum kegiatan yang luas, mulai dari perancangan dan validasi sistem penilaian berbasis AI hingga pemantauan berkelanjutan terhadap keluaran dan dampaknya dalam konteks dunia nyata. Pendidik, profesional di bidang penilaian, dan pembuat kebijakan harus tetap memegang otoritas tertinggi atas keputusan-keputusan yang berdampak besar, memastikan bahwa AI berfungsi sebagai alat bantu untuk pengambilan keputusan yang berdasar (*informed judgment*), bukan sebagai sumber kebenaran mutlak yang tidak boleh dipertanyakan. Pendekatan yang melibatkan manusia secara aktif dalam proses ini (*human-in-the-loop*) tidak hanya mendukung deteksi dan koreksi kesalahan atau bias yang mungkin luput dari sistem otomatis, tetapi juga menjunjung tinggi prinsip-prinsip dasar keadilan, prosedur yang wajar (*due process*), dan akuntabilitas dalam pendidikan.

Selain itu, penggunaan AI yang bertanggung jawab dalam penilaian bukan sekadar masalah teknis atau prosedural; hal ini merupakan kewajiban etis dan profesional yang berkelanjutan. Hal ini melibatkan pengembangan budaya refleksi kritis, perbaikan berkelanjutan, dan pelibatan pemangku kepentingan, di mana pengembang dan pihak yang menerapkan teknologi berkomunikasi secara terbuka mengenai kemampuan serta keterbatasan AI, serta mekanisme yang tersedia untuk pengajuan banding, pemulihan hak, dan transparansi. Dengan mengedepankan pengawasan manusia dan pengelolaan yang etis, institusi pendidikan dapat memastikan bahwa penerapan teknologi penilaian berbasis AI justru memajukan bukan melemahkan misi utama mereka dalam mendukung pembelajaran, kesetaraan, dan kesejahteraan mahasiswa.

Peran Penilaian Manusia

Meskipun sistem AI dalam penilaian pendidikan semakin canggih, nilai penilaian manusia yang tak tergantikan tetap bertahan. Para ahli manusia membawa nuansa interpretatif, pemahaman kontekstual, dan pertimbangan etis yang tidak dapat direplikasi sepenuhnya oleh AI, terlepas dari seberapa besar kemajuannya. Meskipun AI mampu memproses kumpulan data yang sangat besar dengan cepat, mengidentifikasi pola statistik yang samar, dan memberikan penilaian yang konsisten dalam skala luas, teknologi ini tidak dapat mempertimbangkan faktor-faktor seperti konteks budaya, riwayat pembelajaran individu, atau respons mahasiswa yang tidak standar yang mungkin memerlukan pertimbangan khusus.

Penilaian manusia memainkan peran krusial dalam beberapa tahapan proses asesmen. Pada tahap perancangan, pendidik dan ahli psikometri menetapkan tujuan pembelajaran, memilih fitur yang relevan, serta menentukan batasan penggunaan model yang tepat keputusan-keputusan yang menuntut keahlian pedagogis dan penalaran berbasis nilai. Dalam konteks operasional, pengawasan manusia sangat penting untuk memantau keluaran yang tidak wajar, memutuskan kasus-kasus yang berada di ambang batas kriteria, serta meninjau situasi di mana skor yang dihasilkan AI bertentangan dengan penilaian pengajar atau laporan mandiri mahasiswa. Sebagai contoh, ketika sistem penilaian esai otomatis menandai esai

sebagai tulisan yang melenceng dari topik atau memberikan skor yang sangat rendah secara tak terduga, peninjau manusia dapat mempertimbangkan faktor-faktor kontekstual seperti kemahiran berbahasa, interpretasi terhadap instruksi tugas, atau bukti ekspresi kreatif yang mungkin tidak sepenuhnya tertangkap oleh algoritma.

Selain itu, penilaian manusia sangat vital untuk menjaga keadilan dan kesetaraan. Sistem otomatis, jika tidak diawasi, berisiko melanggengkan atau memperparah bias yang sudah ada dalam data pendidikan. Dengan menyediakan ruang untuk banding, audit, dan intervensi, peninjau manusia dapat mengidentifikasi dan menangani hasil yang tidak adil, serta memastikan bahwa AI berfungsi sebagai pendukung bukan pengganti bagi keahlian profesional dan tanggung jawab etis. Pada akhirnya, sistem asesmen yang paling dapat diandalkan dan efektif adalah sistem yang memadukan efisiensi serta kemampuan analitis AI dengan daya pertimbangan, empati, dan wawasan kritis dari pendidik serta pengambil keputusan manusia.

Seiring dengan semakin integralnya peran AI dalam asesmen pendidikan, memastikan adanya pengawasan manusia yang kuat tetap menjadi hal yang sangat penting. AI menjanjikan peningkatan efisiensi, skalabilitas, dan terkadang objektivitas; namun, penelitian dan panduan kebijakan menegaskan bahwa penilaian manusia sangat diperlukan untuk menjamin penggunaan teknologi asesmen yang etis, adil, dan valid.

Bias Otomatisasi dan Ketergantungan AI

Integrasi AI yang pesat ke dalam asesmen pendidikan membawa risiko signifikan terkait bias otomatisasi dan ketergantungan berlebihan pada pengambilan keputusan berbasis algoritma. Bias otomatisasi mengacu pada kecenderungan pengguna untuk menerima begitu saja atau lebih memercayai rekomendasi yang diberikan oleh sistem otomatis, bahkan ketika terdapat bukti yang bertentangan atau hal tersebut tidak sejalan dengan akal sehat. Dalam konteks pendidikan, bias ini dapat mendorong pengajar, administrator, dan pembuat kebijakan untuk mengandalkan skor atau klasifikasi yang dihasilkan AI tanpa mempertanyakan validitas, akurasi, atau keadilannya secara memadai.

Ketergantungan berlebihan semacam itu dapat menimbulkan konsekuensi serius, termasuk melanggengkan bias yang sudah ada, meminggirkan peserta didik dengan karakteristik non-standar, serta mengurangi peran penilaian manusia dalam menafsirkan respons mahasiswa yang kompleks atau bergantung pada konteks. Meskipun sistem otomatis memiliki kemampuan yang hebat, tingkat keadilan dan efektivitasnya sangat bergantung pada data dan asumsi yang digunakan; kepercayaan tanpa kendali terhadap hasil sistem tersebut dapat mengikis mekanisme perlindungan penting seperti prosedur banding dan peninjauan individual yang krusial untuk menjamin kesetaraan dan akuntabilitas dalam pendidikan. Pemanfaatan AI secara etis menuntut kewaspadaan berkelanjutan terhadap bias otomatisasi, serta memastikan bahwa pendidik tetap terlibat aktif dalam pengawasan, refleksi kritis, dan proses pengambilan keputusan akhir.

Keterlibatan Manusia dalam Penilaian AI

Untuk benar-benar memanfaatkan keunggulan AI dalam penilaian pendidikan sekaligus memitigasi risikonya, diperlukan pendekatan yang terencana dan sistematis terkait pengawasan oleh manusia. Keterlibatan yang kuat dari pendidik, administrator, dan pemangku

kepentingan lainnya harus diintegrasikan ke dalam setiap tahapan perancangan, penerapan, dan penggunaan sistem AI. Hal ini dimulai dengan pengembangan profesional yang komprehensif: sekolah dan distrik pendidikan perlu berinvestasi dalam pembelajaran berkelanjutan yang bersifat praktis untuk membangun literasi AI, memperjelas kemampuan serta keterbatasan teknologi penilaian, dan memberdayakan staf agar mampu menafsirkan hasil algoritma secara kritis serta bila diperlukan mempertanyakan hasil tersebut. Pelatihan berbasis skenario dapat membantu pendidik menyelaraskan rekomendasi hasil AI dengan realitas kelas yang kompleks, sekaligus mengembangkan kemampuan penilaian yang diperlukan untuk mengenali kapan intervensi perlu dilakukan.

Selain pelatihan, protokol pengawasan yang efektif sangatlah penting. Kebijakan harus mewajibkan agar keputusan penting terkait aspek akademik, penempatan, atau disiplin ditinjau oleh pendidik yang kompeten, guna memastikan bahwa hasil keluaran AI tidak dianggap sebagai keputusan final tanpa adanya pertimbangan manusia. Mekanisme penyelesaian masalah yang transparan harus tersedia agar mahasiswa, keluarga, atau pendidik dapat menyanggah atau mengajukan banding atas hasil yang dihasilkan oleh AI. Panel peninjau yang terstruktur harus diberi wewenang untuk melakukan intervensi jika diperlukan, sehingga memberikan perlindungan penting terhadap potensi kesalahan atau ketidakadilan. Proses-proses ini tidak hanya melindungi peserta didik secara individu, tetapi juga menjunjung tinggi integritas dan akuntabilitas sistem pendidikan.

Selain itu, keterlibatan manusia yang bermakna bergantung pada akuntabilitas yang terdistribusi dan kolaboratif. Pengawasan terhadap keputusan AI harus menjadi tanggung jawab bersama, dengan memanfaatkan beragam keahlian dari pengajar, administrator, pakar teknis, serta jika relevan mahasiswa dan keluarga mereka sendiri. Institusi perlu melakukan audit berkala terhadap keputusan yang dihasilkan AI untuk mengidentifikasi pola kesalahan atau bias yang sistematis, serta secara aktif meminta masukan dari para pemangku kepentingan guna memandu perbaikan berkelanjutan. Pendekatan partisipatif ini membangun kepercayaan terhadap teknologi dan memastikan bahwa proses penilaian tetap adil, transparan, dan tanggap terhadap kebutuhan semua peserta didik yang terus berkembang.

Penggunaan AI yang etis dan efektif dalam penilaian pendidikan tidak bergantung pada penggantian keahlian profesional, melainkan pada penguatannya. Sistem penilaian berbasis AI harus menyediakan bukti berharga untuk pertimbangan manusia; AI tidak bertindak sebagai penentu keputusan akhir (*arbiter*), melainkan sebagai mitra. Melalui pengembangan profesional yang mumpuni, kebijakan institusional yang kuat, dan pengawasan yang cermat, dunia pendidikan dapat mewujudkan potensi AI sembari tetap menjaga keadilan, otonomi, dan kepercayaan sebagai inti dari praktik penilaian. Dengan menciptakan struktur yang mendukung refleksi berkelanjutan, komunikasi transparan, dan keterlibatan pemangku kepentingan, sistem pendidikan dapat memastikan bahwa teknologi AI memajukan bukan justru melemahkan misi utama mereka dalam mendukung pembelajaran, kesetaraan, dan kesejahteraan mahasiswa.

2.6 PRIVASI DAN PERLINDUNGAN DATA DALAM PENILAIAN AI

Seiring sistem AI menjadi landasan dalam pengukuran pendidikan, sistem tersebut semakin bergantung pada pengumpulan, penyimpanan, dan pembelajaran data peserta didik yang sangat banyak dan bersifat sensitif. Data ini mencakup tidak hanya nilai tes dan jawaban tertulis, tetapi juga data perilaku, biometrik, atau multimedia yang kaya informasi (misalnya, video wajah, rekaman suara). Meskipun aliran data ini memungkinkan personalisasi yang lebih baik dan analitik yang canggih, hal ini secara bersamaan meningkatkan risiko privasi dan keamanan risiko yang melampaui sekadar kegagalan teknis dan menyentuh persoalan etika serta sosial yang mendasar.

Konteks unik penilaian pendidikan membuat taruhannya menjadi lebih tinggi. Peserta didik yang sering kali masih di bawah umur beserta keluarga mereka mungkin tidak selalu memiliki pengetahuan, kuasa, atau sumber daya untuk memberikan persetujuan berdasarkan informasi yang memadai (*informed consent*), menyanggah pengumpulan data yang tidak tepat, atau mengenali ancaman yang tidak kentara. Kebocoran, penyalahgunaan, atau pembagian data penilaian yang tidak disengaja dapat berujung pada pencurian identitas, kerusakan reputasi, diskriminasi, atau bahkan dampak psikologis yang merugikan. Selain itu, data pendidikan berskala besar semakin menjadi sasaran pihak-pihak yang berniat jahat (misalnya, serangan *ransomware*), dan berbagai regulasi (seperti FERPA dan GDPR) menetapkan kewajiban hukum yang harus dipenuhi.

Implikasi Etis Privasi Data

Pihak yang menerapkan sistem penilaian berbasis AI harus memahami privasi dan keamanan bukan sekadar sebagai daftar periksa kepatuhan atau tantangan teknis semata, melainkan sebagai keharusan etis yang mendasar bagi terciptanya kepercayaan dan keadilan dalam lingkungan pendidikan. Perlindungan data peserta didik berkaitan erat dengan prinsip-prinsip otonomi, martabat, dan keadilan.

- **Otonomi**: Peserta didik harus memiliki kendali yang bermakna atas data apa saja yang dikumpulkan mengenai diri mereka serta bagaimana sistem menggunakan, membagikan, dan menyimpan data tersebut. Dalam praktiknya, hal ini berarti perlunya kebijakan yang kuat terkait persetujuan berdasarkan informasi yang memadai (*informed consent*), pemberitahuan privasi yang disesuaikan dengan usia, serta opsi untuk tidak berpartisipasi (*opt-out*) jika memungkinkan.
- **Martabat dan non-diskriminasi**: Penanganan data sensitif yang tidak tepat dapat membuat peserta didik rentan terhadap perlakuan yang tidak adil. Sebagai contoh, catatan penilaian yang bocor atau dibagikan tanpa persetujuan dapat menimbulkan bias dalam penempatan, penerimaan perpengajaran tinggi, atau kesempatan kerja.
- **Kekuasaan dan kesetaraan**: Kelompok yang terpinggirkan mungkin lebih rentan terhadap dampak negatif terkait privasi, kurang memiliki kemampuan untuk menentang pengumpulan atau penggunaan data yang tidak tepat, serta menanggung dampak yang lebih berat akibat kebocoran data. Oleh karena itu, pengelolaan data yang bertanggung jawab menuntut kewaspadaan ekstra dan dukungan proaktif bagi para pembelajar ini.

Prinsip Perlindungan Data

Pengelolaan data yang bertanggung jawab merupakan proses berkelanjutan yang memerlukan ketelitian teknis, investasi kelembagaan, serta budaya transparansi dan kewaspadaan. Daftar berikut menyajikan praktik-praktik yang direkomendasikan untuk penanganan data secara etis.

- **Keamanan sejak tahap perancangan (*secure by design*):** Mengikuti panduan dari badan keamanan siber, sistem harus dirancang sejak awal untuk mencegah akses tidak sah, meminimalkan penyimpanan data sensitif, dan memastikan isolasi komponen-komponen yang krusial bagi keamanan. Enkripsi, kontrol akses, dan penilaian kerentanan secara berkala merupakan persyaratan dasar.
- **Minimalisasi dan pembatasan tujuan**: Hanya kumpulkan data yang benar-benar diperlukan untuk tujuan penilaian atau pembelajaran, serta batasi masa penyimpanan data sesuai dengan kebutuhan untuk memenuhi tujuan tersebut. Akumulasi data yang tidak perlu tidak hanya meningkatkan risiko kebocoran, tetapi juga dapat melanggar standar etika dan hukum.
- **Transparansi dan persetujuan**: Berikan penjelasan dalam bahasa yang mudah dipahami kepada pembelajar dan wali mengenai data apa yang dikumpulkan, bagaimana data tersebut akan digunakan, dan kepada siapa data tersebut akan dibagikan. Persetujuan harus didasarkan pada informasi yang memadai, bersifat sukarela, dan dapat diubah kembali; kebijakan harus mematuhi norma hukum sekaligus berupaya sungguh-sungguh untuk menghormati otonomi individu.
- **Pemantauan berkelanjutan dan respons insiden**: Organisasi pendidikan harus mengembangkan, menyimulasikan, dan memperbarui secara berkala protokol mereka untuk mendeteksi, melaporkan, serta menanggapi kebocoran atau penggunaan data yang tidak tepat. Respons insiden bukan sekadar masalah teknis; komunikasi dengan pembelajar dan keluarga yang terdampak merupakan hal yang krusial.
- **Audit privasi berkala**: Lakukan penilaian dampak privasi dan pemantauan keamanan oleh pihak ketiga secara berkala, terutama ketika sistem AI mengalami pembaruan signifikan atau saat diperkenalkannya modalitas data baru.

2.7 STUDI KASUS ETIKA PLATFORM PENILAIAN AI

Protokol Evaluasi Etis

Penerapan platform penilaian berbasis AI di lingkungan pendidikan berkembang pesat, namun penggunaannya memunculkan pertanyaan penting terkait aspek etika, teknis, dan praktis. Sebelum implementasi atau pengadaan, sangatlah penting untuk menilai platform semacam itu secara sistematis menggunakan protokol yang mencakup dimensi etis dan operasional utama. Tabel berikut menyajikan protokol terstruktur beserta pertanyaan-pertanyaan pemandu (Tabel 2.1).

Tabel 2.1 Protokol untuk Mengevaluasi Penerapan AI secara Etis

Tahap Evaluasi	Pertanyaan Kunci
Dokumentasi, transparansi, dan kemampuan untuk dijelaskan	Apakah tersedia dokumentasi teknis dan prosedural? Apakah logika sistem, alur data, dan rincian model dijelaskan dengan jelas? Apakah pengajar dan mahasiswa menerima penjelasan yang jelas mengenai nilai?
Validitas dan reliabilitas	Apakah terdapat bukti yang telah dipublikasikan yang menunjukkan bahwa nilai yang dihasilkan oleh AI akurat dan konsisten dibandingkan dengan penilai manusia yang berkualifikasi?
Keadilan dan bias	Apakah pembelajaran subkelompok atau audit bias telah dilakukan dan dilaporkan? Apakah terdapat mekanisme untuk memitigasi bias yang ditemukan?
Pengawasan oleh pengguna	Bisakah pengajar mengesampingkan hasil keluaran AI? Apakah tersedia mekanisme banding?
Privasi dan keamanan	Apakah langkah-langkah perlindungan privasi tangguh dan transparan? Apakah kepatuhan terhadap peraturan seperti FERPA ditunjukkan secara jelas?
Keterbatasan dan akuntabilitas	Apakah batasan-batasan sistem tersebut dinyatakan dengan jelas? Apakah terdapat pengakuan secara terbuka mengenai potensi kesalahan, audit independen, serta penetapan tanggung jawab yang jelas atas kesalahan atau dampak merugikan? Apakah tersedia mekanisme untuk peninjauan oleh pihak ketiga dan akuntabilitas yang berkelanjutan?

Penerapan Protokol Evaluasi

Untuk menelusuri penawaran terkini dalam penilaian esai otomatis, kami melakukan pencarian Google menggunakan frasa "*AI Grading for Teachers*" (Penilaian AI untuk Pengajar). Kami memilih hasil pertama yang bukan merupakan konten bersponsor (tidak dipromosikan), yang mengarahkan kami ke halaman utama (*landing page*) sebuah platform penilaian berbasis AI yang dirancang untuk para pendidik. Halaman tersebut segera memosisikan *platform* ini sebagai solusi untuk mengurangi beban kerja pengajar serta meningkatkan kualitas dan kecepatan pemberian umpan balik terhadap tulisan mahasiswa. Judul utamanya menekankan penghematan waktu yang signifikan ("menilai esai dengan waktu 80% lebih singkat"), sementara teks pendukung menyoroti bahwa alat ini telah digunakan di lebih dari seribu sekolah. Halaman tersebut menyajikan sistem ini sebagai "Penilai Esai AI" yang memungkinkan pengajar memberikan umpan balik yang dipersonalisasi dan selaras dengan standar secara lebih efisien, dengan berbagai fitur, antara lain:

- Integrasi dengan sistem manajemen pembelajaran (LMS) yang umum digunakan (*Google Classroom, Canvas, Schoology*)
- Dukungan untuk rubrik ujian negara bagian dan kriteria penilaian yang disesuaikan
- Wawasan berbasis data untuk perencanaan pembelajaran dan identifikasi kesenjangan pembelajaran

- Penekanan pada "keadilan dalam penilaian" dan minimalisasi bias manusia melalui penilaian objektif berbasis rubrik
- Umpan balik terdiferensiasi yang disesuaikan dengan kebutuhan setiap mahasiswa

Platform ini juga menarik minat calon pengguna melalui serangkaian testimoni pengajar, yang memuji kemampuan alat tersebut dalam memberikan umpan balik yang rinci dan tepat waktu, mengurangi subjektivitas penilaian, serta menghemat waktu. Sejumlah pengajar menyebutkan adanya dampak transformatif terhadap beban kerja dan praktik pembelajaran mereka di kelas. Halaman tersebut juga menyediakan informasi praktis mengenai cara kerja sistem: pengajar mengimpor tugas, menentukan kriteria penilaian atau rubrik, menerima nilai dan umpan balik yang dihasilkan AI, serta tetap memiliki kemampuan untuk meninjau dan menyesuaikan keduanya sebelum dibagikan kepada Mahasiswa.

Platform ini menegaskan bahwa pendidik tetap memegang kendali penuh atas nilai akhir. Pada bagian pertanyaan yang sering diajukan (FAQ), platform membahas topik-topik seperti kompatibilitas dengan berbagai standar negara bagian, privasi dan keamanan data (termasuk kepatuhan terhadap FERPA dan NIST, enkripsi data, dan pengumpulan data minimal), serta penggunaan data yang telah dianonimkan bersama penyedia AI pihak ketiga. Perusahaan menegaskan bahwa mereka (dan pihak ketiga) tidak menggunakan data mahasiswa untuk melatih model dan bahwa tanggung jawab penilaian akhir tetap berada di tangan pengajar.

Secara keseluruhan, halaman landas (*landing page*) tersebut memasarkan platform penilaian ini sebagai asisten AI yang menghemat waktu, objektif, dan ramah bagi pengajar, serta mendukung praktik terbaik dalam pendidikan; *platform* ini juga menyinggung isu privasi, keandalan, dan keadilan, meskipun hanya sedikit detail teknis yang diberikan terkait aspek-aspek tersebut.

1) Hasil Evaluasi Etis

Dengan meninjau berbagai dimensi penerapan AI yang etis, kami mencatat hal-hal berikut:

- *Dokumentasi, kemampuan penjelasan (explainability), dan transparansi*: Platform ini menjelaskan alur data dasar namun tidak memberikan rincian mengenai model yang mendasarinya, data pelatihan, atau logika penilaian. Tidak jelas bagaimana skor AI dihasilkan atau penjelasan apa jika ada yang diberikan untuk membenarkan nilai individu. Dokumentasi teknis dan kemampuan penjelasannya terbatas.
- *Validitas dan keandalan*: Tidak ada bukti independen atau yang telah ditinjau oleh rekan sejawat (*peer-reviewed*) untuk membandingkan akurasi atau konsistensi nilai hasil AI dengan penilaian manusia yang ahli. Klaim platform ini didukung oleh testimoni pengguna, bukan oleh studi empiris atau evaluasi yang dipublikasikan mengenai keandalan atau tingkat kesalahan.
- *Keadilan dan bias*: Bahasa pemasaran menekankan pada "kesetaraan dalam penilaian" dan minimalisasi subjektivitas manusia, namun tidak ada bukti pembelajaran subkelompok atau audit keadilan/bias (misalnya, perbandingan kinerja berdasarkan ras/etnis, gender, atau kemahiran bahasa Inggris). Tidak ada studi mengenai fungsi

algoritma yang berbeda (*differential algorithmic functioning*) atau strategi mitigasi yang dijelaskan.

- *Pengawasan manusia*: Pengajar mengimpor tugas mahasiswa, menentukan rubrik penilaian, meninjau nilai dan umpan balik hasil AI, serta tetap memiliki kemampuan untuk menyesuaikan atau mengubah keputusan tersebut sebelum menyampaikannya kepada mahasiswa. Platform menegaskan bahwa "pengajar memegang kendali penuh atas nilai akhir." Selain itu, tidak ada mekanisme standar bagi mahasiswa untuk mengajukan banding atau mempertanyakan nilai.
- *Privasi dan keamanan*: Platform ini mengikuti langkah-langkah perlindungan data standar dan mengklaim kepatuhan terhadap FERPA dan NIST, namun mengandalkan pengajar dan sekolah untuk mengelola persetujuan (*consent*); sistem ini juga mengirimkan sebagian data kepada penyedia AI pihak ketiga.
- *Keterbatasan dan akuntabilitas*: Platform ini mengakui bahwa nilai dan umpan balik yang dihasilkan oleh AI mungkin mengandung kesalahan dan menyarankan pengajar untuk meninjau hasilnya. Namun, platform ini belum menjalani evaluasi ketat, pelaporan publik, ataupun audit independen terkait keandalan, keadilan, dan praktik privasinya. Keanggotaan dalam organisasi industri memang menunjukkan adanya niat baik, namun tanpa bukti konkret atau mekanisme akuntabilitas yang transparan, para pemangku kepentingan terpaksa hanya mengandalkan jaminan dari perusahaan. Hasil evaluasi dan rekomendasi langkah selanjutnya disajikan dalam Tabel 2.2.

Tabel 2.2 Hasil Protokol

Area Evaluasi	Status Platform	Rekomendasi Utama
Dokumentasi, kemampuan untuk dijelaskan, dan transparansi	Hanya mencakup gambaran umum; tidak memuat rincian teknis, dasar penilaian, maupun hasil audit	Merilis dokumentasi model dan laporan audit secara publik; menyajikan dasar pertimbangan pemberian skor
Validitas dan reliabilitas	Tidak ada laporan validasi empiris	Menugaskan/melaporkan studi perbandingan dengan penilai manusia
Keadilan dan bias	Tidak ada pembelajaran subkelompok atau pembelajaran bias yang dipublikasikan	Melakukan dan memublikasikan audit bias/keadilan secara berkala
Pengawasan manusia	Pengajar dapat melakukan peninjauan; tidak ada prosedur banding yang jelas	Meresmikan prosedur banding
Privasi dan keamanan	Mengklaim kepatuhan dasar; namun rinciannya tidak memadai	Mengungkapkan kebijakan secara rinci; memungkinkan audit oleh pihak ketiga
Keterbatasan dan akuntabilitas	Pernyataan penafian dasar mengenai kemungkinan	Melaporkan keterbatasan secara rutin; memungkinkan pemantauan independen dan

	kesalahan; tidak ada audit publik atau mekanisme pemulihan	menetapkan mekanisme pemulihan

Penerapan platform penilaian berbasis AI secara etis menuntut lebih dari sekadar komitmen umum dan afiliasi industri. Hal ini memerlukan dokumentasi yang konkret dan transparan mengenai logika sistem, keandalan, evaluasi keadilan, serta langkah-langkah perlindungan privasi. Meskipun platform yang ditinjau telah membuat kemajuan dalam mengungkapkan alur data dasar dan keterbatasannya, masih terdapat pertanyaan yang belum terjawab mengenai kemampuan menjelaskan hasil (*explainability*), ketangguhan ukuran keadilan dan keandalannya, serta memadainya perlindungan privasi yang diterapkan. Selama pengembang dan pihak yang menerapkan sistem belum menjawab kekhawatiran ini secara substansial melalui bukti terbuka dan validasi independen, sikap hati-hati sangat diperlukan sebelum mengandalkan sistem semacam itu untuk pengambilan keputusan pendidikan yang berdampak besar.

2.8 ARAH PENGEMBANGAN PENILAIAN AI YANG ETIS

Prinsip Utama

Integrasi AI ke dalam penilaian pendidikan menghadirkan peluang besar sekaligus tantangan yang berkelanjutan. Sebagaimana telah dibahas dalam bab ini, sistem berbasis AI memiliki kemampuan untuk memperkaya pengukuran pendidikan dengan meningkatkan efisiensi, memperluas akses, dan memfasilitasi bentuk-bentuk pembelajaran yang sebelumnya tidak terbayangkan. Namun, manfaat ini bergantung pada komitmen yang cermat dan berkelanjutan terhadap nilai-nilai etika inti: keadilan, transparansi, kemampuan untuk dijelaskan (*explainability*), akuntabilitas, privasi, dan pengelolaan yang bertanggung jawab.

Bias algoritmik baik yang berasal dari data historis, tertanam dalam desain, maupun muncul saat penerapan menimbulkan risiko yang terus ada, terutama bagi peserta didik dari latar belakang yang kurang terwakili atau kelompok minoritas. Transparansi dan kemampuan untuk dijelaskan tetap krusial dalam membangun serta mempertahankan kepercayaan para pemangku kepentingan; namun, kemajuan dalam kinerja teknis tidak boleh mengorbankan aspek pemahaman atau kemungkinan untuk melakukan peninjauan ulang (*contestability*).

Pengawasan manusia sangatlah penting: kita tidak boleh memosisikan pendidik dan administrator hanya sebagai konsumen pasif dari keluaran algoritma, melainkan sebagai mitra aktif yang mempertanyakan, menafsirkan, dan bila perlu memperbaiki keputusan yang dihasilkan melalui AI. Oleh karena itu, inovasi etis di bidang ini harus dipahami sebagai upaya multidimensi yang memadukan kewaspadaan teknologi, perumusan kebijakan, pelatihan institusional, penelitian empiris berkelanjutan, serta dialog terus-menerus dengan pihak-pihak yang paling terdampak oleh hasil penilaian.

Rekomendasi Penerapan

Berdasarkan pembelajaran dalam bab ini serta panduan terpadu dari berbagai kerangka kerja internasional dan organisasi di bidang pengukuran, kami mengajukan rekomendasi praktik berikut:

- **Berkomitmen pada pemantauan dan perbaikan berkelanjutan**: Sistem penilaian AI harus menjalani audit rutin terkait keadilan dan akurasi di seluruh subkelompok, serta dikalibrasi ulang seiring munculnya risiko baru atau perubahan dalam populasi peserta didik.
- **Mengutamakan transparansi dan penjelasan yang bermakna**: Merancang keluaran dan antarmuka yang tidak hanya menjamin keterbukaan informasi teknis, tetapi juga menyediakan penjelasan yang dapat ditindaklanjuti dan sesuai dengan kebutuhan audiens baik bagi pendidik, peserta didik, maupun keluarga.
- **Melembagakan pengawasan manusia**: Menyusun protokol yang jelas untuk peninjauan oleh manusia terhadap kasus-kasus yang berdampak besar atau ambigu, serta membekali seluruh staf yang bertugas melakukan pengawasan dengan literasi AI dan keterampilan interpretasi kritis.
- **Melibatkan pemangku kepentingan dalam perancangan dan evaluasi**: Mendorong proses inklusif yang mengangkat suara dari beragam peserta didik, praktisi, dan komunitas, mulai dari tahap perumusan masalah hingga penerapan dan pemantauan pascapeluncuran.
- **Mengedepankan privasi dan keamanan data**: Memandang keharusan etis dan hukum terkait privasi sebagai hal yang mutlak, menerapkan prinsip keamanan sejak tahap perancangan (*secure-by-design*), serta memastikan individu tetap memiliki kendali atas data pendidikan mereka sendiri.
- **Menetapkan mekanisme pemulihan hak**: Menyediakan jalur yang jelas dan mudah diakses untuk pengajuan banding dan koreksi termasuk peninjauan oleh pihak ketiga jika diperlukan agar pengembang dapat segera menangani kesalahan dan bias.

Arah Pengembangan Masa Depan

Meskipun telah dicapai kemajuan yang signifikan, masih banyak pertanyaan terbuka mengenai integrasi etis AI dalam penilaian pendidikan. Upaya di masa mendatang perlu difokuskan pada pengembangan metrik dan metode yang tangguh untuk menjamin keadilan dan transparansi, memahami dampak jangka panjang terhadap kesetaraan, serta meningkatkan kolaborasi antara sistem AI dan penilaian manusia. Memperluas jangkauan pengembangan profesional yang efektif serta menerjemahkan pedoman etika ke dalam kebijakan dan praktik sehari-hari juga merupakan hal yang krusial.

Pada akhirnya, potensi AI dalam penilaian tidak hanya bergantung pada inovasi teknis, tetapi juga pada komitmen kolektif kita terhadap tanggung jawab etis, refleksi berkelanjutan, dan pelibatan inklusif seluruh pemangku kepentingan. Dengan menempatkan etika sebagai praktik partisipatif yang terus berkembang, penilaian pendidikan dapat memajukan ilmu pengukuran sekaligus mewujudkan keadilan dalam pendidikan.

BAB 3
EVOLUSI PENILAIAN BERBASIS TEKNOLOGI AI

3.1 TRANSFORMASI PENILAIAN PENDIDIKAN

Bab ini mengeksplorasi peran transformatif kecerdasan buatan (AI) dalam pengukuran pendidikan, sebuah bidang yang telah mengalami perubahan signifikan selama satu abad terakhir. Penilaian pendidikan telah lama digunakan untuk mengevaluasi pengetahuan dan kemampuan mahasiswa, namun model-model tradisional kini semakin dipandang tidak memadai dalam menjawab tuntutan lanskap pendidikan abad ke-21. Integrasi AI ke dalam pengukuran pendidikan merupakan kemajuan teknologi sekaligus pergeseran mendasar dalam cara pendidik, pembuat kebijakan, dan peneliti memandang serta melaksanakan penilaian.

Bab ini menyajikan gambaran menyeluruh mengenai inovasi berbasis AI dalam pengukuran pendidikan, dengan menelaah konteks historis maupun signifikansi praktik penilaian di masa kini. Bab ini akan menyoroti bagaimana AI mengubah pendekatan tradisional terhadap pengujian serta menguraikan potensi munculnya model penilaian baru yang lebih dinamis. Model-model ini menjanjikan peningkatan akurasi, efisiensi, dan kesetaraan dalam pendidikan, sekaligus menjawab berbagai tantangan lama yang terkait dengan tes standar dan kerangka kerja penilaian tradisional.

Cakupan bab ini meliputi eksplorasi tema-tema krusial yang menjadi inti pemahaman mengenai dampak AI terhadap penilaian pendidikan. Pertama, bab ini akan membahas keterbatasan penilaian tradisional, seperti masalah terkait keadilan, aksesibilitas, dan fokus yang sempit pada pengetahuan akademis. Kedua, bab ini akan mengupas tantangan modern yang dihadapi institusi pendidikan, misalnya kebutuhan akan penilaian yang mampu mengukur keterampilan non-kognitif yang kompleks serta menyediakan pengalaman belajar yang dipersonalisasi. Terakhir, bab ini akan mengulas solusi inovatif yang dimungkinkan oleh AI, seperti pengujian adaptif, sistem umpan balik waktu nyata (*real-time*), dan algoritma pembelajaran mesin (*machine learning*) untuk pembahasan kumpulan data berskala besar.

3.2 PERKEMBANGAN PENILAIAN PENDIDIKAN

Penilaian pendidikan memiliki sejarah yang panjang dan kompleks, dengan akar yang berasal dari ujian lisan dan tertulis pada masyarakat kuno dan abad pertengahan. Pada abad ke-20, tes standar menjadi pilar utama sistem pendidikan di seluruh dunia, didorong oleh kebutuhan akan tolok ukur objektif mengenai pencapaian akademik. Tes-tes ini, seperti SAT dan ACT di Amerika Serikat, telah digunakan untuk mengambil keputusan krusial yang berdampak besar terhadap masa depan akademik dan profesional mahasiswa. Namun, pengujian standar menuai kritik karena ketidakmampuannya dalam mencakup keseluruhan spektrum kemampuan mahasiswa serta karena melanggengkan ketimpangan berdasarkan status sosial-ekonomi, ras, dan latar belakang budaya.

Menanggapi keterbatasan tersebut, para peneliti pendidikan dan pembuat kebijakan menyerukan model penilaian yang lebih komprehensif. Munculnya teknologi AI telah membuka peluang baru bagi pengukuran pendidikan, memungkinkan dilakukannya penilaian yang lebih efisien, personal, dan responsif terhadap kebutuhan belajar individu. Penilaian berbasis AI mampu pembahasan data dalam jumlah besar untuk memberikan wawasan yang sebelumnya sulit diperoleh, sehingga berpotensi merevolusi praktik pendidikan. Bab ini akan mengkaji bagaimana AI digunakan untuk mengatasi tantangan dalam penilaian tradisional dan menciptakan pengalaman belajar yang lebih bermakna. Bab ini juga akan membahas implikasi inovasi tersebut bagi mahasiswa, pendidik, dan pembuat kebijakan, dengan menekankan perlunya integrasi teknologi AI yang cermat ke dalam sistem pendidikan.

3.3 TRANSFORMASI PENILAIAN PENDIDIKAN MELALUI AI

Tema utama bab ini berkisar pada perbandingan antara penilaian tradisional dan peluang yang dihadirkan oleh AI. Penilaian tradisional, seperti tes standar, sering dikritik karena sifatnya yang seragam (menggunakan satu metode untuk semua) serta kegagalannya dalam mengukur keterampilan berpikir tingkat tinggi atau kemampuan non-kognitif, seperti kreativitas dan kecerdasan emosional. Penilaian semacam ini juga cenderung memperparah ketimpangan, mengingat mahasiswa dari latar belakang kurang beruntung mungkin tidak memiliki akses terhadap sumber daya dan kesempatan persiapan tes yang sama dengan rekan-rekan mereka yang lebih mampu.

Tantangan pendidikan modern menuntut solusi inovatif, seperti kebutuhan akan penilaian yang mampu beradaptasi dengan beragam gaya belajar dan memberikan umpan balik secara waktu nyata (*real-time*). AI memiliki potensi untuk memenuhi kebutuhan tersebut melalui pengujian adaptif yang menyesuaikan tingkat kesulitan soal berdasarkan kinerja mahasiswa serta melalui wawasan berbasis data yang menjadi landasan bagi pembelajaran yang dipersonalisasi. Selain itu, AI dapat mengembangkan asesmen yang mengukur kompetensi kompleks, seperti pemecahan masalah dan kolaborasi, yang sangat penting bagi keberhasilan di dunia kerja abad ke-21.

3.4 KONSEP DASAR PENILAIAN BERBASIS AI

Pemahaman yang jelas mengenai istilah dan konsep utama sangat penting untuk mengikuti pembahasan dalam bab ini. Penilaian pendidikan mengevaluasi pembelajaran mahasiswa secara sistematis melalui tes, tugas, atau metode lain untuk mengukur pengetahuan dan keterampilan. Penilaian tradisional mencakup tes standar dan evaluasi sumatif yang dilaksanakan pada akhir periode pembelajaran. Penilaian semacam ini telah banyak digunakan namun kini semakin dipandang memiliki keterbatasan dalam hal cakupan dan efektivitas. Sebaliknya, metode penilaian inovatif mencakup berbagai praktik yang bertujuan memberikan pemahaman lebih komprehensif mengenai pembelajaran mahasiswa. Metode-metode ini meliputi penilaian formatif yang memantau kemajuan mahasiswa dan memandu proses pembelajaran serta penilaian berbasis kinerja (*performance-based*

assessments atau PBA) yang mengharuskan mahasiswa menunjukkan keterampilan mereka melalui tugas-tugas dunia nyata.

Penilaian berbasis AI merupakan bagian dari metode inovatif yang memanfaatkan teknologi untuk meningkatkan praktik penilaian. Metode ini mencakup perangkat seperti tes adaptif terkomputerisasi (*computerized adaptive testing* atau CAT) dan penilaian esai otomatis, yang menggunakan algoritma pembelajaran mesin (*machine learning*) untuk pembahasan jawaban mahasiswa dan memberikan umpan balik. Kerangka pembelajaran kritis dalam bab ini akan mengkaji manfaat dan keterbatasan penilaian berbasis AI. Bab ini akan membahas aspek etika dan praktis, seperti privasi data, bias algoritma, dan perlunya transparansi dalam sistem AI. Selain itu, bab ini akan membahas peran AI dalam mendorong kesetaraan pendidikan, serta menelaah peluang dan tantangan terkait penerapan teknologi ini dalam skala besar.

3.5 PERKEMBANGAN PENILAIAN PENDIDIKAN

Bentuk-Bentuk Awal Penilaian

Akar penilaian pendidikan dapat ditelusuri hingga ribuan tahun yang lalu, dengan beragam bentuk dan fungsi yang muncul di berbagai peradaban. Penilaian pada masa awal umumnya bersifat informal dan berpusat pada demonstrasi keterampilan praktis serta tanya jawab lisan. Sebagai contoh, di Mesir Kuno, sekolah juru tulis menggunakan metode pelafalan lisan untuk mengevaluasi penguasaan mahasiswa terhadap tulisan hieroglif. Mahasiswa diharuskan menghafal dan mereproduksi teks-teks yang rumit, dan kemahiran mereka dinilai oleh juru tulis yang lebih berpengalaman; hal ini mencerminkan bentuk evaluasi yang sangat subjektif namun efektif dalam konteksnya.

Di Yunani Kuno, pendidikan sangat dipengaruhi oleh metode Socrates, yang menekankan dialog lisan untuk merangsang pemikiran kritis. Diskusi-diskusi ini berfungsi baik sebagai sarana pengajaran maupun penilaian, karena respons mahasiswa terhadap pertanyaan-pertanyaan yang menggali pemahaman mereka mengungkapkan sejauh mana mereka menguasai konsep-konsep filosofis dan matematis. Demikian pula di Eropa abad pertengahan, penilaian pembelajaran sangat erat kaitannya dengan sistem magang, di mana demonstrasi keterampilan secara praktis bukan sekadar pengetahuan akademis menjadi fokus utama. Pendekatan ini tidak terbatas pada bidang kerajinan dan perdagangan saja, tetapi juga mencakup profesi seperti kedokteran dan hukum; dalam bidang-bidang tersebut, tingkat penguasaan dinilai melalui ujian lisan yang dikenal sebagai *disputation* atau debat ilmiah.

Di Eropa abad pertengahan, model magang menjadi sistem yang dominan. Evaluasi dilakukan berdasarkan unjuk kerja, dengan fokus utama pada penguasaan keterampilan praktis dalam bidang kerajinan dan perdagangan. Kemajuan seorang pemagang dinilai berdasarkan kemampuannya meniru keterampilan sang ahli (*master*), yang puncaknya adalah pembuatan karya akhir yang disebut sebagai "*master-piece*" (karya mahakarya); karya ini digunakan untuk menentukan apakah pemagang tersebut layak menjadi seorang ahli atau pengrajin utama.

Perubahan dari penilaian lisan ke penilaian tertulis menandai pergeseran mendasar dalam praktik pendidikan. Sistem ujian kekaisaran Tiongkok, yang bermula sejak tahun 605 M pada masa Dinasti Sui, merupakan salah satu contoh paling awal dari ujian tertulis berskala besar. Dirancang untuk menyeleksi kandidat bagi jabatan pegawai negeri, ujian-ujian ini menguji pengetahuan tentang naskah klasik Konfusianisme dan menjadi model bagi sistem tes terstandarisasi. Pada abad ke-19, para pendidik di Eropa dan Amerika mulai mengadopsi metode serupa, sehingga penggunaan ujian tertulis meluas di sekolah maupun universitas.

Munculnya ujian tertulis dapat ditelusuri kembali ke Eropa dan Amerika pada abad ke-19. Terinspirasi oleh sistem ujian kekaisaran Tiongkok, sistem pendidikan di Eropa dan Amerika mulai mengadopsi penilaian tertulis untuk menstandarisasi kualifikasi dan menciptakan kriteria objektif bagi kemajuan akademis. Menjelang akhir tahun 1800-an, ujian tertulis telah menjadi hal yang umum dalam pendidikan menengah maupun tinggi. Pergeseran ini menandai perubahan signifikan dari metode evaluasi lisan yang bersifat personal di masa lalu, sekaligus meletakkan dasar bagi sistem penilaian yang lebih formal dan terstruktur.

Periode ini juga menandai formalisasi penilaian melalui pengembangan sistem pemberian nilai dan penggunaan soal esai untuk mengukur kemampuan analitis. Menjelang akhir abad ke-19, universitas-universitas di Amerika Serikat, seperti Harvard dan Yale, menggunakan ujian tertulis sebagai kriteria utama penerimaan mahasiswa; hal ini menandai peralihan menuju pendekatan evaluasi yang lebih terstruktur dan terstandarisasi. Perubahan-perubahan tersebut membuka jalan bagi munculnya tes standar modern pada abad ke-20, yang kemudian mengubah praktik penilaian pendidikan di seluruh dunia.

Perkembangan Tes Standar

Munculnya tes standar pada abad ke-20 berawal dari perkembangan pesat bidang psikologi pendidikan serta upaya untuk menciptakan alat ukur yang objektif bagi kecerdasan dan bakat akademik. Karya Alfred Binet dan Théodore Simon pada awal tahun 1900-an meletakkan dasar bagi tes kecerdasan modern. Tujuan mereka adalah menciptakan instrumen yang dapat mengidentifikasi mahasiswa yang membutuhkan dukungan pendidikan khusus, yang kemudian menghasilkan pengembangan skala Binet-Simon pada tahun 1905. Lewis Terman, seorang psikolog asal Amerika, mengadaptasi skala tersebut untuk menciptakan Skala Kecerdasan Stanford-Binet pada tahun 1916, sekaligus memperkenalkan konsep *intelligence quotient* (IQ) atau tingkat kecerdasan.

Penerapan tes IQ secara luas di Amerika Serikat pada awal abad ke-20 bertepatan dengan meningkatnya keyakinan akan kekuatan sains dan pengukuran dalam memajukan pendidikan dan masyarakat. Penggunaan tes kecerdasan untuk mengklasifikasikan dan mengelompokkan mahasiswa menjadi tren utama yang memengaruhi kebijakan serta praktik pendidikan. Tokoh penting lainnya, Robert Yerkes, mengembangkan tes Army Alpha dan Beta selama Perang Dunia I untuk mengevaluasi kemampuan kognitif para calon anggota militer. Tes-tes ini menunjukkan potensi penilaian terstandar dalam mengevaluasi kelompok besar secara efisien, sehingga semakin memperkuat peran penilaian tersebut dalam pengambilan keputusan di bidang pendidikan dan masyarakat.

Pada masa pascaperang, cakupan tes terstandar meluas dari sekadar tes kecerdasan hingga mencakup pengukuran bakat dan prestasi akademik. Pengembangan SAT oleh *College Entrance Examination Board* pada tahun 1926 merupakan tonggak sejarah penting dalam bidang penilaian pendidikan. SAT bertujuan menyediakan tolok ukur standar untuk membandingkan mahasiswa dari berbagai latar belakang pendidikan, sehingga turut mendemokratisasi akses ke perpengajaran tinggi. Sementara itu, ACT yang diperkenalkan pada tahun 1959 hadir sebagai alternatif dengan lebih menitikberatkan pada prestasi akademik dibandingkan bakat bawaan.

Maraknya penggunaan tes terstandar tidak lepas dari kontroversi. Para kritikus berpendapat bahwa tes-tes tersebut memperparah ketimpangan sosial karena lebih mengutamakan pengetahuan dan keterampilan kognitif tertentu dibandingkan yang lain. Kendati demikian, tes terstandar mendapatkan penerimaan luas berkat objektivitas, reliabilitas, dan efisiensinya. Menjelang pertengahan abad ke-20, tes terstandar telah menjadi elemen sentral dalam sistem pendidikan Amerika, yang memengaruhi berbagai aspek mulai dari pengelompokan mahasiswa dan perancangan kurikulum hingga penerimaan mahasiswa serta kebijakan pendidikan.

Tonggak Perkembangan Penilaian

Periode pertengahan hingga akhir abad ke-20 menyaksikan perkembangan signifikan dalam penilaian pendidikan, yang didorong oleh perubahan kebijakan pendidikan, kemajuan dalam bidang psikometri, serta meningkatnya sorotan publik terhadap hasil pendidikan. Salah satu perkembangan yang paling berdampak adalah munculnya sistem penilaian berisiko tinggi (*high-stakes testing*), yang mengaitkan hasil tes dengan pendanaan pendidikan, langkah-langkah akuntabilitas, dan keputusan kebijakan.

Undang-Undang *No Child Left Behind* (NCLB), yang disahkan menjadi undang-undang oleh Presiden George W. Bush pada tahun 2002, merupakan contoh nyata dari tren ini. NCLB mewajibkan pelaksanaan tes tahunan dalam bidang membaca dan matematika bagi siswa kelas 3 hingga 8 serta satu kali di tingkat sekolah menengah atas, dengan hasil tes yang digunakan untuk mengevaluasi kinerja sekolah. Sekolah yang gagal mencapai kemajuan tahunan yang memadai (*adequate yearly progress*) menghadapi sanksi, termasuk hilangnya pendanaan dan potensi restrukturisasi. Kebijakan ini menandai pergeseran ke arah penggunaan penilaian standar untuk mengukur pembelajaran mahasiswa serta sebagai sarana untuk menegakkan akuntabilitas dan mendorong reformasi pendidikan.

Era NCLB menyoroti baik kekuatan maupun keterbatasan dari tes berisiko tinggi (*high-stakes testing*). Meskipun para pendukung berpendapat bahwa kebijakan tersebut memberikan perhatian yang sangat dibutuhkan terhadap kesenjangan prestasi dan ketimpangan pendidikan, para kritikus berpendapat bahwa kebijakan itu memicu praktik pengajaran yang hanya berorientasi pada tes (*teaching to the test*), menyempitkan kurikulum, serta meningkatkan tekanan bagi mahasiswa dan pendidik. Perdebatan ini terus mewarnai diskusi mengenai peran penilaian standar dalam dunia pendidikan.

Pada saat yang sama, penilaian internasional seperti *Programme for International Student Assessment* (PISA) dan *Trends in International Mathematics and Science Study* (TIMSS)

mulai mengemuka. PISA, yang pertama kali diselenggarakan oleh *Organisation for Economic Cooperation and Development* (OECD) pada tahun 2000, menilai kemampuan remaja berusia 15 tahun dalam menerapkan pengetahuan pada konteks dunia nyata. TIMSS, yang dikelola oleh *International Association for the Evaluation of Educational Achievement*, mengevaluasi pengetahuan matematika dan sains mahasiswa kelas 4 dan 8.

Penilaian internasional ini telah memengaruhi kebijakan pendidikan dengan menyediakan tolok ukur untuk membandingkan kinerja sistem pendidikan antarnegara. Negara-negara yang memperoleh hasil rendah dalam PISA atau TIMSS sering kali menghadapi tekanan untuk mereformasi sistem pendidikan mereka, yang berujung pada perubahan standar kurikulum, pelatihan pengajar, dan praktik penilaian. Penyebaran global penilaian standar juga memunculkan pertanyaan mengenai bias budaya, validitas perbandingan lintas negara, serta dampak penilaian tersebut terhadap kesetaraan pendidikan.

Perkembangan Penilaian Kontempore

Abad ke-21 telah menghadirkan tantangan dan peluang baru bagi penilaian pendidikan. Kemajuan teknologi telah mengubah praktik penilaian, memungkinkan pengembangan tes berbasis komputer dan tes adaptif. Inovasi-inovasi ini memungkinkan pengukuran kemampuan mahasiswa yang lebih presisi, pemberian umpan balik secara waktu nyata (*real-time*), serta kemampuan untuk menilai keterampilan kognitif yang kompleks. Penilaian berbasis komputer, seperti yang digunakan dalam PISA sejak tahun 2015, menawarkan cara baru untuk mengevaluasi keterampilan pemecahan masalah dan kolaborasi mahasiswa, yang kian dihargai dalam dunia kerja modern.

Terlepas dari kemajuan tersebut, peran tes standar dalam pendidikan masih menjadi perdebatan. Para kritikus berpendapat bahwa tes standar sering kali gagal menangkap keseluruhan kemampuan mahasiswa, khususnya keterampilan yang sangat kreatif dan keterampilan non-kognitif. Munculnya gerakan "*opt-out*" (menolak partisipasi) di Amerika Serikat di mana orang tua menolak mengizinkan anak-anak mereka mengikuti tes standar mencerminkan meningkatnya ketidakpuasan terhadap kondisi penilaian pendidikan saat ini.

Sebagai tanggapan, para pendidik dan pembuat kebijakan mulai menjajaki pendekatan penilaian alternatif. Penilaian formatif, yang memberikan umpan balik berkelanjutan untuk mendukung pembelajaran alih-alih sekadar mengukurnya, semakin banyak diterapkan guna menciptakan sistem penilaian yang lebih seimbang dan suportif. Demikian pula, penilaian berbasis kinerja (*Performance-Based Assessments* atau PBA) yang mengharuskan mahasiswa menunjukkan pemahaman mereka melalui proyek, presentasi, dan tugas-tugas dunia nyata kini mulai diintegrasikan ke dalam sejumlah sistem pendidikan.

` Masa depan penilaian pendidikan kemungkinan besar akan melibatkan kombinasi antara pendekatan tradisional dan inovatif, yang menyeimbangkan kebutuhan akan standardisasi dengan pengakuan bahwa pendidikan merupakan proses yang kompleks dan memiliki banyak dimensi. Seiring dengan terus berkembangnya sistem pendidikan, tantangannya adalah mengembangkan penilaian yang adil, valid, dan selaras dengan tujuan pendidikan yang lebih luas di dunia yang berubah dengan cepat.

3.6 KETERBATASAN PENILAIAN KONVENSIONAL

Penilaian konvensional, yang sering kali dicirikan oleh tes standar, telah menjadi landasan evaluasi pendidikan selama beberapa dekade. Namun, keterbatasannya semakin nyata dalam lanskap pendidikan yang kompleks saat ini. Keterbatasan ini dapat dikelompokkan ke dalam tiga kategori utama: kekhawatiran mengenai akurasi dan validitas, masalah kesetaraan dan aksesibilitas, serta pertanyaan mengenai relevansi dalam konteks pendidikan abad ke-21. Bab ini mengupas tuntas keterbatasan tersebut dengan merujuk pada berbagai sumber ilmiah yang menyoroti perlunya praktik penilaian yang lebih efektif dan inklusif.

Akurasi dan Validitas Penilaian

Penilaian konvensional sering kali gagal mencerminkan kemampuan mahasiswa secara akurat. Fokus pada skor tes standar sebagai tolok ukur pembelajaran mahasiswa, efektivitas pengajar, dan kinerja sekolah telah banyak dikritik. Menurut studi, tes standar sering kali hanya mengukur cakupan pengetahuan mahasiswa yang sempit, serta mengabaikan beragam keterampilan yang diperlukan untuk meraih kesuksesan di dunia nyata. Ketidakselarasan ini menimbulkan pertanyaan mengenai validitas penilaian tersebut sebagai alat evaluasi pendidikan.

1) Ketidaksesuaian antara Hasil Penilaian dan Kompetensi Dunia Nyata

Tujuan utama penilaian pendidikan seharusnya adalah mengukur sejauh mana mahasiswa telah memperoleh kompetensi yang relevan di luar ruang kelas. Namun, sebagaimana yang dikemukakan, paradigma penilaian yang dominan lebih mengutamakan kemampuan mengingat pengetahuan dan keterampilan prosedural dasar dibandingkan berpikir kritis dan kreativitas. Fokus yang sempit ini dapat berujung pada praktik "mengajar demi tes" (*teaching to the test)*, di mana waktu pembelajaran dialihkan untuk persiapan tes alih-alih untuk pembelajaran yang lebih mendalam.

Pergeseran ini menghambat pengembangan keterampilan yang dapat diterapkan dalam berbagai situasi (*transferable skills*), seperti kolaborasi, inovasi, dan pemecahan masalah, yang semakin diakui sebagai hal krusial bagi kesuksesan di abad ke-21. Situasi ini semakin rumit karena berbagai studi menunjukkan bahwa hasil tes standar sering kali memiliki korelasi yang lebih kuat dengan latar belakang sosial-ekonomi mahasiswa dibandingkan dengan kemampuan akademik mereka yang sebenarnya. Hubungan ini mengisyaratkan bahwa apa yang diukur mungkin lebih mencerminkan faktor eksternal seperti akses terhadap sumber daya dan dukungan daripada kompetensi yang telah dikembangkan oleh mahasiswa.

2) Masalah Desain Tes dan Kelemahan Psikometrik

Desain dan karakteristik psikometrik dari asesmen konvensional juga menghadirkan tantangan yang signifikan. Tes tradisional didasarkan pada teori tes klasik yang mengutamakan reliabilitas dan standardisasi. Namun, studi menyoroti bahwa pendekatan ini dapat menyebabkan *construct underrepresentation* (kurangnya representasi konstruk) dan *construct-irrelevant variance* (varians yang tidak relevan dengan konstruk), di mana aspek-aspek penting dari kompetensi terabaikan atau faktor-faktor yang tidak relevan memengaruhi skor tes. Sebagai contoh, tingkat keakraban mahasiswa dengan format tes atau bahasa yang

digunakan dalam butir soal dapat memengaruhi kinerja mereka secara tidak adil, sehingga skor yang diperoleh tidak benar-benar mencerminkan kemampuan mereka yang sesungguhnya.

Masalah lain yang menjadi perhatian adalah adanya bias pada butir soal, di mana pertanyaan tes mungkin lebih menguntungkan kelompok demografis tertentu. Misalnya, studi menemukan bahwa mahasiswa yang sedang mempelajari bahasa Inggris (*English Language Learners* atau ELL) sering kali memperoleh skor lebih rendah pada tes standar; hal ini terjadi bukan karena kurangnya penguasaan materi, melainkan akibat hambatan linguistik yang terkandung dalam butir-butir soal tes tersebut. Kelemahan-kelemahan ini mengganggu keadilan dan validitas asesmen, sehingga menjadikannya alat yang kurang memadai untuk mengevaluasi populasi mahasiswa yang beragam.

Selain itu, lingkungan pengujian yang memiliki konsekuensi tinggi (*high-stakes testing*) turut memicu fenomena yang dikenal sebagai "kecemasan menghadapi tes" (*test anxiety*). Kondisi ini dapat berdampak secara tidak proporsional terhadap kinerja mahasiswa, khususnya bagi mereka yang berasal dari latar belakang terpinggirkan atau memiliki kesulitan belajar. Kecemasan semacam itu dapat mendistorsi hasil tes, sehingga menghasilkan gambaran yang tidak akurat mengenai kemampuan mahasiswa yang sebenarnya dan memperparah ketimpangan pendidikan yang sudah ada.

Kesetaraan dan Aksesibilitas Penilaian

Kesetaraan dan aksesibilitas merupakan pertimbangan krusial dalam asesmen pendidikan. Asesmen konvensional sering kali gagal mengakomodasi keragaman latar belakang dan kebutuhan mahasiswa, sehingga menimbulkan kesenjangan kinerja yang mencerminkan ketimpangan sosial yang lebih luas. Menurut studi, faktor sosial-ekonomi seperti akses terhadap sumber daya persiapan tes memainkan peran penting dalam menentukan hasil kinerja mahasiswa, yang pada gilirannya melanggengkan siklus ketimpangan.

1) Kesenjangan Sosioekonomi dalam Persiapan Tes dan Akses

Mahasiswa dari keluarga mampu memiliki akses ke berbagai sumber daya, termasuk bimbingan belajar privat, kursus persiapan tes, dan kegiatan ekstrakurikuler yang menunjang prestasi akademik. Sebaliknya, mahasiswa dari latar belakang ekonomi rendah sering kali tidak memiliki dukungan tersebut, sehingga menimbulkan kesenjangan kinerja yang lebih mencerminkan ketimpangan ekonomi daripada potensi akademik. Masalah ini diperparah oleh "Efek Matthew" (*Matthew Effect*), di mana mahasiswa yang mengawali dengan keuntungan tertentu akan terus mengakumulasi lebih banyak pencapaian pendidikan seiring berjalannya waktu, sementara mereka yang tidak memiliki keuntungan tersebut justru semakin tertinggal.

Selain itu, meningkatnya komersialisasi layanan persiapan tes telah memperparah kesenjangan ini. Berbagai studi menunjukkan bahwa partisipasi dalam program persiapan tes intensif dapat meningkatkan skor secara signifikan, sehingga memberikan keunggulan kompetitif bagi mereka yang mampu membiayainya. Dinamika ini memunculkan kekhawatiran etis mengenai keadilan penggunaan tes standar untuk pengambilan keputusan yang berdampak besar, seperti penerimaan perpengajaran tinggi dan alokasi beasiswa.

2) Bias terhadap Peserta Didik Beragam

Penilaian konvensional sering kali perlu diselaraskan kembali dengan kebutuhan peserta didik non-tradisional, termasuk mahasiswa neurodivergen dan ELL (*English Language Learners* atau pembelajar bahasa Inggris). Kelompok-kelompok ini sering menghadapi tes yang tidak mengakomodasi profil pembelajaran unik mereka. Sebagai contoh, mahasiswa neurodivergen seperti mereka yang memiliki ADHD atau disleksia mungkin mengalami kesulitan dengan format tes standar yang mengutamakan kecepatan dan ketepatan, yang pada akhirnya menyebabkan kemampuan mereka dinilai lebih rendah dari yang sebenarnya.

Bagi ELL, kompleksitas linguistik dalam butir soal dapat mengaburkan pemahaman materi, sehingga menghasilkan skor yang lebih mencerminkan kemahiran berbahasa daripada penguasaan materi. Bias semacam itu menyoroti perlunya desain penilaian yang lebih inklusif, yang mampu mengakomodasi keberagaman peserta didik serta mencerminkan pengetahuan dan keterampilan mereka secara akurat.

Relevansi Penilaian Abad ke-21

Relevansi penilaian konvensional semakin dipertanyakan seiring dengan perubahan lanskap pendidikan. Saat masyarakat beralih menuju ekonomi berbasis pengetahuan, kemampuan berpikir kritis, literasi digital, dan pemecahan masalah secara kolaboratif menjadi semakin penting. Namun, sebagian besar tes standar masih berakar pada model pendidikan abad ke-20 yang lebih mengutamakan pengetahuan statis dibandingkan kompetensi-kompetensi dinamis tersebut.

1) Kerangka Penilaian Usang yang Mengabaikan Literasi Digital dan Berpikir Kritis

Kerangka penilaian saat ini harus lebih memperhatikan literasi digital, kreativitas, dan keterampilan abad ke-21 lainnya yang krusial bagi keberhasilan di dunia yang berubah dengan cepat. Sebagai contoh, meskipun perangkat digital telah lazim digunakan di tempat kerja modern, banyak penilaian masih mengandalkan format kertas dan pensil yang tidak mengukur kemampuan mahasiswa dalam menavigasi dan pembahasan informasi menggunakan sumber daya digital. Kesenjangan ini telah memicu seruan untuk menciptakan model penilaian baru yang mampu mencakup spektrum keterampilan yang utuh, yang dibutuhkan mahasiswa agar dapat berkembang di era digital.

2) Penekanan Berlebihan pada Hafalan dan Pengetahuan yang Terlepas dari Konteks

Penilaian konvensional juga terlalu menekankan pada hafalan dan pengetahuan yang terlepas dari konteks praktik yang kini semakin dipandang tidak memadai sebagai tolok ukur pembelajaran mahasiswa. Sebagaimana yang dikemukakan, penilaian yang hanya berfokus pada kemampuan mengingat materi gagal menangkap kemampuan mahasiswa dalam menerapkan pengetahuan secara bermakna. Fokus pada keterampilan kognitif tingkat rendah ini berkontribusi pada rendahnya keterlibatan dan motivasi mahasiswa, sehingga mengurangi nilai edukatif dari penilaian-penilaian tersebut.

Singkatnya, keterbatasan penilaian konvensional terlihat jelas dari ketidakmampuannya dalam mengukur kompetensi mahasiswa secara akurat, dampaknya yang tidak adil terhadap populasi mahasiswa yang beragam, serta kurangnya relevansi dengan pendidikan modern. Mengatasi keterbatasan ini menuntut peralihan ke model penilaian yang

lebih holistik dan inklusif, yang mengutamakan validitas, kesetaraan, dan relevansi. Seiring dengan berkembangnya sistem pendidikan, terdapat kebutuhan mendesak untuk mengembangkan pendekatan baru yang lebih selaras dengan kompleksitas proses belajar-mengajar di abad ke-21.

3.7 PENDEKATAN KONTEMPORER DALAM PENILAIAN PENDIDIKAN

Penilaian Formatif dan Diagnostik

Penilaian formatif dan diagnostik telah menjadi instrumen penting dalam praktik pendidikan modern. Berbeda dengan penilaian sumatif tradisional yang mengevaluasi pembelajaran mahasiswa di akhir periode pengajaran, penilaian formatif dilaksanakan sepanjang proses pembelajaran untuk memberikan umpan balik yang dapat digunakan guna memperbaiki kualitas pengajaran maupun pembelajaran. Di sisi lain, penilaian diagnostik digunakan pada awal unit pembelajaran atau tahun akademik untuk mengidentifikasi pengetahuan, keterampilan, dan miskonsepsi yang telah dimiliki mahasiswa. Secara bersama-sama, pendekatan-pendekatan ini menciptakan sistem umpan balik dinamis yang memungkinkan pengajar untuk menyesuaikan pengajaran agar dapat memenuhi kebutuhan setiap mahasiswa secara individual.

1) Penggunaan Penilaian Berkelanjutan untuk Mendukung Pembelajaran

Penilaian formatif yang berkelanjutan merupakan bagian integral dari pembelajaran yang efektif. Teknik-teknik seperti kuis, peta konsep, *exit ticke*t (kartu keluar kelas), dan jurnal reflektif memungkinkan pengajar untuk mengukur pemahaman mahasiswa secara langsung (*real-time*) dan menyesuaikan strategi pengajaran mereka. Sebagai contoh, peta konsep membantu pengajar memvisualisasikan pemahaman mahasiswa mengenai suatu topik dan mengidentifikasi kesenjangan pengetahuan. Informasi ini kemudian dapat menjadi acuan bagi pembelajaran selanjutnya untuk mengatasi area-area kelemahan tersebut.

Umpan balik berkelanjutan yang diberikan melalui penilaian formatif mendorong pola pikir berkembang (*growth mindset*) pada mahasiswa dengan menekankan pembelajaran sebagai sebuah proses. Alih-alih hanya berfokus pada nilai, mahasiswa termotivasi untuk melakukan perbaikan melalui umpan balik yang konstruktif, yang sekaligus menumbuhkan kemampuan pengaturan diri dan keterampilan metakognitif. Penilaian diagnostik memainkan peran penting dalam mengidentifikasi kebutuhan belajar yang spesifik. Misalnya, dalam lingkungan pendidikan K–12, tes diagnostik dapat mengidentifikasi bidang-bidang di mana mahasiswa mungkin memerlukan dukungan tambahan seperti keterampilan dasar matematika atau pemahaman bacaan sehingga memungkinkan dilakukannya intervensi yang tepat sasaran.

2) Implementasi Penilaian Formatif dan Diagnostik

Berbagai studi kasus menyoroti keberhasilan penerapan asesmen formatif dan diagnostik di jenjang pendidikan K–12 maupun pendidikan tinggi. Dalam sebuah studi, sekelompok pengajar sekolah menengah pertama mengintegrasikan asesmen formatif ke dalam pembelajaran matematika dan sains mereka, yang secara signifikan meningkatkan prestasi mahasiswa. Contoh lain ditemukan di pendidikan tinggi, di mana Universitas

Edinburgh menerapkan asesmen formatif dalam mata kuliah teknik. Para dosen menggunakan kuis mingguan dan kegiatan penilaian antar-teman sejawat (*peer-assessment*), yang menghasilkan peningkatan nyata dalam keterlibatan serta pemahaman mahasiswa terhadap konsep-konsep inti.

Demikian pula, inisiatif "*Formative Assessment for Teachers*" di Sekolah Negeri Kota New York melatih para pengajar untuk mengintegrasikan asesmen diagnostik ke dalam pembelajaran sehari-hari. Program tersebut menunjukkan bahwa mahasiswa di kelas yang menerapkan asesmen formatif secara rutin memiliki kinerja lebih baik dalam tes standar, yang mencerminkan potensi asesmen ini dalam mendorong pembelajaran yang bermakna.

Asesmen Berbasis Kinerja

Asesmen Berbasis Kinerja (PBA) mewakili pergeseran dari pengujian tradisional menuju evaluasi kemampuan mahasiswa dalam menerapkan pengetahuan pada situasi dunia nyata. PBA mengharuskan mahasiswa untuk terlibat dalam tugas-tugas yang menunjukkan pengetahuan dan keterampilan mereka, seperti proyek, presentasi, dan portofolio. Asesmen ini mendorong kemampuan berpikir tingkat tinggi dan kreativitas, karena mahasiswa sering kali dituntut untuk pembahasan, menyintesis, dan mengevaluasi informasi, alih-alih sekadar mengingat fakta.

1) Pembelajaran Berbasis Proyek dan Asesmen Portofolio

Pembelajaran berbasis proyek (PBL) telah menjadi strategi pembelajaran populer yang sangat selaras dengan PBA. Dalam PBL, mahasiswa mengerjakan proyek dunia nyata yang kompleks dalam jangka waktu yang cukup lama, yang bermuara pada produk akhir atau presentasi. Metode ini mengembangkan penguasaan materi serta membangun keterampilan penting seperti kolaborasi, pemecahan masalah, dan manajemen waktu. Pengajar menilai mahasiswa berdasarkan keterlibatan mereka dalam proyek, kualitas produk akhir, dan proses refleksi.

Asesmen portofolio memberikan gambaran menyeluruh mengenai kemajuan mahasiswa dari waktu ke waktu. Dalam pendekatan ini, mahasiswa mengumpulkan dan mengurasi sejumlah karya mereka, serta melakukan refleksi terhadap perkembangan dan perjalanan belajar mereka. Portofolio memberikan gambaran yang lebih kaya dan menyeluruh mengenai kemampuan mahasiswa, dengan mencakup aspek kekuatan maupun area yang perlu ditingkatkan. Metode ini sangat efektif untuk mata pelajaran yang menuntut ekspresi kreatif seperti seni, penulisan, dan desain serta banyak diterapkan dalam jenjang pendidikan K–12 maupun pendidikan tinggi.

2) Desain Rubrik dan Kriteria Evaluasi

Komponen penting dari PBA (Penilaian Berbasis Kinerja) yang efektif adalah penggunaan rubrik yang dirancang dengan baik. Rubrik berfungsi sebagai panduan bagi pengajar dan mahasiswa dengan menyediakan kriteria yang jelas untuk setiap tingkat kinerja. Rubrik harus bersifat spesifik, adil, dan selaras dengan tujuan pembelajaran, sehingga membantu menstandarisasi penilaian dan mengurangi subjektivitas. Sebagai contoh, rubrik untuk proyek sains dapat mencakup kriteria seperti "menunjukkan pemahaman tentang prinsip-prinsip ilmiah," "penerapan penelitian," dan "kreativitas dalam penyajian."

Pengembangan rubrik melibatkan identifikasi keterampilan dan bidang pengetahuan esensial yang ingin diukur oleh penilaian tersebut. Rubrik yang disusun dengan baik membantu mahasiswa memahami ekspektasi dan memungkinkan mereka melakukan penilaian mandiri terhadap pekerjaan mereka. Rubrik juga memfasilitasi pemberian umpan balik yang bermakna, karena pengajar dapat memberikan komentar spesifik pada setiap kriteria, yang membimbing mahasiswa menuju perbaikan.

Penilaian Digital dan Adaptif

Teknologi digital telah merevolusi penilaian pendidikan dengan memungkinkan pelaksanaan tes yang dipersonalisasi dan adaptif. Penilaian berbasis komputer menawarkan fleksibilitas, interaktivitas, dan kemampuan untuk mengumpulkan data kinerja mahasiswa guna memberikan informasi bagi proses belajar-mengajar secara waktu nyata (*real-time*). Salah satu bentuk penilaian digital yang paling inovatif adalah CAT (*Computerized Adaptive Testing*), yang menyesuaikan tingkat kesulitan soal berdasarkan respons Mahasiswa.

1) Penggunaan Teknologi untuk Tes yang Dipersonalisasi dan Adaptif

Tes adaptif memberikan pengalaman penilaian yang lebih bersifat individual dengan menyesuaikan tingkat kesulitan soal sebagai respons terhadap setiap jawaban. Tes adaptif memungkinkan pengukuran kemampuan mahasiswa yang lebih presisi, khususnya pada mata pelajaran dengan variasi tingkat keterampilan yang luas, seperti matematika dan seni bahasa. Sebagai contoh, jika mahasiswa menjawab soal dengan benar, soal berikutnya mungkin lebih menantang, sedangkan jawaban yang salah dapat mengarah pada soal yang lebih mudah. Kemampuan adaptasi ini memberikan gambaran yang lebih akurat mengenai tingkat keterampilan mahasiswa dan mengurangi kecemasan saat tes.

Penilaian digital semakin banyak digunakan dalam tes standar, seperti terlihat pada program GRE dan GMAT, di mana model CAT memberikan pengalaman tes yang efisien sekaligus menjaga akurasi. Dalam pendidikan K–12, perangkat seperti asesmen diagnostik i-Ready menyesuaikan diri dengan respons setiap mahasiswa, sehingga memberikan wawasan mendetail kepada pendidik mengenai kekuatan dan kelemahan mahasiswa. Wawasan ini memungkinkan dilakukannya intervensi yang terarah dan pembelajaran berdiferensiasi, yang mendukung pengalaman belajar yang lebih personal.

2) Potensi Asesmen berbasis AI untuk Umpan Balik Real-Time

AI memiliki potensi untuk mentransformasi asesmen pendidikan lebih jauh dengan menyediakan umpan balik berbasis data secara *real-time*. *Platform* berbasis AI, seperti sistem pembelajaran adaptif, terus-menerus pembahasan kinerja mahasiswa dan memberikan panduan yang dipersonalisasi berdasarkan kebutuhan individu. Sebagai contoh, perangkat asesmen berbasis AI dapat mendeteksi pola kesalahan mahasiswa dan menawarkan latihan khusus untuk mengatasi kelemahan-kelemahan tersebut.

AI juga memungkinkan penilaian otomatis untuk tugas-tugas kompleks, seperti esai dan pertanyaan terbuka, yang biasanya membutuhkan banyak waktu dan tenaga dari pendidik. Teknologi Pemrosesan Bahasa Alami (*Natural Language Processing*) dapat menilai kualitas tulisan, koherensi, dan tata bahasa, sehingga memberikan umpan balik langsung kepada mahasiswa dan memungkinkan pengajar untuk fokus pada kegiatan pembelajaran

lainnya. Asesmen berbasis AI sangat menjanjikan dalam konteks pembelajaran jarak jauh, di mana umpan balik *real-time* dapat membantu menjembatani kesenjangan antara pengajar dan mahasiswa.

3.8 PENDIDIKAN BERBASIS KOMPETENSI

Pendidikan Berbasis Kompetensi (*Competency-Based Education* atau CBE) merupakan pergeseran paradigma dalam asesmen, yang berfokus pada penguasaan keterampilan dan pengetahuan tertentu oleh mahasiswa, bukan pada waktu yang dihabiskan di kelas. Model pendidikan tradisional menekankan pada "durasi kehadiran" (*seat-time*), di mana kemajuan mahasiswa didasarkan pada seberapa lama mereka mengikuti kursus, bukan pada pembelajaran atau penguasaan materi yang sesungguhnya. Sebaliknya, CBE memungkinkan mahasiswa untuk maju sesuai dengan kecepatan belajar mereka sendiri, serta memastikan bahwa setiap individu menunjukkan kompetensi sebelum beralih ke topik atau keterampilan berikutnya.

1) Mengalihkan Fokus dari Durasi Kehadiran ke Penguasaan Keterampilan

CBE menantang kerangka kerja asesmen tradisional dengan memprioritaskan penguasaan keterampilan di atas hafalan semata. Dalam sistem CBE, mahasiswa dinilai berdasarkan kemampuan mereka dalam menerapkan pengetahuan untuk memecahkan masalah di dunia nyata. Pendekatan ini selaras dengan Taksonomi Bloom, yang menekankan keterampilan berpikir tingkat tinggi seperti pembelajaran, sintesis, dan evaluasi. Dengan berfokus pada penguasaan kompetensi, CBE mempersiapkan mahasiswa untuk penerapan pengetahuan secara praktis serta membekali mereka dengan keterampilan yang relevan secara langsung bagi karier dan kehidupan mereka.

Sebagai contoh, alih-alih menilai mahasiswa matematika hanya melalui ujian, CBE mungkin mengharuskan mahasiswa untuk menunjukkan keterampilan matematika mereka dalam memecahkan masalah dunia nyata yang kompleks. Demikian pula, seorang mahasiswa bahasa dapat menunjukkan kemampuannya dengan menyusun laporan komprehensif atau berpartisipasi dalam debat terstruktur. Penilaian berbasis kinerja (PBA) ini menuntut Mahasiswa untuk menunjukkan pemahaman mereka melalui proyek dan tugas praktis, bukan melalui tes tradisional.

2) Contoh dari Sistem Sekolah dan Universitas yang Inovatif

CBE semakin mendapat perhatian dalam sistem pendidikan K–12 dan pendidikan tinggi di seluruh Amerika Serikat maupun secara global. Sebagai contoh, *Western Governors University* (WGU) beroperasi sepenuhnya menggunakan model berbasis kompetensi, yang memungkinkan mahasiswa untuk maju sesuai kecepatan masing-masing dan mencapai penguasaan dalam bidang-bidang yang ditargetkan. WGU menilai mahasiswa berdasarkan penyelesaian asesmen dan proyek yang mengukur kemampuan mereka dalam menerapkan pengetahuan pada skenario praktis, bukan berdasarkan persyaratan durasi waktu belajar di kelas (*seat-time*).

Dalam pendidikan K–12, Departemen Pendidikan New Hampshire telah menerapkan CBE di beberapa distrik sekolah. Kemajuan mahasiswa melalui jalur pembelajaran yang

dipersonalisasi disusun berdasarkan kompetensi, bukan berdasarkan tingkat kelas. Model ini memastikan bahwa mahasiswa menguasai keterampilan penting sebelum melanjutkan ke tahap berikutnya, sehingga menciptakan pengalaman pendidikan yang lebih relevan dan disesuaikan dengan kebutuhan individu.

Penilaian yang Holistik dan Responsif secara Budaya

Seiring dengan semakin beragamnya sistem pendidikan, muncul kebutuhan yang kian besar akan penilaian yang mengakui dan menghargai latar belakang budaya mahasiswa. Penilaian yang holistik dan responsif secara budaya memberikan pemahaman yang lebih komprehensif mengenai mahasiswa dengan mempertimbangkan konteks budaya tempat mereka hidup dan belajar. Model-model ini melampaui penilaian tradisional dengan menghargai data kualitatif seperti pengalaman pribadi dan interaksi social yang sering kali mengungkap wawasan tentang pembelajaran mahasiswa yang mungkin terlewatkan oleh penilaian kuantitatif.

1) Mengintegrasikan Latar Belakang Budaya Mahasiswa ke dalam Penilaian

Penilaian yang responsif secara budaya mengakui bahwa identitas budaya mahasiswa turut membentuk pengalaman belajar mereka. Pendidik dapat mengevaluasi pemahaman dan keterampilan mahasiswa secara lebih akurat dengan memasukkan konteks budaya ke dalam proses penilaian. Sebagai contoh, pendekatan responsif budaya dalam mata pelajaran sejarah dapat melibatkan penilaian terhadap kemampuan mahasiswa dalam menafsirkan peristiwa sejarah dari berbagai perspektif budaya, atau meminta mahasiswa untuk mengaitkan materi sejarah dengan latar belakang budaya mereka.

Pendekatan ini juga mencakup penilaian bahasa yang memandang keragaman linguistik mahasiswa multibahasa sebagai sebuah kekuatan, bukan hambatan. Alih-alih memberikan sanksi kepada Mahasiswa karena menggunakan ungkapan khas daerah (vernakular) atau melakukan alih kode (*code-switching*), penilaian responsif budaya mengevaluasi kemahiran berbahasa mahasiswa dalam berbagai konteks serta mengakui kekayaan keterampilan dwibahasa atau multibahasa yang mereka miliki.

2) Metode Kualitatif: Wawancara dan Pengembangan Teknologi Masa Depan Etnografis

Penilaian holistik sering kali mengandalkan metode kualitatif, seperti wawancara dan Pengembangan teknologi masa depan etnografis, untuk menangkap nuansa latar belakang budaya mahasiswa. Metode-metode ini memungkinkan pendidik memperoleh wawasan yang lebih mendalam mengenai proses berpikir, motivasi, dan nilai-nilai budaya mahasiswa. Sebagai contoh, wawancara dapat mengungkap bagaimana mahasiswa memaknai bidang materi tertentu, sehingga memberikan konteks yang berharga untuk memahami kinerja akademik mereka.

Pengembangan teknologi masa depan etnografis, yang melibatkan studi mendalam dan kontekstual mengenai perilaku serta interaksi mahasiswa, berperan penting dalam menilai keterampilan kolaboratif dan sosial. Pengembangan teknologi masa depan ini membantu pendidik memahami bagaimana mahasiswa berinteraksi dengan teman sebaya dari beragam latar belakang dan bagaimana mereka menerapkan pengetahuan budaya mereka dalam situasi sosial.

Pengukuran Keterampilan Sosial-Emosional dan Non-Kognitif

Di dunia yang kompleks saat ini, keberhasilan bergantung pada pengetahuan akademis serta keterampilan sosial-emosional dan non-kognitif. Kolaborasi, kecerdasan emosional, dan kreativitas semakin diakui sebagai hal yang penting bagi pertumbuhan pribadi dan profesional. Akibatnya, penilaian pendidikan terus berkembang dengan mencakup pengukuran keterampilan non-akademis ini, guna memastikan pendekatan yang lebih komprehensif terhadap perkembangan Mahasiswa.

1) Menilai Kolaborasi, Kecerdasan Emosional, dan Kreativitas

Penilaian keterampilan sosial-emosional melibatkan evaluasi terhadap kemampuan mahasiswa dalam memahami dan mengelola emosi, berempati terhadap orang lain, serta menjaga hubungan yang positif. Sebagai contoh, penilaian kecerdasan emosional dapat mengevaluasi seberapa baik mahasiswa mengenali emosi mereka, merespons stres, dan menunjukkan empati. Demikian pula, penilaian keterampilan kolaborasi mengamati kemampuan mahasiswa untuk bekerja secara efektif dalam kelompok, mengelola konflik, dan berkontribusi secara konstruktif terhadap tujuan bersama.

Kreativitas, yang merupakan keterampilan non-kognitif penting lainnya, sering kali dinilai melalui tugas-tugas yang menuntut pemikiran inovatif, pemecahan masalah, dan ekspresi artistik. Sebagai contoh, penilaian kreativitas mungkin meminta mahasiswa merancang prototipe produk atau mengembangkan solusi unik untuk masalah dunia nyata. Rubrik untuk tugas-tugas ini berfokus pada orisinalitas, fleksibilitas, dan kemampuan berpikir di luar kebiasaan (*out-of-the-box*).

3.9 INSTRUMEN PENILAIAN KETERAMPILAN NON-AKADEMIS

Berbagai kerangka kerja dan instrumen telah dikembangkan untuk menilai keterampilan non-kognitif secara efektif. *Collaborative for Academic, Social, and Emotional Learning* (CASEL) menyediakan kerangka kerja komprehensif untuk pembelajaran sosial-emosional (SEL), yang mengidentifikasi lima kompetensi inti: kesadaran diri, manajemen diri, kesadaran sosial, keterampilan membina hubungan, dan pengambilan keputusan yang bertanggung jawab. Sekolah dan pendidik dapat menggunakan kerangka kerja ini untuk merancang asesmen yang mengevaluasi perkembangan mahasiswa di setiap bidang.

Instrumen seperti *Devereux Student Strengths Assessment* (DESSA) menyediakan metode terstruktur untuk menilai kompetensi sosial-emosional mahasiswa jenjang K–12. DESSA mencakup skala penilaian oleh pengajar yang mengukur regulasi diri, ketangguhan, dan kerja sama, sehingga memungkinkan pendidik memantau perkembangan sosial-emosional mahasiswa dari waktu ke waktu. Demikian pula, *Big Five Inventory* mengukur karakteristik non-kognitif seperti keterbukaan, kesadaran (*conscientiousness*), dan keramahan (*agreeableness*), serta memberikan wawasan berharga mengenai kepribadian dan kecenderungan perilaku mahasiswa.

3.10 IMPLIKASI TERHADAP KEBIJAKAN DAN PRAKTIK PENDIDIKAN

Rekomendasi Kebijakan

Pergeseran menuju model penilaian inovatif dalam pendidikan memiliki implikasi signifikan terhadap kebijakan pendidikan dan sistem akuntabilitas. Seiring beralihnya penekanan dari tes standar tradisional ke penilaian yang lebih holistik, berbasis kompetensi, dan formatif, para pembuat kebijakan harus meninjau kembali struktur dan ukuran yang digunakan untuk mengevaluasi hasil pendidikan. Pergeseran ini menuntut kajian kritis terhadap kerangka kerja akuntabilitas yang ada guna memastikan bahwa kerangka kerja tersebut mendukung, dan bukan menghambat, penerapan praktik-praktik inovatif.

1) Meninjau Kembali Sistem Akuntabilitas dalam Konteks Penilaian Inovatif

Sistem akuntabilitas saat ini di banyak konteks pendidikan sangat bergantung pada tes standar, yang terutama mengukur pencapaian akademik dalam serangkaian keterampilan terbatas, seperti matematika dan membaca. Sistem-sistem ini sering kali mengabaikan tujuan pembelajaran yang lebih luas, seperti berpikir kritis, kolaborasi, dan kreativitas. Untuk mendorong pendekatan pendidikan yang lebih komprehensif, pembuat kebijakan sebaiknya mengembangkan sistem akuntabilitas yang mencakup berbagai ukuran pembelajaran mahasiswa, termasuk penilaian berbasis kinerja (PBA) dan metrik pembelajaran *sosial-emosional* (SEL).

Sebagai contoh, beberapa negara bagian di Amerika Serikat, seperti New Hampshire, telah mulai menguji coba model akuntabilitas inovatif melalui inisiatif seperti *Performance Assessment of Competency Education*. Sistem ini mencakup penilaian kinerja yang dikembangkan secara lokal di samping tes standar, sehingga memberikan gambaran yang lebih seimbang mengenai pembelajaran mahasiswa. Pembuat kebijakan perlu mempertimbangkan untuk memperluas inisiatif semacam itu serta menyediakan sumber daya untuk memperluas cakupan model yang berhasil ke berbagai distrik dan negara bagian.

Selain itu, sistem akuntabilitas harus menekankan pada perbaikan berkelanjutan alih-alih tindakan yang bersifat menghukum. Dengan memprioritaskan data penilaian formatif dan refleksi praktik di tingkat sekolah, sekolah dapat berfokus pada perbaikan yang bermakna daripada sekadar "mengajar demi tes" (*teaching to the test*). Hal ini memerlukan kebijakan yang memberikan insentif bagi sekolah untuk bereksperimen dan menerapkan metode penilaian inovatif, serta mekanisme untuk berbagi praktik terbaik di seluruh jaringan pendidikan.

2) Kebutuhan akan Dukungan Kebijakan untuk Inovasi Penilaian

Sekolah dan distrik memerlukan dukungan kebijakan yang kuat agar dapat menerapkan model penilaian inovatif secara sukses. Dukungan kebijakan tersebut mencakup pendanaan untuk penelitian dan pengembangan perangkat penilaian baru, hibah untuk program percontohan, serta investasi dalam infrastruktur teknologi yang diperlukan untuk penilaian digital dan adaptif. Para pembuat kebijakan juga perlu mempertimbangkan untuk menjalin kemitraan dengan perusahaan teknologi pendidikan dan lembaga penelitian guna mengembangkan asesmen yang valid dan reliabel.

Kerangka kerja legislatif harus cukup fleksibel untuk mengakomodasi bentuk-bentuk asesmen baru sembari tetap memastikan bahwa pembelajaran mahasiswa menjadi prioritas utama. Sebagai contoh, kebijakan dapat memberikan pengecualian dari persyaratan tes standar tertentu bagi sekolah yang menerapkan sistem asesmen inovatif dan menunjukkan keberhasilan melalui metode alternatif. Selain itu, diperlukan pedoman yang jelas serta dukungan untuk menjamin kesetaraan dalam asesmen, sehingga seluruh mahasiswa apa pun latar belakang mereka dapat memperoleh akses terhadap evaluasi yang adil dan responsif secara budaya.

Pengembangan Profesional bagi Pendidik

Keberhasilan penerapan model asesmen inovatif sangat bergantung pada keterampilan dan pengetahuan para pendidik. Pengajar harus dibekali kemampuan tidak hanya untuk melaksanakan asesmen baru ini, tetapi juga untuk menafsirkan data yang dihasilkan serta memanfaatkannya sebagai dasar dalam merancang pembelajaran. Hal ini menuntut perubahan mendasar dalam pelatihan pengajar dan pengembangan profesional.

1) Mempersiapkan Pengajar untuk Menerapkan dan Menafsirkan Bentuk-Bentuk Penilaian Baru

Para pendidik memerlukan program pengembangan profesional komprehensif yang berfokus pada teori maupun praktik penilaian inovatif. Sebagai contoh, sesi pelatihan sebaiknya mencakup prinsip-prinsip penilaian formatif dan penilaian berbasis kinerja, perancangan rubrik yang efektif, serta strategi untuk memberikan umpan balik yang konstruktif kepada mahasiswa. Pengajar juga perlu belajar menggunakan data penilaian untuk mengambil keputusan pembelajaran yang tepat, melakukan diferensiasi pengalaman belajar, dan mendukung perkembangan mahasiswa.

Pengembangan profesional harus bersifat berkelanjutan dan terintegrasi dalam praktik keseharian pengajar. Hal ini dapat dicapai melalui komunitas belajar profesional (*Professional Learning Communities* atau PLC), tempat para pengajar berkolaborasi, berbagi pengalaman, dan saling belajar satu sama lain. PLC menyediakan lingkungan yang mendukung bagi pendidik untuk bereksperimen dengan metode penilaian baru, merefleksikan efektivitasnya, dan melakukan penyesuaian yang diperlukan. Selain itu, sekolah sebaiknya menyediakan program pembinaan dan pendampingan yang memasangkan pengajar dengan pakar penilaian yang dapat memberikan bimbingan serta dukungan yang disesuaikan dengan kebutuhan.

2) Kebutuhan Pelatihan untuk Pergeseran Paradigma

Beralih ke penilaian inovatif menuntut adanya pergeseran paradigma mengenai cara pengajar memandang pembelajaran mahasiswa dan peran mereka sebagai fasilitator pembelajaran tersebut. Penilaian tradisional sering kali memperkuat model pembelajaran yang berpusat pada pengajar, sedangkan penilaian formatif dan penilaian berbasis kinerja mendorong pendekatan yang berpusat pada mahasiswa dengan menekankan pembelajaran aktif dan penilaian diri. Untuk menghadapi perubahan ini, pengajar perlu dilatih agar mampu membangun budaya kelas yang menghargai inkuiri, kolaborasi, dan refleksi.

Program pendidikan pengajar juga harus menjawab tantangan emosional dan praktis yang muncul akibat perubahan praktik penilaian. Sebagai contoh, pengajar mungkin merasa

kewalahan dengan kompleksitas penilaian berbasis kinerja (PBA) atau membutuhkan penjelasan mengenai cara mengevaluasi hasil kerja mahasiswa secara adil. Pelatihan harus membekali pengajar dengan perangkat praktis, contoh-contoh, dan kesempatan untuk berlatih secara langsung guna membangun kepercayaan diri serta kompetensi mereka.

Lebih lanjut, pengembangan profesional harus menekankan kompetensi budaya dan kemampuan menciptakan penilaian inklusif yang mencerminkan keberagaman latar belakang mahasiswa. Pengajar perlu belajar mengenali dan mengatasi bias budaya dalam rancangan penilaian serta memastikan bahwa evaluasi yang mereka lakukan bersifat adil dan menghargai pengalaman unik setiap mahasiswa. Dengan berinvestasi dalam program pelatihan yang komprehensif dan tanggap budaya, sistem pendidikan dapat lebih mendukung pengajar dan meningkatkan hasil belajar mahasiswa.

BAB 4
TRANSPARANSI EVALUASI MENULIS BERBASIS AI

Algoritma telah lama mampu memberikan skor pada esai dengan meniru cara manusia melakukannya, yakni dengan mengikuti rubrik penilaian. Melalui algoritma semacam itu, skor otomatis dapat segera memberi tahu mahasiswa di tingkat mana tulisan mereka berada berdasarkan rubrik tersebut. Namun, hal ini saja tidak membantu mahasiswa memahami mengapa mereka berada di tingkat tersebut atau apa yang perlu mereka lakukan untuk naik ke tingkat berikutnya. Panduan yang dapat ditindaklanjuti seperti itu sering kali penting agar umpan balik menjadi efektif.

Meskipun umpan balik terperinci semacam itu sangat penting untuk mendukung perkembangan mahasiswa sebagai penulis, pembuatannya juga memakan banyak waktu. Pengajar sering kali memiliki keterbatasan waktu untuk melakukan sesi diskusi empat mata dengan mahasiswa atau memberikan umpan balik terperinci pada esai dan setiap versi tulisan mahasiswa. Salah satu solusi yang dapat mendukung pendidik maupun mahasiswa dalam proses pemberian umpan balik ini adalah penggunaan alat evaluasi menulis otomatis (*Automated Writing Evaluation* atau AWE).

Kemajuan terkini dalam kecerdasan buatan (AI) memungkinkan alat-alat ini melampaui penilaian berbasis rubrik untuk memberikan umpan balik yang lebih bernuansa mengenai kualitas dan karakteristik tulisan mahasiswa. Seiring berkembangnya kemampuan AI ini, pembuatan alat AWE serta penerapannya oleh mahasiswa dan pengajar di ruang kelas kemungkinan akan semakin cepat. Menulis merupakan aktivitas yang memiliki banyak aspek, dan umpan balik AI mungkin memprioritaskan satu aspek penulisan dibandingkan aspek lainnya. Sebagai contoh, umpan balik mungkin berfokus pada tata bahasa dan konvensi penulisan, atau pada aspek yang lebih substantif seperti koherensi, penalaran, atau struktur organisasi. Terlepas dari fokus spesifik alat tersebut, semua solusi AWE pada dasarnya bertujuan untuk mendukung perkembangan mahasiswa sebagai penulis.

Namun, terdapat risiko nyata bahwa alat-alat ini justru memberikan dampak sebaliknya: yakni secara tidak sengaja menghambat perkembangan kemampuan menulis jika tidak dirancang atau digunakan dengan tepat. Bukti awal dari penggunaan alat AI secara lebih luas menunjukkan perlunya kehati-hatian. Sebuah pemantauan sistematis baru-baru ini menemukan bahwa ketergantungan berlebihan mahasiswa pada sistem AI dapat mengikis keterampilan yang menjadi fondasi komunikasi tertulis yang kuat seperti berpikir kritis, pengambilan keputusan, dan penalaran analitis. Kita juga mengetahui bahwa motivasi merupakan elemen kunci dalam pembelajaran, dan kurangnya keterlibatan manusia dalam proses pemberian umpan balik dapat mengurangi tingkat keterlibatan pengguna terhadap perangkat tersebut. Terkait hal ini, perangkat AWE secara tidak sengaja dapat mengurangi interaksi penting antara mahasiswa dan pengajar, atau antar-Mahasiswa. Kita harus menyadari risiko-risiko ini dan memperhitungkan kemungkinannya saat merancang, menerapkan, dan mengevaluasi perangkat AWE.

Tujuan mengevaluasi alat AWE (*Automated Writing Evaluation*) bukan sekadar untuk memastikan alat tersebut tidak menimbulkan dampak negatif, melainkan untuk menentukan apakah alat itu benar-benar mendukung perkembangan mahasiswa sebagai penulis. Hal ini menuntut pemeriksaan cermat terhadap bukti yang berkaitan dengan desain, pengembangan, penerapan, dan validasi alat-alat tersebut. Pertanyaan kuncinya adalah: elemen apa saja dari proses pengembangan dan evaluasi yang harus ada agar alat AWE dapat menjalani pemantauan eksternal yang ketat?

Untuk menjawab pertanyaan ini, ada baiknya kita membayangkan lima aspek yang mencirikan alat bantu menulis yang sudah matang sepenuhnya:

1. tujuan yang didefinisikan dengan jelas beserta skenario penggunaannya;
2. bukti yang mendukung penggunaan tersebut;
3. transparansi mengenai bagaimana model AI dilatih dan beroperasi di dalam alat tersebut;
4. pelaporan kinerja model yang transparan berdasarkan kumpulan data (*dataset*) yang terdefinisi dengan baik; serta
5. mekanisme perlindungan yang dirancang secara eksplisit untuk menjaga privasi dan keselamatan mahasiswa.

Landasan dari kelima aspek ini adalah konsep tentang menulis itu sendiri yaitu bagaimana mahasiswa berkembang sebagai penulis seiring waktu, dan apa yang mereka butuhkan untuk meningkatkan keterampilan menulis mereka.

Mungkin tidak realistis untuk mengharapkan alat AWE saat ini mencapai tingkat kematangan penuh, mengingat masih barunya teknologi dan produk penyedia umpan balik berbasis AI. Selain itu, tingkat kematangan alat-alat tersebut kemungkinan akan bervariasi di kelima aspek tadi, bergantung pada prioritas pengembang dan masalah spesifik yang ingin diatasi oleh alat tersebut. Meskipun demikian, terlepas dari tingkat kematangannya, alat-alat yang ada saat ini tetap dapat memberikan dukungan yang berarti bagi mahasiswa, khususnya di sekolah-sekolah yang minim atau bahkan tidak memberikan umpan balik sama sekali.

Di tengah upaya komunitas pendidikan untuk mengintegrasikan AI ke dalam ruang kelas secara bertanggung jawab, transparansi mengenai tingkat kematangan alat AWE menjadi pedoman penting. Hal ini memastikan bahwa alat bantu menulis berbasis AI dievaluasi tidak hanya berdasarkan klaim yang disampaikan, tetapi juga berdasarkan bukti yang mendasari klaim tersebut. Transparansi semacam itu dapat membantu instansi pemerintah di tingkat lokal, negara bagian, maupun federal dalam meninjau dan menyetujui alat bantu pengajaran; hal ini juga dapat mendukung pengajar dalam mengambil keputusan di kelas mengenai alat mana yang akan digunakan dan bagaimana cara menggunakannya.

Untuk mendukung hal tersebut, kami mengusulkan sebuah kerangka kerja guna mengevaluasi tingkat kematangan alat AWE berdasarkan kelima dimensi ini. Meskipun tingkat kematangan saja tidak menjamin kualitas, hal tersebut merupakan indikator kuat akan kualitas alat tersebut. Perangkat yang sudah matang yang telah melalui berbagai siklus pengembangan, iterasi dokumentasi, dan beragam studi efikasi dapat ditelaah dan dimintai pertanggungjawaban dengan cara yang tidak dapat dilakukan terhadap perangkat tahap awal.

Sama halnya dengan obat dalam uji klinis Fase I yang tidak serta-merta dianggap "buruk" melainkan hanya terlalu baru untuk diketahui efektivitas atau keamanannya, perangkat AWE pada tahap awal pun belum dapat dinilai secara utuh. Demikian pula, obat Fase III mungkin saja tetap gagal, namun ia menyediakan bukti yang mendukung evaluasi lebih lanjut. Dengan cara yang sama, tingkat kematangan perangkat AWE mencerminkan kadar transparansi dan bukti yang tersedia untuk ditelaah. Dalam konteks pendidikan, matriks kematangan dapat memanfaatkan tingkatan bukti, seperti yang dijabarkan dalam ESSA.

Oleh karena itu, kami memperkenalkan "matriks kematangan" yang menguraikan kriteria pada berbagai tahap pengembangan. Menurut pandangan kami, matriks kematangan menawarkan cara terstruktur untuk menyelaraskan ekspektasi terkait kualitas dan efikasi; hal ini mendorong pengembang dan peneliti untuk mengungkapkan lebih banyak informasi seiring berkembangnya perangkat, serta memungkinkan pendidik dan pembuat kebijakan untuk mengambil keputusan berbasis bukti mengenai perangkat mana yang akan digunakan untuk tujuan apa dan bagi siapa atau oleh siapa.

Matriks kematangan menelaah hal ini secara langsung dengan mengkaji klaim teori tindakan (*theory-of-action*) serta bukti yang mendukung klaim tersebut, di samping pemeriksaan terhadap teknologi AI, validasi, serta pertimbangan mengenai keselamatan, keamanan, dan privasi mahasiswa. Kami mengawalinya dengan menyajikan perangkat AWE utama dan menempatkannya dalam konteks pembahasan ini. Kemudian, kami memperkenalkan matriks kematangan, membahas setiap area fokus, dan menempatkan setiap perangkat pada tingkat kematangan yang diuraikan dalam rubrik untuk mengilustrasikan serta memperjelas kriteria tersebut.

Visi kami mengenai transparansi sangat selaras dengan *AI Risk Management Framework* dari NIST, yang menguraikan elemen-elemen seperti konteks aplikasi, data dan input, model dan validasi AI, serta pertimbangan bagi individu, masyarakat, dan planet. Meskipun kerangka kerja NIST bersifat umum untuk semua aplikasi AI, fokus kami adalah pada desain dan penggunaan perangkat AWE di dalam kelas. Visi kami juga sejalan dengan dokumen *Designing for Education with Artificial Intelligence* dari OET (Kantor Teknologi Pendidikan) Departemen Pendidikan AS, yang menyatakan bahwa perancangan untuk pendidikan memerlukan penyediaan bukti mengenai dasar pemikiran dan dampak perangkat, sembari tetap menjamin keselamatan dan keamanan. Kedua dokumen federal tersebut menguraikan perlunya transparansi sebagai cara utama untuk membangun kepercayaan saat menggunakan perangkat yang diperkuat atau didukung oleh AI dalam pendidikan.

4.1 PERANGKAT EVALUASI PENULISAN OTOMATIS (AWE)

Sebagaimana disebutkan di atas, perangkat AWE dirancang dengan pendekatan yang berbeda-beda untuk mendukung aspek penulisan tertentu. Empat contoh yang saat ini digunakan di ruang kelas menggambarkan beragam tujuan dan teknologi AI yang digunakan. *Khanmigo Writing Coach* tutor AI dari Khan Academy yang melibatkan mahasiswa dalam dialog mengenai draf tulisan mereka menggunakan AI generatif diluncurkan pada tahun 2024. MI Write, sebuah perangkat AWE yang telah digunakan secara luas selama lebih dari sepuluh

tahun, menghasilkan skor penulisan keseluruhan melalui penilaian otomatis menggunakan *Project Essay Grade* dan *Six Trait Rubric*. eRevise+RF menyediakan umpan balik yang berorientasi pada revisi dengan fokus pada penggunaan bukti dan penalaran oleh mahasiswa dalam tulisan argumentatif serta telah diteliti di berbagai ruang kelas di dua negara bagian.

Write-On with Cambi! (selanjutnya disebut Cambi) yang dirilis pada tahun 2024 dan telah digunakan terutama di satu negara bagian oleh lebih dari 2.000 pengajar serta 40.000 mahasiswa memandu mahasiswa dalam tahap penyusunan draf dan revisi dengan menekankan elemen-elemen kunci komposisi. Perangkat-perangkat ini dipilih karena telah lama digunakan di berbagai negara bagian (misalnya, MI Write), karena penggunaan AI yang inovatif (eRevise+RF, Cambi), dan/atau karena tingginya volume penggunaan dalam satu bidang umpan balik (matematika) yang memanfaatkan AI generatif dengan adaptasi baru untuk ranah penulisan (*Writing Coach*).

Fokus kami adalah menjabarkan matriks kematangan menggunakan perangkat-perangkat ini, karena setiap perangkat menawarkan elemen yang mencerminkan area fokus utama dan tingkatan dalam matriks tersebut. Kami bermaksud mengilustrasikan bagaimana matriks dapat ditafsirkan dengan menggunakan perangkat-perangkat ini sebagai contoh, alih-alih memberikan penilaian mutlak terhadap perangkat itu sendiri. Pilihan ini didasari oleh keinginan untuk meminimalkan konflik kepentingan: penulis bab ini turut berkontribusi dalam pengembangan Cambi. Informasi mengenai setiap perangkat AWE saat matriks kematangan diterapkan diperoleh melalui penelitian yang dipublikasikan dalam jurnal dan konferensi, serta tulisan blog.

4.2 MATRIKS KEMATANGAN AWE

Dalam kerangka kerja ini, kami menguraikan apa yang secara wajar dapat diharapkan oleh para evaluator baik itu pendidik, pembuat kebijakan, maupun peneliti pada berbagai tingkat kematangan pengembangan perangkat. Sebagaimana disajikan dalam Tabel 4.1, matriks kematangan memiliki lima area fokus: Teori Tindakan (*Theory of Action*), Evaluasi Hasil (*Outcome Evaluation*), Teknologi AI, Validasi Model, serta Keselamatan, Keamanan, & Privasi. Setiap area memiliki empat tingkat kematangan: Awal (*Early*), Berkembang (*Developing*), Mapan (*Established*), dan Matang (*Mature*). Kami memperkirakan bahwa tingkat kematangan dapat diterapkan secara terpisah pada setiap area; misalnya, sebuah perangkat mungkin memiliki teori tindakan yang kuat namun minim bukti untuk mendukung klaimnya. Sebaliknya, sebuah perangkat mungkin memiliki deskripsi yang sangat rinci mengenai teknologi AI dan proses validasi model, namun tidak memiliki teori tindakan. Hal ini sebagian disebabkan oleh fakta bahwa pengembangan perangkat AWE dapat didorong dan dirancang oleh kelompok yang berbeda, seperti ilmuwan pembelajaran, insinyur AI, atau ahli penilaian.

Dengan demikian, kami memperkirakan perangkat akan menunjukkan pola tingkat kematangan tertentu di kelima area tersebut; pola-pola ini mengindikasikan kekuatan dan kelemahan bukti yang mendukungnya. Kami memperkirakan bahwa perangkat dengan bukti pendukung paling kuat akan mencapai tingkat Mapan (*Established*) atau Matang (*Mature*) di kelima area tersebut. Selanjutnya, kami menjelaskan masing-masing dari lima area fokus

matriks kematangan dan menguraikan secara singkat contoh tingkat kematangan untuk setiap area. Kami juga menyajikan contoh dari keempat perangkat AWE untuk mengilustrasikan tingkat-tingkat tersebut.

Tabel 4.1 Matriks Kematangan AWE

Tingkat Kematangan	**Teori Aksi**	**Evaluasi Hasil**	**Teknologi AI**	**Validasi Model**	**Keselamatan, Keamanan, dan Privasi**
Awal	Tidak ada teori aksi yang dijelaskan.	Tidak ada bukti pendukung yang disertakan.	Tidak ada deskripsi mengenai teknologi AI.	Tidak ada hasil validasi model.	Tidak ada bukti adanya langkah-langkah keselamatan, keamanan, atau privasi.
Berkembang	Tujuan alat disebutkan, namun tidak disertai penjelasan yang jelas mengenai bagaimana tujuan tersebut akan dicapai. Memperkenalkan konsep utama alat tersebut.	Terdapat studi korelasional atau penelitian kualitatif berskala kecil yang berfokus pada tujuan jangka pendek, kegunaan, atau persepsi.	Deskripsi minimal mengenai teknologi AI, pendekatan pemodelan, data pelatihan, dan umpan balik.	Bukti terbatas mengenai kinerja model dan deskripsi data untuk mendukung evaluasi tersebut.	Penyebutan dasar mengenai keamanan dan privasi data Mahasiswa. Deskripsi minimal tentang bagaimana potensi dampak berbahaya dikelola.
Mapan	Tujuan disebutkan dengan penjelasan yang terbatas mengenai cara alat tersebut mencapai tujuannya. Hal ini dapat mencakup model logika yang memuat input, aktivitas, output, dan	Sejumlah bukti berbasis metode campuran (misalnya, kuasi-eksperimental, kualitatif, data proses) dengan ukuran sampel yang memadai di berbagai sekolah atau distrik, yang berfokus pada	Pendekatan pemodelan dijelaskan namun kurang memuat detail penting, seperti alasan pemilihan model, data pelatihan, atau penjelasan	Bukti mengenai kemampuan model dalam memprediksi skor atau umpan balik secara akurat, dengan keselarasan yang terbatas antara prediksi dan definisi konstruk yang	Terdapat informasi mengenai perlindungan privasi, penyimpanan model, dan perjanjian penggunaan data. Terdapat deskripsi mengenai bagaimana potensi dampak berbahaya dikelola.

	hasil (*outcome*).	evaluasi tujuan jangka pendek.	mengenai mekanisme umpan balik.	diuraikan dalam teori tindakan.	
Dewasa	Penjelasan yang jelas mengenai tujuan perangkat termasuk sasaran jangka pendek dan jangka Panjang yang didukung oleh model logika. Menguraikan bagaimana berbagai komponen penyusun teori perubahan saling terhubung.	Bukti yang diperoleh melalui berbagai metode termasuk penelitian kualitatif, data proses, serta desain eksperimental atau kuasi-eksperimental—digunakan untuk mengevaluasi hasil jangka pendek maupun jangka panjang. Ukuran dan keterwakilan sampel mendukung interpretasi hasil.	Penjelasan rasional dan deskripsi yang transparan mengenai pendekatan pemodelan, data, dan fitur desain yang mendukung umpan balik.	Bukti mengenai kemampuan model untuk memprediksi skor atau umpan balik secara akurat, yang selaras dengan definisi konstruk dalam teori tindakan. Bukti akurasi yang mencakup beragam Mahasiswa, kelas, dan wilayah.	Protokol privasi dan keamanan data yang komprehensif: model yang disimpan di lingkungan aman, data terenkripsi, perjanjian kontraktual mengenai penggunaan data, serta mitigasi risiko serangan (adversarial).

Teori Tindakan (*Theory of Action*)

Elemen pertama dalam kerangka kerja ini adalah teori tindakan, sebuah komponen krusial dalam menentukan tujuan dan desain sistem AWE. Teori tindakan adalah "teori eksplisit mengenai bagaimana suatu intervensi seperti proyek, program, strategi, inisiatif, atau kebijakan berkontribusi pada rangkaian hasil antara dan, pada akhirnya, terhadap hasil akhir yang ditargetkan atau teramati".

Pernyataan sikap *National Council on Measurement in Education* (NCME) tahun 2018 mensyaratkan penjabaran eksplisit mengenai "komponen-komponen penyusun program; hasil akhir yang ditargetkan (baik jangka pendek maupun jangka panjang); mekanisme kausal yang diyakini mendasari perubahan yang diharapkan; serta pelaksanaan ideal program yang dipercaya paling berpeluang menghasilkan capaian-capaian tersebut." Model logika memperjelas teori tindakan tersebut dengan menguraikan input, aktivitas, dan output yang terkait dengan hasil akhir yang diinginkan. Teori tindakan merinci bagaimana aktivitas dan keluaran (*output*) tertentu menghasilkan dampak (*outcome*) tertentu, serta bagaimana faktor-

faktor kontekstual memengaruhinya. Dalam konteks alat AWE (*Automated Writing Evaluation*), faktor kontekstual dapat mencakup dukungan dan keterlibatan pengajar saat Mahasiswa menggunakan alat tersebut di dalam kelas.

Pada dasarnya, teori tindakan merupakan pernyataan mengenai tujuan yang ingin dicapai oleh alat tersebut dan bagaimana alat itu mencapainya. Perlu dicatat bahwa teori tindakan ini sejalan dengan panduan OET mengenai perancangan alat pendidikan serta penyediaan bukti dan dasar pemikiran terkait dampaknya; panduan OET itu sendiri merekomendasikan penggunaan teori tindakan dalam perancangan, pengembangan, dan validasi alat pendidikan berbasis AI.

Elemen-elemen kunci dari teori tindakan mencakup definisi konstruk serta umpan balik penilaian yang diberikan kepada mahasiswa dan/atau pengajar. Definisi konstruk mengacu pada elemen-elemen penulisan yang dievaluasi oleh alat AWE. Definisi ini menempatkan penggunaan elemen-elemen tersebut dalam konteks riset penulisan atau penilaian penulisan (misalnya, standar, rubrik, keputusan ahli mengenai rentang penilaian, dan ilmu pembelajaran). Definisi konstruk untuk suatu alat bisa bersifat luas (misalnya, mencakup seluruh aspek penulisan yang ditetapkan dalam standar atau rubrik) ataupun sempit (misalnya, hanya berfokus pada elemen konvensi atau elemen argumentasi tertentu). Apa pun pilihannya, konstruk yang dinilai oleh alat tersebut harus didefinisikan dengan jelas dan didasarkan pada penelitian. Umpan balik penilaian mengacu pada mekanisme yang menerjemahkan elemen-elemen yang dievaluasi ke dalam format yang dapat diakses dan digunakan oleh mahasiswa maupun pengajar. Bagi mahasiswa, hal ini dapat tercermin dari cara mereka berinteraksi dengan umpan balik AWE melalui antarmuka pengguna.

Bagi pengajar, hal ini dapat tercermin dari cara mereka memanfaatkan umpan balik tersebut dalam proses pengajaran dan evaluasi terhadap tulisan mahasiswa. Bagi kedua pihak, umpan balik penilaian juga mencakup informasi tambahan yang membantu pengguna memahami alat tersebut (misalnya, panduan pengguna). Panduan ini memberi tahu pengguna bahwa mereka sedang berinteraksi dengan "model AI" dan berfokus menjelaskan tujuan penggunaan sistem AWE, cara menggunakan fitur-fitur sistem AWE, serta khususnya dalam konteks penilaian K-12 bagaimana alat tersebut selaras dengan standar negara bagian dan rubrik penilaian. Pada intinya, umpan balik penilaian merupakan mekanisme kunci yang mewujudkan teori tindakan tersebut; Hal ini merupakan implementasi terprogram dari asumsi-asumsi yang mendasari bagaimana informasi AWE dapat ditafsirkan dan digunakan oleh pengajar serta mahasiswa.

Untuk mengeksplorasi seperti apa teori tindakan bagi perangkat AWE dalam praktiknya, kami telah menyusun satu teori untuk perangkat penulisan fiktif. Tujuan perangkat fiktif ini adalah membantu mahasiswa meningkatkan strategi transisi mereka. Tabel 4.2 menyajikan model logika yang mencakup input, aktivitas, *output*, serta hasil jangka pendek dan jangka panjang yang diharapkan dari penggunaan perangkat ini. Perlu dicatat bahwa konstruk "transisi" merupakan elemen dasar dalam model logika tersebut; artinya, perangkat harus mampu mengidentifikasi transisi dalam esai mahasiswa secara akurat dan memberikan

evaluasi mengenai kualitas transisi (misalnya, lemah/kuat). Dalam konteks ini, model logika tersebut mencakup dua tingkat klaim:

1) konstruk tersebut dapat dinilai secara akurat.
2) penggunaan perangkat tersebut menghasilkan luaran jangka pendek dan jangka panjang yang diharapkan.

Tabel 4.2 Model Logika Alat AWE Fiktif untuk Mengilustrasikan Teori Tindakan yang Matang

Input	Aktivitas	Hasil	Hasil Jangka Pendek	Hasil Jangka Panjang
- Mahasiswa menyusun draf esai dan menekan tombol "dapatkan umpan balik" di dalam perangkat lunak AWE.	- Alat ini menyoroti transisi yang lemah atau tidak ada serta meminta Mahasiswa untuk melakukan revisi - Mahasiswa melakukan perubahan pada teks transisi yang sudah ada - Mahasiswa menambahkan teks transisi baru	- Esai akhir diserahkan ke sistem	- Peningkatan penggunaan beragam kata dan frasa transisi - Keterhubungan antarparagraf yang lebih kohesif - Skor rubrik yang lebih tinggi pada dimensi organisasi - Peningkatan kepercayaan diri dalam menghubungkan gagasan	- Mahasiswa menerapkan keterampilan penggunaan kata penghubung pada tugas menulis yang tidak melibatkan alat bantu - Esai Mahasiswa menunjukkan alur logika yang lebih kuat - Meningkatkan nilai keseluruhan esai, melampaui aspek koherensi

Teori tindakan kemudian menghubungkan berbagai masukan, aktivitas, dan keluaran ini untuk merumuskan teori mengenai mekanisme kausal yang menghasilkan capaian akhir. Dalam contoh ini, teori tindakan dapat dikembangkan berdasarkan model logika dengan penjelasan sebagai berikut:

Pernyataan dan teks transisi merupakan aspek krusial dalam penulisan yang berkualitas tinggi. Unsur-unsur ini menandakan hubungan antar-gagasan, serta memandu pembaca beralih dari satu kalimat atau paragraf ke kalimat atau paragraf berikutnya. Tanpa unsur tersebut, poin-poin yang sebenarnya kuat sekalipun bisa terasa terputus-putus atau mendadak. Transisi yang efektif tidak sekadar menciptakan alur yang lancar; transisi memperjelas hubungan logis (sebab-akibat, pertentangan, urutan, penekanan), membantu pembaca mengantisipasi apa yang akan disampaikan selanjutnya, serta memperkuat koherensi keseluruhan argumen atau narasi. Saat mahasiswa mengumpulkan draf esai, teknologi AWE mendeteksi keberadaan pernyataan transisi.

Hal ini memungkinkan sistem untuk menyoroti bagian tertentu dan mendorong Mahasiswa agar lebih memperhatikan area yang memiliki transisi lemah atau tidak ada sama sekali. Dorongan ini memicu mahasiswa untuk melakukan metakognisi guna meninjau, mengevaluasi, dan menyunting teks mereka. Dalam konteks penggunaan AWE ini, pengajar

memberikan dukungan dengan menjawab pertanyaan dan memperjelas konsep, yang pada gilirannya membantu mahasiswa memperdalam penerapan strategi transisi. Siklus revisi ini menghasilkan esai akhir yang berujung pada

- Peningkatan penggunaan beragam jenis transisi,
- Hubungan yang lebih kuat antarparagraf,
- Skor rubrik yang lebih tinggi pada aspek-aspek seperti organisasi tulisan, dan
- Meningkatnya kepercayaan diri mahasiswa dalam menghubungkan gagasan.

Seiring berjalannya waktu, manfaat jangka pendek ini akan terbawa ke situasi lain melalui proses generalisasi dan regulasi diri: Mahasiswa mulai mampu mendeteksi sendiri transisi yang lemah atau tidak ada dalam berbagai tugas menulis yang berbeda.

Sebagaimana disebutkan sebelumnya, teori tindakan ini diawali dengan mengidentifikasi dan, sampai batas tertentu, mendefinisikan konsep pernyataan transisi. Sangatlah penting untuk memperjelas aspek penulisan apa yang menjadi fokus alat tersebut; hal ini krusial untuk merinci capaian spesifik, baik dalam jangka pendek maupun jangka panjang. Teori tindakan untuk alat yang belum terlalu matang mungkin tidak mendefinisikan konsep tersebut secara eksplisit, tidak menyertakan model logika sama sekali, atau jika model logika tersedia gagal menghubungkan masukan, aktivitas, dan keluaran dengan mekanisme yang menghasilkan capaian akhir.

Tabel 4.3 menyajikan teori tindakan untuk keempat alat AWE tersebut, yang menggambarkan beragam tingkat kematangan sebagaimana diuraikan dalam matriks. Perangkat dengan tingkat spesifisitas rendah dalam menjabarkan input, aktivitas, dan output dikategorikan pada tingkat *Developing* (Berkembang). Perangkat yang menawarkan spesifisitas lebih tinggi terkait elemen-elemen tersebut namun tanpa detail seperti pada contoh fiktif dikategorikan pada tingkat *Established* (Mapan); sedangkan perangkat dengan tingkat spesifisitas tinggi dalam hal detail umpan balik, serta sistem yang berkontribusi terhadap umpan balik dan hasil tersebut, dikategorikan pada tingkat *Mature* (Matang).

Tabel 4.3 Tingkat Kematangan Perangkat AWE untuk Fokus *Theory-of-Action* (Teori Aksi)

Alat AWE	Keterangan	Tingkat Kematangan
Pelatih Penulisan	Dengan memberikan akses ke Writing Coach, mahasiswa terlibat dalam proses penulisan, yang menghasilkan peningkatan kualitas esai baik pada tugas saat ini maupun pada tugas-tugas mendatang. Mahasiswa menjalani keseluruhan proses pengembangan esai melalui empat tahapan: (1) memahami tugas; (2) menyusun kerangka; (3) membuat draf; serta (4) menerima umpan balik dan melakukan revisi. Mahasiswa mendapatkan umpan balik yang dihasilkan oleh AI terkait lima elemen penulisan:	Mengembangkan

	pendahuluan, bukti, struktur, gaya dan nada, serta kesimpulan.	
MI Write	Dengan meringankan beban dalam memberikan umpan balik, MI Write memungkinkan pengajar untuk memberikan lebih banyak tugas menulis dan memfokuskan upaya pemberian umpan balik mereka. Pada gilirannya, MI Write memberikan kesempatan berlatih menulis yang lebih banyak bagi Mahasiswa sesuatu yang mereka butuhkan untuk meningkatkan kualitas tulisan. Selain itu, skor kualitas tulisan otomatis dari MI Write menyediakan data penilaian yang tepat waktu dan andal yang dapat digunakan untuk memantau perubahan kinerja dari waktu ke waktu serta umpan balik otomatis yang membantu mahasiswa meningkatkan pemahaman mereka mengenai kriteria kualitas tulisan. MI Write menawarkan sejumlah fitur yang secara teoretis dapat meningkatkan keterampilan menulis, termasuk beragam kesempatan untuk berlatih menulis, penggunaan *graphic organizer* untuk mendorong perencanaan yang lebih baik, pelaporan skor berkala guna mendukung pemantauan kemajuan secara mandiri oleh mahasiswa, pemantauan sejawat (*peer review*) yang dapat memicu revisi lebih lanjut, serta pelajaran interaktif untuk pengembangan keterampilan.	Didirikan
eRevisi + RF	Teori tindakan eRevise+RF berlandaskan pada teori sosiokultural mengenai penilaian formatif, yang menekankan pada "sifat pembelajaran yang kontekstual dan sosial." Para penulis berpendapat bahwa perangkat AWE harus mempertimbangkan norma-norma disiplin ilmu terkait argumentasi berbasis teks, nilai serta tujuan pengajar dan mahasiswa, peran pengajar dalam memediasi penggunaan perangkat AWE, dan hasil penggunaan oleh mahasiswa termasuk revisi yang dilakukan mahasiswa untuk	Dewasa

	menyempurnakan esai mereka. Oleh karena itu, perangkat AWE harus mampu menangkap secara akurat aspek-aspek penting dari penggunaan bukti. Pengajar harus memandang perangkat tersebut selaras dengan tujuan pedagogis mereka, sementara mahasiswa harus menganggapnya mudah dipahami dan bermanfaat. Dukungan pengajar terhadap mahasiswa dalam menafsirkan umpan balik eRevise+RF kemungkinan bervariasi, meskipun dialog substantif antara pengajar dan mahasiswa mendukung efektivitas revisi esai serta pengembangan keterampilan. Secara keseluruhan, hal ini menghasilkan esai yang lebih baik dengan penggunaan bukti yang selaras dengan umpan balik yang diberikan serta memungkinkan mahasiswa menjelaskan bagaimana proses pembelajaran mereka sejalan dengan umpan balik tersebut.	
Cambi	Mahasiswa yang dibimbing melalui proses peninjauan esai yang terstruktur dengan umpan balik yang segera, relevan, dan disertai catatan akan menghasilkan esai dengan kualitas keseluruhan yang lebih baik. Bimbingan ini dapat meminimalkan atau bahkan meniadakan perlunya peninjauan berulang oleh pengajar, serta memungkinkan terjadinya percakapan yang lebih substantif (dan kemungkinan lebih menarik serta memikat) antara pengajar dan mahasiswa mengenai tulisan mereka. Umpan balik semacam ini akan sangat bermanfaat bagi mahasiswa dengan tingkat kemampuan yang lebih rendah karena dapat memberikan dukungan bertahap (*scaffolding*) yang sering kali tidak sempat diberikan oleh pengajar. Seiring mahasiswa berinteraksi secara bermakna dengan Cambi dari waktu ke waktu, mereka akan mengembangkan kebiasaan meninjau tulisan sendiri serta memahami elemen-	Didirikan

	elemen dasar esai; hal ini pada gilirannya akan membantu mereka menjadi penulis yang lebih baik secara keseluruhan. Aspek kunci dari pernyataan ini adalah bahwa dukungan yang diberikan oleh Cambi tidak mendorong mahasiswa untuk terlalu bergantung pada AI, dan pengajar tetap terlibat aktif dalam memberikan umpan balik serta arahan terkait.	

Evaluasi Hasil

Teori tindakan (*theory of action*) mengajukan klaim mengenai hasil yang diharapkan serta mekanisme kausal yang mendasari klaim tersebut. Evaluasi hasil mengkaji sejauh mana klaim-klaim ini didukung oleh bukti. Sebagaimana dicatat oleh NCME, "...teori tindakan harus dievaluasi secara berkala menggunakan bukti empiris untuk memeriksa apakah teori tersebut berjalan sebagaimana mestinya." Di bawah ini, kami menguraikan empat aspek yang perlu dipertimbangkan dalam mengevaluasi sistem AWE (*Automated Writing Evaluation*): persepsi, implementasi, dampak langsung/jangka pendek terhadap tulisan mahasiswa, serta dampak jangka panjang/distal terhadap keterampilan menulis mahasiswa.

Alat yang sudah matang akan memiliki bukti yang mencakup seluruh aspek tersebut, yang mengarah pada penilaian dampak jangka panjang; sementara alat yang belum sepenuhnya matang mungkin hanya memiliki bukti mengenai dampak jangka pendek atau bukti terbatas terkait implementasi maupun persepsi. Dalam semua kasus, jumlah, keterwakilan, dan jenis partisipan studi merupakan pertimbangan penting dalam desain studi dan interpretasi hasil.

Sebagian besar penggunaan AWE oleh pengajar kemungkinan bersifat sukarela, sehingga langkah-langkah untuk memitigasi dampak dari sampel *convenience* (sampel yang dipilih berdasarkan kemudahan akses dan bukan secara acak) menjadi perlu dilakukan; hal ini menuntut penggunaan kelompok pembanding yang disetarakan (*matched comparison groups*) atau metode lain untuk mendekati skenario kontrafaktual, yang pada gilirannya membatasi sejauh mana temuan tersebut dapat digeneralisasi. Kendati demikian, pengumpulan bukti dari berbagai studi diharapkan dapat memberikan dukungan bagi atau justru menyanggah klaim-klaim dalam teori tindakan tersebut.

Teori sosiokultural mengenai asesmen formatif menyatakan bahwa hasil pembelajaran mencakup metakognisi dan regulasi diri, serta bahwa pembelajaran terjadi dalam konteks sosial kelas yang dinamis melibatkan pengajar dan mahasiswa maupun dalam konteks yang lebih luas terkait persyaratan dan kebijakan sekolah, distrik, dan negara bagian. Dengan menggunakan sudut pandang ini, persepsi mahasiswa dan pengajar menjadi aspek yang sangat penting dalam teori tindakan yang matang, karena persepsi tersebut dapat menjadi hasil dari teori tindakan itu sendiri (misalnya, mahasiswa merasa lebih percaya diri dalam menulis) atau menjadi faktor perantara dalam mencapai suatu hasil (misalnya, mahasiswa menganggap alat tersebut bermanfaat dan menarik, sehingga mereka menggunakannya).

Memahami bagaimana pengajar memandang alat tersebut dalam kaitannya dengan teori pengajaran mereka, kurikulum yang diterapkan, standar negara bagian, sikap terhadap AI, dan praktik di kelas jelas merupakan sudut pandang penting untuk memahami hasil yang dicapai. Terkait dengan hal tersebut, penting juga untuk memahami bagaimana mahasiswa memandang alat bantu ini dalam kaitannya dengan persepsi mereka mengenai keterampilan menulis, minat terhadap teknologi dan AI, serta kegunaan alat tersebut bagi kegiatan menulis mereka. Studi evaluasi hasil yang komprehensif akan mempertimbangkan persepsi pengajar maupun mahasiswa dalam desain studinya, baik untuk menafsirkan hasil akhir secara lebih baik maupun sebagai bagian dari hasil itu sendiri. Umpan balik semacam itu dapat diperoleh melalui survei dan wawancara yang berfokus pada klaim-klaim utama dari teori tindakan (*theory of action*) yang mendasarinya.

Implementasi mengacu pada apakah dan bagaimana mahasiswa serta pengajar menggunakan alat bantu AWE tersebut. Tingkat implementasi dapat ditelaah melalui berbagai cara, termasuk laboratorium kegunaan (*usability labs*), laboratorium kognitif, pembelajaran data proses/log penulisan, dan Pengembangan teknologi masa depan kelas. Laboratorium kegunaan dan kognitif melibatkan pengamatan terhadap satu atau beberapa mahasiswa saat mereka berinteraksi dengan alat bantu AWE. Dalam studi kegunaan, fokus utamanya biasanya terletak pada apakah mahasiswa memahami cara menggunakan alat tersebut beserta berbagai fiturnya. Sifat dan jenis umpan balik yang diberikan oleh alat berbasis AI dapat mempersulit pembelajaran ini, terutama ketika respons yang muncul bersifat dinamis dan beragam. Hasil dari studi kegunaan ini akan bersifat formatif; artinya, hasil tersebut akan mendorong perubahan pada UI/UX, perubahan pada umpan balik penilaian, atau menghasilkan pengembangan sumber daya pendukung tambahan. Laboratorium kognitif terutama berfokus pada bagaimana mahasiswa berinteraksi dengan alat tersebut untuk menyunting esai mereka.

Dalam laboratorium ini, mahasiswa mungkin diminta untuk "berpikir lantang" (*think aloud*) saat berinteraksi dengan alat tersebut, guna menggambarkan bagaimana mereka menafsirkan umpan balik dan tindakan apa yang mereka ambil terhadap esai mereka sebagai respons atas umpan balik tersebut. Pengaturan, pengodean, dan pembelajaran dalam laboratorium kognitif kemungkinan besar juga akan menjadi rumit akibat sifat dan jenis umpan balik yang diberikan oleh sistem, serta variasi dalam esai mahasiswa dan keterampilan menyunting mahasiswa. Laboratorium ini memberikan wawasan mengenai apa yang dipikirkan dan dilakukan mahasiswa sebagai respons terhadap umpan balik, serta bagaimana mereka mempertimbangkan revisi esai mereka. Hal-hal tersebut kemudian dapat menjadi dasar bagi evaluasi data proses/log penulisan. Data proses menelaah informasi tersimpan yang dikumpulkan mengenai cara mahasiswa menggunakan alat tersebut yaitu layar apa yang mereka akses, kapan mereka menyunting esai (serta hubungan antara umpan balik dan perilaku penyuntingan), bagaimana mereka menyunting esai, berapa lama waktu yang mereka habiskan untuk setiap bagian pada tiap tahapan proses penulisan, serta bagaimana mereka berinteraksi dengan pengajar atau satu sama lain. Yang terpenting, data proses perlu menyelaraskan mekanisme umpan balik dengan penyuntingan itu sendiri, serta berpotensi

menyediakan sumber interaksi dan penyuntingan yang kaya untuk dievaluasi terutama jika dipembelajaran bersamaan dengan data laboratorium kognitif. Dengan demikian, pembelajaran data proses tidak hanya memberikan konteks mengenai cara mahasiswa menggunakan perangkat tersebut, tetapi juga menawarkan kesempatan untuk menguji mekanisme kausal yang dikemukakan dalam teori tindakan.

Dampak jangka pendek berfokus pada pengaruh penggunaan alat AWE terhadap mahasiswa dan pengajar selama proses peninjauan dan revisi esai, dengan menilai apakah dan bagaimana esai mahasiswa mengalami peningkatan setelah berinteraksi dengan sistem, serta mahasiswa mana yang memperoleh manfaat. Dampak langsung ini dapat ditelaah dengan membandingkan skor berbasis rubrik (yang dinilai oleh pengajar, sistem otomatis, atau profesional) antara mahasiswa dengan kemampuan serupa yang berinteraksi dengan alat AWE dan mereka yang tidak. Pembelajaran ini dapat menunjukkan mahasiswa mana yang memperoleh manfaat serta apakah manfaat tersebut konsisten di berbagai kelompok mahasiswa atau kelas.

Selain skor rubrik, pembelajaran dapat mencakup perubahan pada esai sebagaimana diprediksi oleh teori tindakan (misalnya, penambahan frasa transisi saat diminta melakukannya, atau tidak adanya perubahan saat tidak ada dorongan untuk melakukan hal tersebut). Kesimpulan ini bergantung pada desain penelitian yang melibatkan uji coba terkontrol acak (*randomized controlled trials*) atau desain kuasi-eksperimental, dengan fokus pada keterwakilan pengajar dan mahasiswa untuk keperluan pengambilan kesimpulan dan generalisasi.

Dampak jangka panjang menelaah pengaruh penggunaan alat AWE terhadap kemampuan menulis mahasiswa secara keseluruhan dari waktu ke waktu, dengan teori tindakan sebagai panduan untuk menentukan aspek kemampuan spesifik yang dinilai. Rentang waktu penelitian dapat bervariasi, mulai dari penggunaan di kelas selama beberapa minggu hingga penggunaan berulang sepanjang tahun ajaran. Studi-studi ini mengharuskan penggunaan penilaian menulis independen yang dilakukan tanpa pemberian umpan balik. Tolok ukur eksternal ini dapat berupa tugas menulis pada sesi kelas berikutnya, perkembangan skor menulis dari waktu ke waktu, dan/atau kinerja mahasiswa dalam penilaian menulis tingkat negara bagian (baik yang bersifat sumatif maupun lainnya). Terakhir, studi-studi ini perlu mengumpulkan data mengenai persepsi dan praktik mahasiswa serta pengajar sebagai bahan pertimbangan dalam menafsirkan hasil penelitian.

Tabel 4.4 menyajikan tingkat kematangan untuk keempat alat AWE tersebut. Keempat tingkat kematangan ini ditemukan pada berbagai alat AWE, yang mencerminkan kumpulan bukti yang diperoleh untuk setiap alat terkait aspek persepsi, implementasi, serta dampak jangka pendek dan jangka panjang. Perlu dicatat bahwa deskripsi tersebut tidak berfokus pada hasil aktual, melainkan pada tujuan dan metode penelitian dalam menguji klaim teori tindakan. Perbedaan utama di antara alat-alat AWE tersebut terletak pada jumlah bukti yang tersedia dalam keempat aspek saat mengevaluasi hasil. MI Write memiliki pembelajaran bukti yang paling komprehensif, diikuti oleh eRevise+RF. Alat Cambi memiliki rangkaian studi yang terbatas untuk mendukung penggunaannya, dengan fokus pada persepsi dan dampak jangka

pendek. *Writing Coach* belum memublikasikan bukti untuk mendukung klaim mengenai teori tindakannya (*theory-of-action*).

Tabel 4.4 Tingkat Kematangan Alat AWE untuk Fokus Evaluasi Hasil

Alat AWE	Keterangan	Tingkat Kematangan
Pelatih Menulis	Belum ada bukti pendukung yang dipublikasikan untuk Writing Coach. Meskipun terdapat bukti hasil yang mendukung perangkat matematika Khanmig, belum jelas bagaimana studi-studi tersebut dapat diterapkan pada Writing Coach.	Awal
MI Write	MI Write memiliki rangkaian penelitian evaluasi hasil yang ekstensif, mencakup persepsi pengajar dan mahasiswa, penerapan di kelas, serta dampak jangka pendek dan jangka panjang. Di bawah ini, kami menyajikan sebagian studi utama yang mengilustrasikan pembelajaran evaluasi hasil terkait MI Write. Studi mengkaji tiga metode penerapan MI Write di kelas: satu metode menggunakan pendekatan *ad hoc* dalam pengajaran menulis dengan MI Write, satu lagi menggunakan pendekatan *Self-Regulated Strategy Development* (SRSD) dengan MI Write, dan satu kelompok dasar (*baseline*) tanpa penggunaan MI Write. Studi ini juga memantau penggunaan MI Write. Studi metode campuran yang bersifat kuasi-eksperimental ini melibatkan wawancara dan survei terhadap pengajar serta mahasiswa, serta mengumpulkan bukti fidelitas penerapan melalui berkas log untuk menelaah penggunaan sistem tersebut. Skor esai *pre-test*, fidelitas perlakuan, dan jenis perlakuan digunakan sebagai input dalam model bertingkat (*multi-level model*), dengan variabel dependen berupa skor akhir esai, panjang esai akhir, dan jumlah elemen komposisi dalam esai akhir. Data dikumpulkan di satu negara bagian, mencakup lima distrik, dengan melibatkan 19 pengajar dan sekitar 800 mahasiswa sekolah menengah pertama. Studi memperluas pembelajaran ini dengan	Dewasa

	berfokus pada perubahan esai mahasiswa (yang diukur berdasarkan skor keseluruhan, panjang esai, dan penggunaan elemen komposisi) di kelas-kelas yang menggunakan MI Write dengan kedua strategi tersebut, serta bagaimana persepsi mahasiswa dan pengajar berbeda di antara kedua strategi itu. Data juga dipembelajaran berdasarkan ras/etnis. Pembalajaran menggunakan kelompok fokus untuk mengkaji sikap dan pengalaman pengajar terhadap MI Write; studi ini berfokus pada bagaimana 90 pengajar di wilayah Mid-Atlantic memandang pengaruh AWE terhadap kemampuan menulis mereka sendiri maupun mahasiswa mereka, serta pandangan mereka mengenai potensi AWE dalam mentransformasi proses belajar-mengajar. Studi memperluas pembelajaran ini dengan menggunakan pembelajaran bertingkat terhadap data mahasiswa untuk menelaah perubahan kinerja penulisan esai sepanjang tahun ajaran. Studi mengkaji dampak jangka pendek dan jangka panjang penggunaan MI Write di tingkat sekolah dasar (kelas 3–5), dengan fokus pada faktor penerapan dan motivasi. Sampel penelitian terdiri dari 1.935 Mahasiswa dan 135 pengajar dari 14 sekolah dasar. Sikap pengajar dan mahasiswa terhadap MI Write dinilai melalui survei. Kinerja mahasiswa pada esai penilaian tingkat negara bagian dan skor esai pascates di MI Write digunakan sebagai variabel dependen. Variabel prediktor meliputi jumlah esai yang ditulis mahasiswa menggunakan MI Write, jumlah draf esai, rata-rata peningkatan skor antara draf pertama dan akhir, serta total waktu (dalam menit) penggunaan MI Write. Kovariat yang digunakan adalah skor esai pratest MI Write, rata-rata skor draf pertama MI Write, skor efikasi diri menulis mahasiswa, serta data demografis (jenis kelamin, ras, status pendidikan luar biasa, status pembelajar bahasa Inggris, dan persentase mahasiswa yang menerima makan siang	

	gratis atau bersubsidi di sekolah). Persepsi pengajar dan mahasiswa mengenai MI Write ditelaah, dan penggunaannya dinilai melalui berkas log. Pemodelan linear hierarkis digunakan untuk memprediksi kedua variabel dependen tersebut dengan menggunakan prediktor dan kovariat yang ada. Pembelajaran menggunakan uji coba terkontrol secara acak (pada tingkat kelas) untuk meneliti dampak penggunaan MI Write terhadap skor esai sumatif dan menemukan tidak adanya dampak selama kurun waktu satu tahun; namun, mereka pembahasan skor berdasarkan tingkat penggunaan dan menemukan adanya dampak pada sekolah-sekolah yang menggunakan MI Write secara tepat dan sering. Model hierarkis tersebut mencakup kelompok perlakuan, skor pratest, status pendidikan luar biasa, dan kontrol populasi prioritas. Mereka juga meneliti fidelitas perlakuan, efikasi diri menulis, serta tingkat kesenangan mahasiswa dalam menulis. Sampel penelitian terdiri dari 2.547 mahasiswa sekolah menengah (kelas 7 dan 8) dari 14 sekolah di tiga distrik. Para penulis juga pembahasan kinerja berdasarkan distrik. Studi memperluas evaluasi ini dengan berfokus pada faktor-faktor yang memengaruhi persepsi mahasiswa terhadap MI Write.	
eRevisi + RF	eRevise+RF didukung oleh serangkaian penelitian evaluasi hasil yang kuat, yang mencakup persepsi pengajar dan mahasiswa, penerapannya di kelas, serta dampak jangka pendeknya. Penelitian ini terutama berfokus pada dua negara bagian dan menggunakan dua topik penulisan (*prompts*). Pengajar dan mahasiswa mengisi survei mengenai pengalaman serta penggunaan perangkat tersebut. Ke-16 pengajar peserta menjalani wawancara semi-terstruktur yang membahas pengalaman mahasiswa menggunakan perangkat, kesesuaian umpan balik	Didirikan

	dengan tujuan pengajar, potensi dampak perangkat terhadap pengajaran menulis, serta seberapa sering pengajar menggunakan perangkat tersebut. Survei mahasiswa berfokus pada pemahaman mahasiswa terhadap umpan balik, tindakan yang mereka ambil terkait umpan balik tersebut, serta sejauh mana saran yang diberikan mereka terapkan saat merevisi esai. Pengajar juga diminta mencatat log pelaksanaan yang memuat pertanyaan mahasiswa selama penggunaan perangkat serta interaksi mereka dengan mahasiswa; data ini kemudian dibandingkan dengan nilai mahasiswa untuk menentukan dampak interaksi pengajar terhadap kinerja menggunakan model linear hierarkis. Evaluasi terhadap peningkatan kualitas jangka pendek didasarkan pada jumlah dan jenis bukti yang disertakan, alasan di balik penggunaan bukti tersebut, serta integrasi bukti dan alasan di berbagai draf yang ditulis selama tiga. Para peneliti juga mengkaji tingkat peningkatan antar-draf berdasarkan jenis umpan balik yang diberikan dan kualitas draf awal untuk mengeksplorasi bagaimana umpan balik tersebut berkaitan dengan revisi yang dilakukan mahasiswa.	
Cambi	Terdapat sejumlah kecil studi mengenai Cambi yang berfokus pada persepsi pengajar, dampak langsung penggunaan Cambi terhadap nilai esai dalam uji coba terkontrol acak, serta pembelajaran perkembangan esai selama satu sesi penulisan. Wawancara semi-terstruktur dengan pengajar dilakukan untuk mengukur sejauh mana sampel yang dipilih berdasarkan kemudahan akses (*convenience sample*) yang terdiri dari 20 pengajar di dua negara bagian menganggap alat tersebut selaras dengan metode pengajaran mereka sebelum digunakan, serta untuk memahami pandangan mereka setelah menggunakannya di kelas. Data wawancara digunakan untuk menyempurnakan alat tersebut dan	Mengembangkan

	menentukan tingkat kesesuaiannya dengan praktik pengajaran mereka. Sebuah studi berskala kecil yang melibatkan Cambi dilakukan terhadap sampel 11 kelas yang dipimpin oleh pengajar di dua negara bagian; dalam studi ini, mahasiswa secara acak dibagi ke dalam kelompok yang mendapatkan akses ke umpan balik Cambi dan kelompok yang tidak mendapatkan akses tersebut namun menerima panduan statis yang serupa dengan panduan Cambi. Variabel terikat dalam studi ini adalah nilai rubrik yang dihasilkan oleh sistem penilaian otomatis. Hasil dari kedua kelompok dibandingkan menggunakan nilai rata-rata dan distribusi nilai. Studi ini juga meneliti jenis-jenis perubahan yang dilakukan mahasiswa (penambahan teks serta penyuntingan di dalam esai yang mencakup dua kategori: tingkat permukaan dan substantif) selama sesi penulisan pada kedua kelompok tersebut.	

Teknologi AI

Teknologi AI memampukan alat AWE dengan mendasari kemampuan untuk memberikan umpan balik secara langsung. Teknologi AI dapat mencakup berbagai hal, mulai dari pendekatan klasik berbasis fitur hingga model bahasa besar (*large language models* atau LLM) berbasis *transformer* yang disesuaikan (*fine-tuned*) untuk prediksi (misalnya, BERT), hingga penggunaan *prompting* dengan model AI generatif seperti ChatGPT. Terlepas dari teknologi AI yang dipilih, dasar pemikiran untuk semua elemen desain mesin, metode pelatihan atau *prompting* AI, serta penggunaan data sangatlah penting bagi kematangan alat AWE. Meskipun jelas bahwa teknologi AI harus selaras dengan teori tindakan, tujuan, dan metode, kemungkinan terdapat banyak pendekatan kandidat dalam memilih teknologi AI. Alat yang matang akan mempertimbangkan semua faktor tersebut. Alat tahap awal mungkin hanya memberikan sedikit atau bahkan tidak ada penjelasan mengenai setiap faktor. Alat yang sedang dikembangkan akan memiliki penjelasan minimal, sedangkan alat yang sudah mapan akan memiliki penjelasan, namun tingkat alasan di balik pilihan-pilihan tersebut mungkin belum lengkap. Di bawah ini, kami menguraikan elemen-elemen kunci teknologi AI: arsitektur, pelatihan, dan data.

Arsitektur mesin merupakan faktor krusial dalam teknologi AI. Yang dimaksud dengan arsitektur di sini adalah desain mesin yang menghasilkan *output*, serta sifat dari *input* dan *output* tersebut. Versi awal sistem AWE menggunakan fitur yang dirancang oleh pakar dengan fokus pada identifikasi kesalahan ejaan atau tata bahasa, deteksi pola kata yang terkait dengan

esai berkualitas tinggi, dan penilaian karakteristik pembentukan kalimat. Fitur-fitur ini kemudian digunakan dalam model statistik parametrik (misalnya, regresi multivariat) atau model non-parametrik (misalnya, *gradient-boosted regression trees* atau *support vector machines*) untuk memprediksi skor atau nilai lainnya. Metode yang lebih baru memanfaatkan *embedding* model bahasa besar sebagai "fitur implisit" untuk prediksi, atau memanfaatkan AI generatif melalui prompting untuk mendukung mahasiswa. Yang penting, metode-metode ini dapat menggunakan LLM sumber tertutup (misalnya, OpenAI) melalui *application programming interface* (API) atau LLM sumber terbuka (misalnya, Llama 3, deBERTa) yang tersedia untuk diunduh dan disesuaikan (*fine-tuning*) melalui pusat perangkat lunak seperti *HuggingFace*.

Perbedaan penting antara pendekatan *closed-source* dan pendekatan *open-source* atau kustom adalah bahwa arsitektur mesin yang mendasari prediksi pada model *closed-source* tidak dipublikasikan dan dikelola oleh vendor eksternal (misalnya, OpenAI, Amazon, atau Anthropic). Dalam kasus terakhir ini, hanya kode yang melingkupi LLM yang dapat dijabarkan secara eksplisit (misalnya, panggilan API dan *prompt*); namun, LLM itu sendiri tidak dapat dideskripsikan. Alat AWE yang sudah matang menunjukkan dasar pemikiran yang jelas dalam penggunaan pendekatan tertutup (*closed*), terbuka (*open*), atau khusus (*custom*). Untuk alat AWE yang menggunakan pendekatan terbuka atau khusus, alasan pemilihan model itu sendiri harus dijelaskan; biasanya, pemilihan ini didasarkan pada kinerja sebelumnya dalam tugas terkait, data dan metode yang digunakan untuk melatih awal (*pre-train*) model (dalam hal LLM), jumlah parameter model dan arsitektur *transformer*, serta batasan apa pun terkait fitur esai (misalnya, panjang jendela konteks).

Kompleksitas penulisan kemungkinan besar akan menuntut penggunaan beberapa mesin pendukung sistem AWE, dan masing-masing mesin tersebut harus dijelaskan. Terakhir, alat yang sudah matang harus menjelaskan sifat keluaran model, seperti label yang dikaitkan dengan kalimat dalam esai (misalnya, melabeli kalimat sebagai bukti atau alasan), label untuk frasa (misalnya, mengidentifikasi kesalahan tata bahasa dalam kalimat), dan skor yang diberikan untuk keseluruhan esai, yang biasanya dikaitkan dengan deskripsi rubrik.

Pendekatan dalam melatih mesin juga merupakan elemen kunci untuk memahami teknologi AI tersebut. Sistem yang sudah matang akan menjelaskan proses ini secara rinci, dengan menguraikan semua komponennya. Dalam beberapa kasus, mungkin tidak ada proses *fine-tuning* (penyesuaian halus), melainkan hanya penggunaan *prompting* pada LLM. Dalam kasus lain mungkin hanya ada *fine-tuning* minimal, seperti memberikan contoh kepada LLM sebagai bagian dari *prompt* atau mengombinasikannya dengan teknik temu balik informasi (*information retrieval*) untuk memberikan *prompt* secara dinamis kepada LLM.

Pada kasus lainnya, *fine-tuning* atau pelatihan dapat secara langsung memperbarui bobot model guna meningkatkan kinerja tugas prediksi, alih-alih sekadar memberikan *prompt* agar model menghasilkan keluaran. Pendekatan ini biasanya digunakan pada LLM sumber terbuka (*open-source*) dan arsitektur khusus untuk mengoptimalkan kinerja tugas prediksi, serta terbukti cukup berhasil dalam menghasilkan keluaran yang menyerupai hasil kerja manusia. Alat AWE yang sudah matang menyajikan rincian lengkap mengenai hiperparameter

atau fitur yang digunakan untuk melakukan *fine-tuning* atau melatih model-model tersebut. Dalam kasus LLM sumber terbuka (*open-source*), hal ini dapat mencakup jumlah *epoch*, ukuran *batch*, laju pembelajaran (*learning rate*), fungsi kerugian (*loss function*), dan kriteria penghentian.

Terakhir, perangkat AWE yang matang (*mature*) menjabarkan data yang digunakan untuk melakukan *fine-tuning* atau pemberian *prompt* pada model, termasuk alasan di balik pemilihan data dan pendekatan pengelolaan datanya. Sebagai contoh, data harus merepresentasikan populasi sasaran yang akan menerima umpan balik tersebut, menunjukkan kualitas pelabelan dan penilaian oleh manusia yang memadai, serta memiliki ukuran sampel yang cukup untuk pelatihan dan validasi model AI.

Menurut pandangan kami, menghasilkan data berkualitas tinggi, berlabel, dan representatif harus menjadi fokus utama sistem AWE dan sering kali menjadi salah satu komponen biaya terbesar dalam pengembangannya. Bidang ini pula yang menuntut keahlian substansial dalam pelabelan dan penilaian oleh manusia guna memperoleh informasi yang mencerminkan konstruk yang dituju. Mengingat model AI berkembang pesat, data validasi yang sama dapat digunakan untuk memantau evolusi model dan menelaah peningkatan kinerja atau untuk memperoleh jenis umpan balik tambahan.

Akhirnya, metrik dan kriteria untuk mengevaluasi pelabelan manusia (dan pelabelan AI) juga sangat penting dalam proses *fine-tuning*, karena hal-hal tersebut menentukan tolok ukur keberhasilan model serta memengaruhi pemilihan arsitektur mesin dan hiperparameter yang digunakan dalam pelatihan. Data yang dihasilkan secara sintetis seperti yang dibuat oleh AI generatif dan digunakan dalam pelatihan harus dijelaskan secara rinci, begitu pula metode berbasis AI apa pun yang digunakan untuk memantau keluaran mesin. Tabel 4.5 menyajikan tingkat kematangan teknologi AI beserta deskripsinya untuk keempat perangkat AWE tersebut. *Writing Coach* mendapatkan peringkat "Awal" (*Early*) karena minimnya penelitian mengenai teknologi AI yang digunakan dalam perangkat tersebut; namun, kami tetap menyajikan informasi mengenai teknologi AI tutor matematika Khanmigo, dengan asumsi bahwa informasi tersebut akan menjadi acuan bagi penerapannya dalam *Writing Coach*.

Tabel 4.5 Tingkat Kematangan Alat AWE untuk Fokus Teknologi AI

Alat AWE	Keterangan	Tingkat Kematangan
Pelatih Penulisan	Tidak ada deskripsi mengenai teknologi AI yang digunakan, namun diperkirakan *Writing Coach* akan memanfaatkan model GPT-4 yang bersifat *closed-source* dengan teknik *prompting*. Penelitian mengenai bimbingan belajar matematika menunjukkan penggunaan model baik yang bersifat *open-source* maupun *closed-source* saat AI menilai ketepatan respons mahasiswa, serta saat alat AI menghasilkan teks berdasarkan respons	Lebih awal

	mahasiswa dan menilai sejauh mana teks tersebut tepat (misalnya, menafsirkan respons mahasiswa dengan benar, mengajukan pertanyaan yang memadai terkait respons tersebut, atau membimbing mahasiswa menuju jawaban yang benar). Khan Academy telah merilis kumpulan data untuk memodelkan ketepatan respons matematika mahasiswa. Terakhir, para peneliti telah mengkaji cara penggunaan data sintetis untuk mempercepat waktu penerapan soal matematika baru. Kami memperkirakan metodologi-metodologi ini akan diadaptasi untuk *Writing Coach*.	
MI Write	Penilaian dilakukan oleh *Project Essay Grade* (PEG) menggunakan model khusus berbasis fitur klasik yang mengacu pada model *Six Trait* (gagasan, organisasi, gaya, kelancaran kalimat, pemilihan kata, dan konvensi), di mana rubrik untuk setiap aspek memiliki lima tingkatan. Model dilatih untuk setiap aspek, tiga genre (naratif, informatif, argumentatif), dan lima rentang kelas (3–4, 5–6, 7–8, 9–10, 11–12). Model dilatih sedemikian rupa sehingga skor selaras secara vertikal lintas rentang kelas dalam satu genre (artinya, esai yang sama akan mendapatkan skor lebih rendah dari model kelas 3–4 dibandingkan dari model kelas 5–6). Model ini disesuaikan untuk revisi esai dengan memeriksa perubahan pada variabel-variabel kunci untuk setiap aspek dan bagaimana perubahan tersebut memengaruhi gradien; skor AI dimaksudkan untuk merepresentasikan peningkatan skor berdasarkan perubahan-perubahan tersebut. Mahasiswa dapat menerima hingga lima pernyataan umpan balik untuk setiap aspek yang kemungkinan besar terkait dengan skor otomatis serta menerima umpan balik tata bahasa yang disisipkan langsung dalam teks. Deskripsi mengenai data yang	Didirikan

	digunakan untuk melatih sistem tersebut tidak tersedia dalam riset.	
eRevise + RF	Model khusus berbasis fitur digunakan untuk memberikan umpan balik mengenai *Number of Pieces of Evidence* (NPE) yang menunjukkan jumlah topik dalam esai serta *specificity* (SPC) yang menunjukkan jumlah detail bukti. Daftar yang disusun secara manual disediakan untuk mesin NPE dan SPC, dan kata-katanya dikodekan menggunakan *embedding* sumber terbuka *(open-source)*. Teknik *sliding window* diterapkan pada esai untuk memperoleh nilai hitung NPE dan SPC. NPE dan SPC dimaksudkan untuk merepresentasikan keluasan dan kedalaman bukti yang digunakan dalam esai. Saat ini, hanya dua *prompt* yang dievaluasi dalam sistem tersebut. Mahasiswa memiliki dua kesempatan untuk merevisi esai mereka. Pada draf pertama, ukuran NPE dan SPC diberikan kepada mahasiswa sebagai umpan balik. Setelah draf kedua diserahkan, sistem menggunakan model khusus berbasis fitur untuk menilai sifat revisi berdasarkan skor NPE dan SPC. Sistem tersebut membandingkan kedua draf dan mengidentifikasi kalimat yang direvisi atau kalimat baru. Sebuah pengklasifikasi LLM sumber terbuka (*open-source*), yang telah disesuaikan (*fine-tuned*) menggunakan esai tingkat perpengajaran tinggi, digunakan untuk mengklasifikasikan revisi sebagai revisi tingkat permukaan atau revisi konten. Pengklasifikasi sumber tertutup (*closed-source*) terpisah (melalui GPT-4) kemudian digunakan untuk memprediksi apakah revisi yang teridentifikasi merupakan bukti atau penalaran, serta untuk mengidentifikasi ruang konteks dari argumen itu sendiri. Terakhir, pengklasifikasi sumber terbuka yang telah disesuaikan digunakan untuk menentukan	Dewasa

	apakah revisi tersebut berhasil atau tidak berhasil. Klasifikasi bukti/penalaran dan keberhasilan/ketidakberhasilan menghasilkan empat label revisi: bukti berhasil, bukti tidak berhasil, penalaran berhasil, dan penalaran tidak berhasil. Label-label ini kemudian dipetakan ke umpan balik yang diberikan kepada mahasiswa menggunakan kombinasi skor NPE dan SPC serta skor revisi: umpan balik untuk memberikan lebih banyak bukti, umpan balik untuk memberikan lebih banyak detail, atau umpan balik untuk menawarkan penjelasan lebih lanjut mengenai hubungan antara bukti dan argumen. Mahasiswa kemudian dapat menggunakan umpan balik ini dalam revisi ketiga mereka. Rincian yang menguraikan teknologi AI tersebut dapat ditemukan dalam studi. Penelitian yang mendasari pengembangan berbagai elemen sistem ini termuat dalam sejumlah artikel: pengembangan rubrik revisi, pemodelan keberhasilan revisi, pengembangan NPE dan SPC serta penyusunan daftar kata.	
Cambi	Model *transformer* sumber terbuka yang telah disesuaikan secara khusus (*fine-tuned*) digunakan untuk seluruh umpan balik AI. Model-model tersebut dilatih untuk mengidentifikasi elemen komposisi dan elemen konvensi. Model elemen komposisi dilatih menggunakan data dari dua rentang jenjang kelas (3–5 dan 6–8) serta dua genre tulisan (argumentatif dan informatif). Data pelatihan mencakup respons mahasiswa dari program penilaian sumatif di suatu negara bagian selama beberapa tahun. Model-model tersebut dilatih menggunakan kalimat yang telah diberi label oleh penilai manusia berdasarkan prosedur penilaian standar. Berbagai model dilatih menggunakan elemen-elemen konvensi penulisan. Satu	Dewasa

	model *transformer* disesuaikan (*fine-tuned*) menggunakan kalimat mahasiswa yang telah diberi label oleh manusia dari negara bagian yang sama untuk mengidentifikasi apakah kalimat tersebut merupakan *run-on* (kalimat yang menyambung tanpa tanda baca yang tepat), fragmen, atau kalimat normal. Model *transformer* lainnya disesuaikan menggunakan dataset yang tersedia untuk umum untuk menerima masukan berupa kalimat dan menghasilkan keluaran berupa kalimat yang telah diperbaiki. Keluaran dari model ini kemudian diproses menggunakan alat tata bahasa berbasis aturan yang mengidentifikasi jenis dan sifat perbaikan dalam kalimat tersebut. Kedua model ini kemudian menghasilkan pembelajaran konvensi berbasis kalimat yang mengidentifikasi jenis kesalahan konvensi dalam suatu kalimat. Seluruh keluaran AI untuk suatu respons kemudian dipetakan ke kode umpan balik yang selaras dengan rubrik standar tingkat negara bagian. Umpan balik yang disusun khusus untuk Mahasiswa disediakan dan ditautkan ke bagian-bagian tertentu yang disorot dalam respons tersebut. Rincian yang menguraikan teknologi AI ini dapat ditemukan dalam pembelajaran.	

Baik eRevise+RF maupun Cambi memperoleh peringkat "Matang" (*Mature*) karena dokumentasi pengembangan AI-nya sangat komprehensif, mencakup data yang digunakan untuk melatih model, alasan dan pembelajaran pemilihan model, serta keselarasan antara keluaran model dan umpan balik yang diberikan. Perangkat MI Write hanya mendapatkan tingkat "Mapan" (*Established*) karena data yang digunakan untuk melatih model tidak dijelaskan dalam riset, dan pemodelannya hanya dipaparkan secara garis besar. Kendati demikian, proses pemodelan secara keseluruhan telah dijelaskan, begitu pula kaitan antara pemodelan tersebut dan umpan balik.

Validasi Model

Evaluasi kinerja model AI biasanya dilakukan dengan membandingkannya terhadap kinerja manusia menggunakan metrik dan kriteria yang terdefinisi dengan jelas serta diselaraskan dengan definisi konstruk yang diuraikan dalam teori tindakan (*theory of action*). Tujuan validasi model adalah untuk memahami sejauh mana keluaran AI mendekati kinerja

manusia dengan cara yang dapat digeneralisasikan ke populasi yang lebih luas. Pembelajaran ini biasanya didasarkan pada data yang sama dengan yang digunakan untuk melatih mesin AI, atau pada data yang dikumpulkan setelah mesin tersebut diterapkan dalam perangkat AWE. Untuk memastikan model tidak bias, pembelajaran harus mencakup tidak hanya kinerja secara keseluruhan, tetapi juga kinerja pada subkelompok mahasiswa utama (misalnya, ras/etnis, jenis kelamin, bahasa utama, disabilitas, status sosial-ekonomi), kelas, atau wilayah. Hasil evaluasi dapat menjadi dasar pengambilan keputusan mengenai apakah, apa, dan bagaimana keluaran yang dihasilkan AI digunakan dalam umpan balik penilaian.

Dalam kasus di mana prediksi bersifat kategorikal atau kontinu, pendekatan validasinya relatif mudah dan jelas. Di sini, strategi berfokus pada metrik dan kriteria kesepakatan seperti akurasi, korelasi, nilai kappa, skor F1, dan selisih rata-rata terstandarisasi (*standardized mean difference*), dengan menggunakan kesepakatan manusia sebagai tolok ukur. Dalam kasus di mana prediksi bersifat kualitatif dan/atau iteratif seperti pada penggunaan chatbot atau umpan balik teks yang dihasilkan AI proses validasinya belum terdefinisi dengan jelas dan merupakan bidang penelitian yang sedang berkembang. Untuk pendekatan ini, proses validasi perlu mempertimbangkan interaksi dinamis antara mahasiswa dan sistem AI selama sesi pemberian umpan balik, sembari mengklasifikasikan respons mahasiswa secara tepat di berbagai jenis jawaban mahasiswa, memberikan jawaban yang akurat namun tidak mengarahkan (*non-leading*), serta mengelola pergeseran konseptual mahasiswa dalam percakapan.

Penggunaan chatbot memerlukan teori tindakan yang kuat yang menjabarkan cakupan seluruh masukan dan keluaran yang wajar dalam interaksi, serta penggunaan rubrik dan metode yang jelas untuk mengevaluasi akurasi dan kesesuaian masukan dan keluaran tersebut. Bidang pekerjaan ini masih dalam tahap awal, mengingat baru munculnya chatbot AI dan integrasinya ke dalam perangkat AWE. Kami mengharapkan alat AWE yang memiliki validasi model tingkat *Mature* (matang) untuk mempertimbangkan semua aspek yang dijelaskan di atas, disertai bukti ekstensif yang mendukung klaim bahwa model AI tersebut berkinerja baik.

Alat AWE yang tergolong *Established* (mapan) akan mempertimbangkan metrik-metrik utama namun mungkin belum mencakup pembelajaran berdasarkan kelompok demografis. Alat AWE yang sedang dikembangkan mungkin hanya menggunakan sebagian dari metrik-metrik utama tersebut. Alat AWE tahap awal (*Early*) tidak memiliki metrik terkait kinerja model. Tingkat kematangan alat AWE untuk validasi model disajikan dalam Tabel 4.6. *Writing Coach* dikategorikan sebagai *Early* karena belum ada publikasi hasil mengenai kinerjanya. Tiga alat AWE lainnya dikategorikan sebagai *Established*; hal ini dikarenakan hasil kinerja tidak tersedia secara spesifik untuk model yang digunakan dalam alat AWE tersebut (contohnya MI Write) melainkan hanya tersedia secara umum untuk mesin (*engine*) dasarnya atau karena hasil tersebut tidak meninjau kinerja berdasarkan kelompok mahasiswa (contohnya eRevise+RF dan Cambi).

Tabel 4.6 Tingkat Kematangan Perangkat AWE untuk Fokus Validasi Model

Alat AWE	Keterangan	Tingkat Kematangan
Pelatih Penulisan	Hasil evaluasi model untuk Writing Coach belum tersedia. Kami memperkirakan bahwa hasilnya akan mengacu pada temuan dari evaluasi bidang matematika, termasuk akurasi penilaian hasil kerja mahasiswa serta sifat dan kualitas umpan balik otomatis. Untuk setiap aspek tersebut, label manusia dan respons operasional digunakan sebagai acuan kebenaran (*ground truth*), dengan evaluasi yang mencakup kesesuaian persis (*exact agreement*), *Krippendorff's alpha*, koefisien korelasi *Matthews*, *Fleiss' kappa*, presisi, dan *recall*.	Awal
MI Write	Meskipun tampaknya tidak ada bukti yang dipublikasikan mengenai model-model yang digunakan dalam MI Write, para penulis berpendapat bahwa model-model tersebut memanfaatkan PEG. Kualitas penilaian PEG sebagaimana diukur melalui *quadratic weighted kappa*, kesesuaian persis (*exact agreement*), dan selisih rata-rata terstandarisasi (*standardized mean difference*) pada berbagai tingkat kelas, genre, dan rubrik telah dilaporkan dalam sejumlah artikel.	Didirikan
eRevisi + RF	Akurasi metrik eRevise+RF NPE, SPC, dan revisi dibandingkan dengan penilaian manusia; kesepakatan persis (*exact agreement*), korelasi intra-kelas, dan *quadratic weighted kappa* digunakan untuk menguji tingkat kesepakatan berdasarkan respons dan skor operasional. Skor NPE dan SPC juga dipembelajaran sensitivitasnya terhadap perubahan dalam revisi mahasiswa dengan mengamati peningkatan skor dikaitkan dengan tingkat dan jenis perubahan tersebut. Pekerjaan ini dilakukan melalui serangkaian studi seiring dengan evaluasi model AI untuk tujuan penggunaan. Studi dilaksanakan di dua negara bagian menggunakan dua topik penulisan (*prompt*)	Didirikan

	serta melibatkan jumlah pengajar dan mahasiswa yang relatif terbatas. Hasil penelitian ini dipaparkan dalam berbagai artikel.	
Cambi	Data yang diberi label oleh manusia dari satu negara bagian digunakan untuk mengevaluasi model elemen komposisi. Rubrik anotasi dirancang dan digunakan selama proses pelabelan oleh manusia, dan 15% esai mendapatkan pelabelan kedua oleh manusia. Kinerja mesin AI dievaluasi dengan membandingkannya terhadap label manusia menggunakan metrik kappa dan kesepakatan persis (*exact agreement*). Model lintas genre dan tingkat kelas juga dinilai untuk menguji ketangguhan kinerja model di berbagai rentang tingkat kelas dan genre. Untuk rincian lebih lanjut. Model pembentukan kalimat dievaluasi dalam dua tahap: pertama, menggunakan kumpulan data kalimat berlabel manusia yang disisihkan (*held-out set*), sekali lagi menggunakan kappa dan kesepakatan persis; kedua, menggunakan kumpulan data sumber terbuka serta memeriksa persentase positif palsu dan positif benar. Model konvensi akhir juga dievaluasi dalam dua tahap: pertama, menggunakan kumpulan data uji yang tersedia untuk umum dan metrik *Generalized Language Evaluation Understanding* (GLEU); kedua, menggunakan kumpulan data sumber terbuka serta memeriksa positif palsu dan positif benar.	Didirikan

Keselamatan dan Keamanan Data

Perangkat AWE yang matang harus menjamin keselamatan, privasi, dan keamanan mahasiswa serta data mereka. Aspek-aspek ini ditekankan dalam pedoman AI OET, *Standards*, pedoman NIST, peraturan perundang-undangan federal (misalnya, *Family Educational Rights and Privacy Act* [FERPA], *Protection of Pupil Rights Amendment* [PPRA], dan *Children's Online Privacy Protection Act* [COPPA]), serta pedoman dan hukum tingkat negara bagian (misalnya, *Connecticut Data Privacy Act*). Faktanya, *Standards* itu sendiri menguraikan perlunya melindungi data mahasiswa sebagai bagian dari praktik pengukuran yang baik terkait hak dan

tanggung jawab peserta tes (lihat Standar 8.5 dan 8.6). Isu-isu ini menjadi sangat penting di era LLM dan AI.

Penggunaan model sumber tertutup (*closed-source*) dan lanskap hukum yang terus berkembang di sekitarnya menimbulkan kekhawatiran serius. Vendor mungkin memiliki pendekatan yang berbeda-beda dalam menangani privasi dan keamanan data pengguna; oleh karena itu, diperlukan kehati-hatian untuk memastikan bahwa sistem tidak menggunakan data mahasiswa atau data lainnya untuk melatih model dan bahwa data tersebut tidak bocor dalam penggunaan hilir (*downstream uses*).

Tabel 4.7 Tingkat Kematangan Perangkat AWE untuk Fokus Keselamatan, Keamanan, dan Privasi

Alat AWE	Keterangan	Tingkat Kematangan
Pelatih Penulisan	Khan Academy menyediakan informasi mengenai pengelolaan informasi yang dapat mengidentifikasi individu menggunakan teknik enkripsi. Situs web Khan Academy juga membahas fitur-fitur keamanan seperti pendeteksian konten yang tidak pantas, berbahaya, atau tidak aman; pembatasan interaksi dengan alat AI; penyediaan pesan dalam produk mengenai keterbatasan AI; serta pemberian akses bagi orang tua dan pengajar untuk melihat seluruh riwayat percakapan. Khan Academy juga menyediakan kebijakan privasi yang terperinci di situs webnya. Meskipun tingkat akurasi umpan balik dari Writing Coach belum diketahui, kami berharap penelitian serupa dapat diterapkan pada alat tersebut.	Mapan
MI Write	MI Write menyediakan kebijakan privasi yang terperinci mengenai privasi, keamanan, dan penanganan data serta ketentuan layanan yang jelas. Penggunaan alat AI yang terbatas untuk mengotomatisasi penilaian yang dikaitkan dengan skor dan umpan balik tertulis serta penyorotan kesalahan tata bahasa tertentu, mengindikasikan dampak negatif yang minimal bagi mahasiswa, selain potensi pemberian skor atau umpan balik yang keliru. Serangkaian studi evaluasi hasil mengenai MI Write, yang mengkaji persepsi mahasiswa dan pengajar, implementasi, serta dampak jangka pendek dan jangka panjang, menunjukkan bahwa MI Write	Dewasa

	memiliki risiko keamanan yang sangat kecil atau bahkan tidak ada sama sekali.	
eRevisi + RF	eRevise+RF tidak menyediakan informasi berbasis web mengenai bagaimana perangkat lunak tersebut mengelola privasi data. Meskipun demikian, sebagian besar penelitian tampaknya telah ditinjau oleh Dewan Peninjau Institusional (*Institutional Review Board*), yang mengindikasikan adanya kepatuhan terhadap penggunaan data yang etis dan perlindungan terhadap subjek manusia. Penggunaan alat AI yang terbatas untuk otomatisasi penilaian yang mencakup pemberian skor, umpan balik tertulis, serta penyorotan kesalahan tata Bahasa mengisyaratkan dampak negatif yang minim bagi mahasiswa, selain potensi risiko pemberian skor atau umpan balik yang keliru. Serangkaian studi evaluasi hasil yang mengkaji persepsi mahasiswa dan pengajar, pelaksanaan program, serta dampak jangka pendek dan jangka panjang menunjukkan bahwa eRevise+RF memiliki risiko keamanan yang sangat kecil atau bahkan tidak ada sama sekali.	Didirikan
Cambi	Cambi memiliki sertifikasi ISO 27001/27018, yang menjamin bahwa seluruh data Cambi disimpan dengan aman dan privasi terlindungi sepenuhnya. Seluruh data demografis dan respons mahasiswa menjalani proses penghapusan identitas secara ketat sebelum digunakan untuk melatih mesin AI. Cambi menggunakan model bahasa besar (*large language models*) sumber terbuka yang di-hosting pada infrastruktur *Cambium Assessment Inc.* (CAI), sehingga memitigasi risiko penggunaan API pihak ketiga yang mungkin memanfaatkan data tersebut untuk tujuan lain. Cambi menggunakan antarmuka terstruktur yang dirancang untuk tujuan khusus guna memastikan umpan balik yang diberikan tepat sasaran. Demi menjaga konsistensi umpan balik, mekanisme dan bahasa yang digunakan telah ditentukan sebelumnya (menggunakan skrip) dan bukan dihasilkan secara langsung oleh AI generatif.	Dewasa

	Umpan balik yang ditujukan bagi mahasiswa disampaikan dengan nada positif dan bahasa yang mudah dipahami, serta mendorong mahasiswa untuk melakukan penyuntingan sendiri. Cambi juga menyajikan umpan balik dengan bahasa yang tidak mutlak (tidak menyiratkan kepastian mutlak) agar mahasiswa tidak berasumsi bahwa sistem tersebut selalu benar. Pendekatan ini mendukung mahasiswa dalam proses penyuntingan tanpa memberikan jawaban "yang benar" ataupun mengerjakan tugas tersebut untuk mereka. Mahasiswa dapat memilih untuk mengabaikan panduan tersebut jika mereka mau. Umpan balik juga tersedia dalam bentuk lisan menggunakan teknologi *text-to-speech* otomatis.	

Sebagaimana disebutkan sebelumnya, penggunaan AI jika tidak divalidasi dengan tepat dapat berdampak pada keselamatan mahasiswa. Sistem chatbot dapat diretas (*jailbroken*) agar berperilaku menyimpang dan berpotensi membahayakan; sistem ini juga dapat dengan meyakinkan mengarahkan mahasiswa ke jalan yang salah, mengerjakan tugas mahasiswa sehingga menciptakan ketergantungan berlebih yang berujung pada hasil pembelajaran yang lebih buruk, atau sekadar memberikan informasi yang keliru. Evaluasi terhadap dampak buruk semacam itu sekali lagi mencerminkan teori tindakan yang kuat, yang "mengantisipasi bagaimana penerapan program yang tidak tepat dapat menyebabkan konsekuensi negatif yang tidak disengaja" (NCME).

Perangkat AI seharusnya "tidak menyebabkan kondisi yang membahayakan nyawa manusia, kesehatan, properti, atau lingkungan dalam situasi tertentu yang telah ditetapkan". Perlu dicatat bahwa dampak merugikan dapat terjadi pada manusia, organisasi, maupun ekosistem. Pedoman OET menguraikan contoh potensi dampak buruk bagi mahasiswa, termasuk umpan balik yang bias terhadap mahasiswa tertentu atau sekadar keliru, serta kecenderungan orang untuk terlalu mengandalkan AI sehingga melimpahkan proses berpikir kepada sistem AI tersebut. Oleh karena itu, kerangka kerja evaluasi AWE harus mempertimbangkan potensi dampak buruk ini dalam pengukuran dan evaluasi efektivitasnya. Keamanan organisasi, seperti sekolah, juga dapat terdampak oleh alat AWE jika organisasi tersebut terlalu bergantung pada keluaran sistem tanpa masukan dan pengawasan guru, mencoba menggantikan peran guru dengan alat tersebut, atau berupaya mengurangi sifat kolaboratif dalam pembelajaran di kelas. Keamanan ekosistem, seperti sistem pendidikan dan komunitas, dapat terancam jika penggunaan alat AWE secara luas berdampak negatif terhadap pembelajaran mahasiswa secara keseluruhan. Alat AWE yang sudah matang harus mempertimbangkan dan mengevaluasi potensi dampak negatif dari penggunaannya terhadap mahasiswa, guru, ruang kelas, dan sekolah.

Kerangka kerja NIST mendefinisikan privasi sebagai "norma dan praktik yang membantu melindungi otonomi, identitas, dan martabat manusia". Dalam konteks alat AWE, hal ini berarti memastikan bahwa data mahasiswa dan guru tetap terjaga kerahasiaannya selama interaksi dengan alat tersebut. Data dalam hal ini mencakup informasi pengenal apa pun, respons mahasiswa, dan umpan balik. Bagi sistem yang menggunakan AI sumber tertutup (*closed-source*), hal ini berarti memastikan bahwa data yang dikirimkan ke dan diterima dari alat tersebut juga tetap menjaga privasi tersebut. Ada tiga contoh jaminan terkait hal ini: data yang dikirim ke versi yang digunakan pengguna tidak digunakan untuk melatih model yang lebih umum dengan kegunaan lain; data yang dikirim ke versi pengguna tidak dapat dihasilkan oleh model lebih umum yang tersedia bagi pihak lain di luar alat AWE tersebut; serta keluaran yang dihasilkan oleh alat tersebut mematuhi aturan privasi data yang sama.

Tabel 4.7 menyajikan peringkat tingkat kematangan untuk keempat perangkat AWE. MI Write dan Cambi dikategorikan pada tingkat *Mature* (Matang) berdasarkan kebijakan organisasi terkait penanganan data dan perangkat lunak, penggunaan AI yang menghasilkan prediksi skor atau umpan balik yang terbatas, serta desain sistem atau penelitian yang mendukung penggunaan AI secara aman. *Writing Coach* dikategorikan pada tingkat *Established* (Mapan) berdasarkan pernyataan kebijakan privasi Khan Academy dan penelitian yang mendasari *chatbot* matematika Khanmigo. eRevise+RF dikategorikan pada tingkat *Established* (Mapan) berdasarkan penggunaan tinjauan *Institutional Research Board* (Dewan Peninjau Penelitian Institusional), penggunaan AI yang menghasilkan prediksi skor atau umpan balik yang terbatas, serta penelitian yang mendasarinya.

Tabel 4.7 Tingkat Kematangan Alat AWE untuk Fokus Keselamatan, Keamanan, dan Privasi

Alat AWE	Deskripsi	Tingkat Kematangan
Writing Coach	Khan Academy menyediakan informasi mengenai pengelolaan informasi yang dapat mengidentifikasi individu (*personally identifiable information*) dengan menggunakan teknik enkripsi. Situs web Khan Academy juga membahas fitur keselamatan, seperti mendeteksi konten yang tidak pantas, berbahaya, atau tidak aman; membatasi interaksi dengan alat AI; menyediakan pesan dalam produk mengenai keterbatasan AI; serta memungkinkan orang tua dan guru untuk melihat seluruh riwayat percakapan. Khan Academy juga menyediakan kebijakan privasi yang terperinci di situs webnya. Meskipun tingkat akurasi umpan balik dari *Writing Coach* belum diketahui, kami memperkirakan bahwa penelitian serupa telah diterapkan pada alat tersebut.	***Established***
MI Write	MI Write menyediakan kebijakan privasi yang terperinci mengenai privasi data, keamanan, dan penanganan data (MI Write, t.t.-a; MI Write, t.t.-b), serta ketentuan layanan yang telah ditetapkan (MI Write, t.t.-c). Penggunaan alat AI yang terbatas untuk mengotomatisasi penilaian, yang dikaitkan dengan skor dan umpan balik teks, serta penandaan beberapa kesalahan tata bahasa, menunjukkan dampak	***Mature***

	negatif yang minimal terhadap siswa selain kemungkinan pemberian skor atau umpan balik yang tidak benar. Serangkaian studi evaluasi hasil pada MI Write yang mengkaji persepsi siswa dan guru, implementasi, serta dampak jangka pendek dan jangka panjang menunjukkan bahwa MI Write menimbulkan sedikit atau tidak ada risiko keselamatan.	
eRevise + RF	eRevise+RF tidak menyediakan informasi berbasis web mengenai bagaimana perangkat lunak tersebut mengelola privasi data. Meskipun demikian, sebagian besar penelitian tampaknya ditinjau oleh *Institutional Review Board*, yang menunjukkan kepatuhan terhadap penggunaan data secara etis dan perlindungan subjek manusia. Penggunaan alat AI yang terbatas untuk mengotomatisasi penilaian yang dikaitkan dengan skor dan umpan balik teks, serta penandaan beberapa kesalahan tata bahasa, menunjukkan dampak negatif yang terbatas terhadap mahasiswa selain kemungkinan pemberian skor atau umpan balik yang tidak benar. Serangkaian studi evaluasi hasil yang mengkaji persepsi mahasiswa dan dosen, implementasi, serta dampak jangka pendek dan jangka panjang menunjukkan bahwa eRevise+RF menimbulkan sedikit atau tidak ada risiko keselamatan.	***Established***
Cambi	Cambi tersertifikasi ISO 27001/27018 (Cambium Assessment, t.t.), yang memastikan bahwa seluruh data Cambi disimpan dengan aman dan privasi sepenuhnya terlindungi. Seluruh data demografis dan respons siswa dihapus identitasnya (*de-identified*) secara ketat sebelum digunakan dalam pelatihan mesin AI. Cambi menggunakan model bahasa besar (*large language models*) sumber terbuka yang dihosting pada infrastruktur *Cambium Assessment Inc.* (CAI), sehingga mengurangi risiko penggunaan API pihak ketiga yang mungkin menggunakan data tersebut untuk tujuan lain. Cambi menggunakan antarmuka terstruktur yang dirancang untuk tujuan tertentu guna memastikan bahwa umpan balik yang diberikan sesuai. Untuk memastikan umpan balik selalu konsisten, mekanisme dan bahasa umpan balik dibuat dalam bentuk skrip, bukan dihasilkan oleh AI generatif. Umpan balik yang ditujukan kepada siswa disampaikan dengan bahasa yang positif dan mudah dipahami serta menempatkan tanggung jawab kepada siswa untuk melakukan penyuntingan. Cambi juga memberikan penyangga (*hedges*) pada umpan balik sehingga siswa tidak menganggap bahwa mesin selalu benar. Pendekatan ini mendukung mahasiswa dalam proses penyuntingan tanpa memberikan jawaban yang “benar” atau mengerjakan tugas untuk mereka. Mahasiswa dapat memilih untuk mengabaikan panduan tersebut jika mereka menghendakinya. Umpan balik juga tersedia melalui suara dengan menggunakan teknologi *text-to-speech* otomatis.	***Mature***

Kami memperkenalkan matriks kematangan AWE untuk menyediakan kerangka kerja terstruktur bagi para pendidik dan pembuat kebijakan dalam mengambil keputusan berbasis

bukti mengenai alat mana yang sebaiknya digunakan untuk tujuan tertentu, sekaligus mendorong pengembang alat AWE untuk mempertimbangkan aspek-aspek kunci yang dapat memengaruhi desain, pengembangan, implementasi, dan penelitian terkait alat-alat tersebut. Kami menguraikan setiap alat AWE untuk masing-masing dari lima area fokus dan mengusulkan tingkat kematangan bagi setiap alat serta area fokusnya. Tujuan evaluasi ini adalah untuk mengilustrasikan penerapan rubrik tersebut menggunakan informasi yang tersedia untuk umum mengenai alat-alat AWE. Tabel 4.8 merangkum peringkat yang kami usulkan untuk setiap alat AWE di kelima area fokus tersebut. Perlu dicatat bahwa tingkat peringkat alat AWE bervariasi di berbagai area fokus. Alat-alat yang sudah lama ada (misalnya, MI Write, erevise+RF) cenderung memiliki peringkat kematangan yang lebih tinggi (misalnya, *Established* atau *Mature*). Alat AWE yang baru dirilis memiliki tingkat yang bervariasi bergantung pada kematangan proses pengembangan dan evaluasinya. Sebagai contoh, Cambi menjalani proses validasi yang substansial sebelum dirilis sehingga memiliki tingkat kematangan yang cukup baik untuk beberapa area fokus, namun hanya memiliki sedikit penelitian yang mendukung evaluasi hasilnya. *Writing Coach* baru saja dirilis dan belum memiliki artikel yang dipublikasikan mengenai desain serta pengembangannya; namun, alat ini mencakup teori tindakan (*theory of action*) dan dibangun di atas fondasi teknis Khan Academy.

Sebuah metafora yang banyak digunakan dan telah mendapat perhatian luas di luar bidang pendidikan dan penilaian adalah "Label Nutrisi AI" (*AI Nutrition Label*). Label ini berfungsi sebagai heuristik yang mudah dipahami sebuah ajakan untuk mengungkapkan apa yang ada "di dalam" alat AI, termasuk asal-usul data, arsitektur model, keterbatasan yang diketahui, dan potensi bias, yang disajikan secara terstandar. Meskipun berbagai organisasi menafsirkan dan menerapkan konsep ini dengan cara yang berbeda, tujuan dasarnya tetap konsisten: mendorong keterbukaan informasi agar konsumen dapat lebih memahami aspek apa yang bermanfaat maupun yang merugikan.

Namun, dalam konteks penilaian pendidikan, alat AI pada dasarnya memiliki kompleksitas yang lebih tinggi. Alat-alat ini tidak hanya membuat inferensi mengenai karakteristik laten yang tidak dapat diamati secara langsung, tetapi juga memberikan umpan balik yang bertujuan untuk mendukung pengembangan karakteristik tersebut pada diri Mahasiswa. Akibatnya, bukti yang diperlukan untuk mengevaluasi penggunaan perangkat semacam itu tidak dapat dikategorikan secara sederhana ke dalam satu label atau dokumen yang seragam. Meskipun demikian, semangat di balik *AI Nutrition Label* yaitu komitmen terhadap transparansi, kejelasan, dan evaluasi yang berlandaskan informasi memadai tetap sangat relevan. Alih-alih mengusulkan sebuah label, kami menawarkan sebuah matriks yang membantu pengembang perangkat mengidentifikasi informasi penting untuk diungkapkan, serta memandu evaluator dalam menentukan ekspektasi yang wajar saat menilai manfaat perangkat tersebut bagi mahasiswa dan pendidik.

Tabel 4.8 Tingkat Matriks Kematangan AWE untuk Keempat Perangkat AWE

Tingkat Kematangan	**Teori Aksi**	**Evaluasi Hasil**	**Teknologi AI**	**Validasi Model**	**Keselamatan, Keamanan, dan Privasi**
Pelatih Menulis	Sedang Dikembangkan	Tahap Awal	Tahap Awal	Tahap Awal	Mapan
MI Write	Sudah Mapan	Tahap Matang	Sudah Mapan	Sudah Mapan	Matang
eRevise+RF	Matang	Tahap Mapan	Matang	Sudah Mapan	Mapan
Cambi	Sudah Mapan	Tahap Pengembangan	Matang	Sudah Mapan	Matang

BAB 5
PERSONALISASI PENILAIAN UNTUK KINERJA DAN KETERLIBATAN

5.1 URGENSI PERSONALISASI DALAM PENILAIAN

Dalam pengukuran pendidikan, standardisasi umumnya memiliki dua makna. Makna pertama berkaitan dengan pembuatan norma untuk acuan skor, seperti dalam "sampel standardisasi tes". Makna kedua yang menjadi fokus pembahasan di sini adalah pemberian perlakuan yang setara kepada peserta tes. Dalam bentuknya yang ketat, jenis standardisasi ini memberikan konten tes yang pada dasarnya sama kepada setiap peserta dalam format dan kondisi yang sama, serta menilai jawaban mereka menggunakan aturan yang sama. Makna standardisasi ini dibawa ke bidang pengukuran pendidikan yang sedang berkembang pada awal abad ke-20 dari ilmu fisika melalui bidang psikologi eksperimental yang saat itu juga masih baru. Dalam konteks eksperimental, gagasannya sederhana: memanipulasi satu variabel sembari menjaga faktor-faktor lain tetap konstan untuk mengidentifikasi pengaruh variabel yang dimanipulasi tersebut terhadap suatu hasil. Dalam pengujian, standardisasi bertujuan untuk mengisolasi variabel yang diukur dengan menjaga semua kondisi lain tetap konstan. Tentu saja, hal yang tidak dapat dilakukan oleh standardisasi adalah menjaga karakteristik peserta tes tetap konstan. Praktik ini bekerja sangat efektif ketika populasi peserta tes relatif homogen; artinya, ketika karakteristik yang tidak relevan dengan konstruk (namun mungkin berkorelasi dengan variabel yang diukur) sebenarnya tidak banyak bervariasi (misalnya, semua peserta tes berasal dari kelas sosial yang serupa). Dalam kasus ideal, ketika semua peserta tes memiliki karakteristik yang persis sama dalam segala hal kecuali pada aspek yang sedang diukur, satu-satunya faktor yang secara signifikan memengaruhi skor tes adalah variabel target tersebut.

Ironisnya, salah satu orang pertama yang menyoroti perlunya homogenitas relatif adalah Carl Brigham. Dalam bukunya, *A Study of American Intelligence* (1923), ia membandingkan kinerja berbagai kelompok ras/etnis dalam tes kecerdasan Army Alpha dan Beta. Karena direkrut untuk dinas militer selama Perang Dunia I, populasi yang diuji sangat heterogen dalam berbagai dimensi. Dengan menggunakan bahasa yang jelas-jelas rasis, Brigham seorang penganut eugenika menafsirkan perbedaan kelompok ras/etnis yang teramati tersebut sebagai tanda akan terjadinya penurunan kecerdasan populasi nasional secara progresif di masa depan. Tujuh tahun kemudian, ia menerbitkan sebuah artikel di *Psychological Review* yang diakhiri dengan pernyataan pencabutan pendapat dalam satu kalimat. Sebelum pernyataan pencabutan tersebut, ia menulis,

> Untuk tujuan membandingkan individu atau kelompok, jelaslah bahwa tes yang menggunakan bahasa sehari-hari setempat (*vernacular*) hanya boleh digunakan pada individu yang memiliki kesempatan setara untuk menguasai bahasa tersebut. Persyaratan ini menutup kemungkinan penggunaan tes semacam itu dalam studi perbandingan yang

melibatkan individu yang dibesarkan di lingkungan rumah di mana bahasa tes tersebut tidak digunakan, atau di mana dua bahasa digunakan.

Penggunaan tes standar yang memiliki dampak signifikan setelah Perang Dunia I sebagian besar terbatas pada penerimaan mahasiswa di perguruan tinggi dan universitas; pada saat itu, populasi peserta tes cenderung sangat homogen, terutama di institusi-institusi selektif yang didominasi oleh pelamar laki-laki kulit putih dari kelas atas. Seiring berjalannya waktu, populasi yang menjalani tes-tes berdampak signifikan tersebut menjadi lebih beragam. Menjelang Perang Dunia II, Harvard mulai menggunakan SAT untuk mengevaluasi pelamar beasiswa sebagai upaya menarik mahasiswa baru dengan latar belakang ekonomi yang lebih beragam. Setelah Perang Dunia II, kebijakan GI Bill mendanai pendidikan tinggi bagi kelompok pelamar dari kelas sosial yang lebih luas, yaitu para tentara yang telah menyelesaikan masa tugasnya. Namun, baik dalam kasus Harvard maupun GI Bill, populasi peserta tes pada dasarnya tetap didominasi oleh laki-laki kulit putih.

Populasi AS sendiri relatif homogen dari segi ras/etnis dan bahasa hingga masa yang belum lama berselang. Pada tahun 1980, komposisi penduduk mencakup 20% bukan kulit putih, 6,2% kelahiran luar negeri, dan 11% yang menggunakan bahasa selain bahasa Inggris di rumah. Pada tahun 2020 hanya empat dekade kemudian angka-angka tersebut telah meningkat dua kali lipat: 42% bukan kulit putih, 13,7% kelahiran luar negeri, dan 19% menggunakan bahasa selain bahasa Inggris di rumah. Per tahun 2022, populasi mahasiswa sekolah negeri sebagian besar terdiri dari kelompok bukan kulit putih dan mahasiswa yang memenuhi syarat untuk mendapatkan makan siang gratis atau dengan harga subsidi (*National Center for Education Statistics* [NCES], 2023, Tabel 203.70 dan 204.10), yang merupakan indikator kelas ekonomi. Kondisi ini disebut oleh Vertovec sebagai "superdiversitas" (*superdiversity*), yang merujuk pada tingkat keberagaman yang belum pernah terjadi sebelumnya. Selain keragaman latar belakang budaya dan bahasa, peserta ujian juga berasal dari konteks sosial-budaya serta pengalaman pendidikan yang beragam. Konteks-konteks ini membentuk cara individu belajar, cara mereka memecahkan masalah, serta tingkat motivasi dan keterlibatan yang mereka bawa saat mengikuti asesmen.

Meningkatnya keragaman populasi serta konteks sosial-budaya dan pendidikan secara signifikan baik di Amerika Serikat maupun di tempat lain menunjukkan perlunya pendekatan asesmen yang memperhitungkan perbedaan dalam hal bahasa, cara memperoleh pengetahuan, identitas budaya, minat, pengetahuan awal, dan faktor sosial-budaya lainnya. Faktor-faktor ini sebelumnya sering dianggap setara atau tidak relevan dalam konteks tes standar. Faktanya, langkah-langkah yang menyimpang dari standardisasi sebagai respons terhadap keragaman populasi telah mulai diterapkan secara bertahap selama beberapa waktu. Langkah-langkah tersebut mencakup tes adaptif untuk mengakomodasi perbedaan tingkat kompetensi dengan lebih baik, penyediaan akomodasi bagi individu dengan kebutuhan khusus, penyediaan versi bahasa bagi pembelajar bahasa Inggris, pemberian pilihan soal untuk mengakomodasi variasi minat dan pengetahuan awal, serta pelaksanaan tes di rumah bagi mereka yang merasa lingkungan di luar pusat tes lebih kondusif. Semua langkah ini bertujuan

untuk menghasilkan pengukuran yang lebih baik dan lebih adil dengan memberikan kesempatan kepada individu untuk menunjukkan pengetahuan dan kemampuan mereka melalui cara-cara yang sesuai bagi mereka. Selain itu, melalui berbagai dasar pertimbangan dan terkadang melalui teknik penghubungan statistik, skor yang dihasilkan dapat dianggap sebanding dengan skor dari pelaksanaan tes standar.

5.2 BENTUK DAN JENIS PERSONALISASI

Bentuk-Bentuk Personalisasi

Semua pendekatan yang disebutkan di atas, pada dasarnya, merupakan bentuk personalisasi mikro. Sesuai dengan namanya, personalisasi mikro memiliki cakupan yang cukup terbatas. Sebagai contoh, tes adaptif pada umumnya hanya melakukan personalisasi berdasarkan satu variable yaitu kompetensi yang sedang diukur dengan memilih butir soal (atau rangkaian soal) berikutnya berdasarkan estimasi kompetensi yang diperbarui secara dinamis. Penerapan personalisasi dalam penilaian yang lebih menyeluruh jauh lebih kompleks; dan serupa dengan pembelajaran yang dipersonalisasi, hal ini dapat dicapai melalui berbagai cara.

Bentuk personalisasi yang paling sederhana adalah merancang penilaian sedemikian rupa sehingga individu dapat menyesuaikan aspek-aspek tertentu dari penilaian tersebut dengan minat, pengetahuan awal, identitas budaya, dan karakteristik relevan lainnya yang mereka miliki. Bentuk personalisasi atau adaptivitas yang digerakkan oleh manusia ini melibatkan peserta didik secara aktif dalam menentukan pilihan terkait konten dan pengalaman penilaian, yang dapat mendorong agensi, motivasi intrinsik, dan keterlibatan. Sebagai contoh, peserta tes dapat memilih tingkat kesulitan butir soal (misalnya, tes adaptif mandiri) atau format respons. Perlu dicatat bahwa personalisasi yang digerakkan oleh manusia ini selaras dengan teori determinasi diri (*self-determination theory*), yang menekankan agensi sebagai kebutuhan dasar manusia dan faktor utama bagi motivasi intrinsik. Agensi merupakan konstruk multidimensi yang lebih dari sekadar penyediaan pilihan dan dianggap krusial bagi pembelajaran yang diatur secara mandiri (*self-regulated learning*). Meskipun penyediaan pilihan dalam penilaian dapat meningkatkan motivasi intrinsik dan keterlibatan peserta tes, penerapan pilihan tersebut secara efektif memerlukan pertimbangan yang cermat. Penelitian menunjukkan bahwa terlalu banyak pilihan justru dapat menurunkan keterlibatan, sehingga mendorong sejumlah ahli untuk menyarankan pembatasan jumlah pilihan menjadi antara dua hingga empat opsi. Salah satu tantangan utama dalam penerapan pilihan pada asesmen berskala besar adalah kompleksitas psikometris untuk memastikan kesetaraan (*equivalence*) di antara berbagai opsi tugas sebuah isu yang akan dibahas kembali pada bagian "Validasi dan Isu Terkait untuk Asesmen yang Dipersonalisasi". Asesmen berbasis penelitian dalam program A*dvanced Placement* (AP) merupakan contoh personalisasi yang digerakkan oleh manusia.

Tugas kinerja yang berlangsung sepanjang kursus ini mencakup penulisan makalah akademik mengenai topik pilihan mahasiswa, presentasi, dan sesi pertahanan lisan (*oral defense*); seluruh komponen tersebut berkontribusi terhadap nilai akhir mahasiswa. Personalisasi di sini terletak pada kebebasan yang diberikan kepada peserta ujian untuk

menentukan topik, metode, serta aspek-aspek lain terkait substansi dan pelaksanaan proyek, yang memungkinkan mereka menampilkan karakteristik diri melalui cara-cara yang dianggap paling efektif dalam menunjukkan kemampuan mereka. Salah satu kompleksitas yang menyertai pendekatan ini terletak pada proses penilaian, mengingat setiap peserta ujian menghasilkan kinerja yang unik.

Bentuk personalisasi lainnya melibatkan pendekatan yang digerakkan oleh mesin. Alih-alih peserta ujian yang melakukan penyesuaian agar asesmen sesuai dengan diri mereka, mesinlah yang menggunakan informasi mengenai peserta ujian untuk merancang asesmen yang unik bagi mereka. Dengan kata lain, berdasarkan vektor nilai yang mendeskripsikan karakteristik peserta ujian, diciptakanlah asesmen yang "optimal" bagi mereka di mana definisi tepat mengenai optimalitas itu sendiri merupakan salah satu aspek yang kompleks. Pendekatan ini memanfaatkan model kognitif komputasional atau metode kecerdasan buatan (AI) termasuk AI generative untuk mempersonalisasi penilaian secara *real-time*. Idealnya, proses ini didasarkan pada model pembelajar (yang dapat ditelaah atau terintegrasi) yang menyajikan pemahaman transparan dan kohesif mengenai kemajuan, kekuatan, serta kebutuhan pembelajar di berbagai dimensi, sekaligus mendukung intervensi yang lebih personal, tepat waktu, dan efektif tanpa mengorbankan kepercayaan terhadap sistem. Meskipun personalisasi berbasis mesin memiliki berbagai keunggulan, sistem yang sepenuhnya otomatis berisiko tidak selaras dengan kebutuhan pembelajar terutama bagi populasi neurodivergen yang perilakunya mungkin menyimpang dari model normatif. Model pembelajar juga mungkin gagal menangkap nuansa kondisi kognitif, afektif, atau kontekstual, sehingga menghasilkan adaptasi yang tidak efektif. Untuk memitigasi risiko terkait ketidakpastian dalam adaptasi berbasis mesin tersebut, pendekatan yang melibatkan manusia dalam prosesnya (*human-in-the-loop*) menjadi sangat penting. Mewujudkan jenis penilaian personal seperti ini terutama dalam konteks penilaian yang berdampak tinggi merupakan tantangan besar; wajar jika belum ada contoh operasional yang diterapkan dalam skala luas. Namun, berkat kemajuan pesat dalam teknologi AI selama beberapa tahun terakhir, contoh-contoh penerapannya mungkin akan segera muncul.

Pendekatan terakhir untuk personalisasi adalah menggabungkan kedua metode tersebut secara hibrida. Alih-alih memandang pemberian kendali antara manusia dan mesin secara satu dimensi, kerangka kerja AI yang berpusat pada manusia seperti model dua dimensi yang diusulkan oleh Shneiderman menekankan pentingnya menyeimbangkan otomatisasi tingkat tinggi yang digerakkan oleh mesin dengan kendali manusia yang kuat demi menjamin transparansi, keamanan, dan kepuasan pengguna. Dalam konteks penilaian pendidikan, sistem rekomendasi mencerminkan keseimbangan ini dengan menggunakan algoritma AI untuk menyarankan satu atau beberapa pilihan, sembari tetap memberikan keleluasaan bagi peserta didik untuk mengabaikan atau memilih dari rekomendasi yang disediakan. AI generatif (genAI) dapat semakin mendukung model personalisasi hibrida berkat kemampuannya menghasilkan konten secara langsung (*on-the-fly*); dalam hal ini, peserta didik memberikan masukan kepada model genAI, dan model tersebut mempersonalisasi konten secara waktu nyata (misalnya, Arslan). Selain itu, *chatbot* berbasis genAI dapat membantu peserta didik dengan menjawab

pertanyaan terkait kegunaan sistem atau memberikan dukungan Bahasa. Namun, perancang penilaian harus berhati-hati saat menerapkan personalisasi *on-the-fly* menggunakan genAI, khususnya dalam konteks penilaian yang memiliki konsekuensi tinggi (*high-stakes assessment*). Sangatlah penting untuk memastikan bahwa prinsip-prinsip utama penilaian pendidikan (yaitu validitas, reliabilitas, dan keadilan) tetap dijunjung tinggi serta alat genAI digunakan secara bertanggung jawab dan etis (Arslan; lihat juga Johnson, mengenai prinsip dan praktik AI yang bertanggung jawab untuk pengukuran dan pembelajaran).

Karakteristik Mahasiswa Apa yang Menjadi Dasar Personalisasi?

Selain berbagai pendekatan terhadap personalisasi, penting juga untuk mempertimbangkan karakteristik tingkat mahasiswa atau perbedaan individu yang menjadi dasar personalisasi asesmen. Faktor-faktor yang disasar ini dapat membentuk profil mahasiswa yang mencakup informasi mengenai karakteristik mahasiswa (misalnya, pengetahuan, minat, serta latar belakang linguistik dan budaya) dan faktor kontekstual seperti kesempatan belajar mereka sebelumnya. Profil mahasiswa ini dapat digunakan untuk menyesuaikan asesmen guna memberikan pengalaman yang lebih personal, yang selaras dengan karakteristik yang dibawa mahasiswa saat mengikuti asesmen. Dalam pembahasan ini, kami berfokus pada karakteristik mahasiswa berikut sebagai pendorong personalisasi, mengingat relevansinya yang krusial terhadap pembelajaran dan asesmen: kinerja sebelumnya, pengetahuan awal, upaya kognitif dan minat, serta latar belakang bahasa dan budaya.

- **Kinerja sebelumnya**. Secara tradisional, pendekatan adaptivitas terutama mempertimbangkan kinerja mahasiswa sebelumnya sebagai dasar personalisasi. Dalam konteks pembelajaran yang dipersonalisasi, teknologi pembelajaran adaptif biasanya menyesuaikan tingkat atau jenis konten agar selaras dengan tingkat kompetensi mahasiswa saat ini berdasarkan kinerja mereka pada tugas-tugas terkait sebelumnya guna membantu mahasiswa berkembang menuju penguasaan keterampilan. Dalam asesmen pendidikan, contoh terbaik dari personalisasi berbasis kinerja mahasiswa adalah *computerized adaptive testing*. Pada CAT pada umumnya, tingkat kompetensi peserta tes yang diestimasi dari kinerja pada butir soal sebelumnya digunakan untuk menentukan butir soal apa yang akan disajikan berikutnya; penentuan ini biasanya didasarkan pada tingkat kesulitan yang tepat (yaitu tidak terlalu mudah maupun terlalu sulit) sesuai dengan estimasi kemampuan peserta tes saat itu. Pendekatan semacam itu umum diterapkan dalam program tes standar yang bersifat operasional. Namun, mengandalkan faktor kognitif seperti kinerja sebelumnya saja mungkin tidak memadai sebagai sarana personalisasi yang lebih mendalam dalam penilaian pendidikan yang mencerminkan konsepsi lebih holistik mengenai pengalaman mahasiswa.
- **Pengetahuan awal**. Penilaian yang dipersonalisasi harus didasarkan pada pemahaman mengenai pengetahuan awal setiap mahasiswa. Pengetahuan spesifik-domain dapat memberikan dampak yang memfasilitasi (misalnya, pengelompokan informasi/ *chunking*, penggunaan strategi, evaluasi) maupun menghambat (misalnya,

miskonsepsi, bias) proses pembelajaran dan kinerja. Mahasiswa memasuki lingkungan pendidikan dengan pengalaman, pengetahuan, dan tingkat keterampilan yang beragam, yang memengaruhi cara mereka memproses informasi selama penilaian berlangsung. Ketika dihadapkan pada tugas penilaian yang baru, untuk memberikan respons, mahasiswa perlu memanggil kembali pengetahuan awal dari memori jangka panjang mereka dan menggunakannya bersama keterampilan prosedural serta strategi yang telah mereka pelajari sebelumnya. Oleh karena itu, lingkungan penilaian yang dipersonalisasi harus melakukan adaptasi dengan mempertimbangkan baik pengetahuan deklaratif (faktual) maupun keterampilan prosedural guna memandu pemilihan konten, pemberian umpan balik, dan penjadwalan. Pengembang perlu mempertimbangkan aspek kuantitas pengetahuan (tinggi vs. rendah) maupun kualitas pengetahuan (spesifik-domain vs. umum, akurat vs. tidak akurat) dalam menentukan cara mempersonalisasi tugas penilaian.

- **Upaya kognitif dan minat**. Motivasi dan keterlibatan mahasiswa sangat penting untuk menilai secara lebih baik apa yang diketahui dan dapat dilakukan oleh mahasiswa. Ketika mahasiswa tidak mengerahkan upaya kognitif secara penuh saat merespons tugas penilaian, validitas interpretasi skor penilaian dapat terancam. Dalam konteks penilaian yang tidak memiliki konsekuensi besar (*low-stakes*), peserta tes mungkin tidak selalu terlibat secara kognitif dalam tugas penilaian karena tidak adanya konsekuensi langsung bagi mereka. Dalam konteks ini, optimalisasi keterlibatan menjadi hal yang penting untuk mendukung interpretasi hasil penilaian yang valid. Salah satu cara untuk mengoptimalkan keterlibatan adalah dengan mempersonalisasi konteks tugas penilaian berdasarkan minat mahasiswa. Sebagai contoh, mahasiswa dapat memasukkan minat mereka di luar sekolah, dan AI generatif (*genAI*) dapat secara dinamis menyesuaikan konteks soal cerita matematika secara *real-time*. Cara lain untuk mengoptimalkan keterlibatan adalah dengan mendeteksi respons yang menunjukkan kurangnya keterlibatan atau kurangnya upaya maksimal, lalu memberikan peringatan atau dorongan (*nudges*) kepada mahasiswa selama penilaian berlangsung, atau memilih pertanyaan maupun konten penilaian dengan mempertimbangkan keterlibatan mahasiswa
- **Latar belakang bahasa**. Personalisasi berdasarkan latar belakang bahasa mahasiswa merupakan praktik operasional yang umum diterapkan, khususnya dalam konteks penilaian internasional berskala besar seperti *Programme for International Student Assessment* (PISA), *Program for the International Assessment of Adult Competencies*, *Trends in International Mathematics and Science Study*, dan *Progress in International Reading Literacy Study*. Penyajian format tes dalam berbagai bahasa yang disesuaikan dengan populasi sampel memungkinkan dilakukannya perbandingan lintas negara terkait pengetahuan, keterampilan, dan kemampuan yang menjadi sasaran penilaian. Dalam kasus-kasus ini, desain penilaian dikontrol secara ketat untuk memastikan bahwa butir-butir soal tetap dapat diperbandingkan, relevan secara lokal, dan mudah dipahami selama proses penerjemahan tes ke dalam berbagai bahasa tersebut. Analisis

statistik yang cermat juga dilakukan setelah pelaksanaan tes untuk memfasilitasi perbandingan antar kelompok.

- **Latar belakang budaya**. Dalam beberapa dekade terakhir, minat terhadap personalisasi penilaian pendidikan yang disesuaikan dengan latar belakang budaya dan identitas mahasiswa kembali menguat, mengarah pada pendekatan penilaian yang relevan, responsif, dan mendukung budaya (atau sosiokultural) mahasiswa. Secara kolektif, pendekatan-pendekatan ini bertujuan untuk "merancang penilaian yang mempertimbangkan karakteristik sosial, budaya, dan karakteristik relevan lainnya dari mahasiswa yang beragam serta konteks tempat mereka tinggal". Tujuannya adalah menciptakan penilaian yang selaras dengan pengetahuan, pengalaman, minat, dan nilai-nilai mahasiswa dari beragam latar belakang mahasiswa yang mungkin tidak terakomodasi dengan baik atau tidak merasa terwakili dalam penilaian tradisional yang mencerminkan budaya mayoritas yang dominan. Dengan demikian, merancang asesmen yang responsif secara sosiokultural menuntut pertimbangan yang cermat terhadap beragam karakteristik mahasiswa termasuk pengetahuan awal, minat, bahasa, dan identitas serta memasukkan konten dan konteks asesmen yang menghargai karakteristik budaya tersebut sebagai aset dan meresponsnya dengan cara yang secara bermakna mencerminkan keterkaitan di antara aspek-aspek tersebut.

Menerapkan Personalisasi dalam Desain Butir Soal, Tugas, dan Tes

Setelah pengembang menentukan karakteristik dan perilaku yang akan menjadi dasar personalisasi, mereka harus mengambil keputusan desain mengenai cara menerapkan personalisasi sebagai respons terhadap kualitas-kualitas tersebut. Di sini, kami membahas secara singkat beberapa kerangka kerja dan pendekatan yang relevan untuk mendukung desain personalisasi dalam asesmen.

Asesmen yang peduli (*caring assessment*). Penerapan personalisasi dalam asesmen mengharuskan perancangan dan penerapan adaptasi pada konten serta konteks asesmen guna memenuhi kebutuhan individu mahasiswa dan merespons karakteristik serta perilaku mereka. Personalisasi ini dapat dilakukan pada tingkat butir soal melalui penyesuaian konten, tingkat kesulitan, atau urutan ataupun pada tingkat asesmen secara keseluruhan, melalui dukungan seperti akomodasi bahasa, dorongan (*nudges*), atau umpan balik. Kerangka kerja untuk asesmen yang peduli menetapkan bahwa adaptasi, dukungan, dan umpan balik yang dipersonalisasi dapat diterapkan pada tiga tahapan waktu: sebelum, selama, dan setelah mahasiswa mengikuti asesmen. Adaptasi sebelum asesmen dapat memanfaatkan informasi yang termuat dalam profil mahasiswa atau model pembelajar terintegrasi untuk menyajikan versi asesmen atau tugas yang berbeda kepada peserta ujian, berdasarkan karakteristik mahasiswa yang telah diketahui. Pemberian versi asesmen dalam bahasa yang berbeda berdasarkan bahasa ibu mahasiswa atau bahasa pengantar pembelajaran merupakan contoh adaptasi semacam ini. Sebagai contoh lain mahasiswa dengan pengetahuan awal yang terbatas mengenai suatu topik dapat diarahkan ke bentuk asesmen alternatif dengan topik yang lebih familier, atau ke versi yang menyertakan materi latar belakang tambahan yang memungkinkan mereka untuk memahami topik utama yang dituju.

Dukungan yang dipersonalisasi dapat diterapkan secara dinamis selama asesmen dengan memantau tindakan dan interaksi mahasiswa dalam tugas (misalnya, respons terhadap butir soal, keterlibatan perilaku) serta memberikan intervensi yang tepat waktu (misalnya, dorongan halus atau pesan) sebagai tanggapan atas kondisi yang terdeteksi. Umpan balik yang dipersonalisasi dapat diberikan setelah asesmen melalui penyesuaian laporan skor yang mengontekstualisasikan kinerja mahasiswa berdasarkan karakteristik mereka (misalnya, pengetahuan awal, minat, kemahiran berbahasa) dan perilaku yang ditunjukkan (misalnya, tingkat keterlibatan, kendala teknis), guna mendukung interpretasi yang lebih mendalam serta langkah tindak lanjut yang lebih akurat untuk pembelajaran atau asesmen.

Desain universal dalam penilaian. Secara lebih umum, penilaian yang dipersonalisasi harus dirancang untuk memaksimalkan keadilan bagi semua peserta tes. *Standards for Educational and Psychological Testing* (AERA, APA, & NCME) menekankan penerapan prinsip-prinsip desain universal untuk mendorong keadilan dengan memaksimalkan aksesibilitas penilaian bagi beragam individu dan subkelompok, sembari meminimalkan munculnya varians yang tidak relevan dengan konstruk (*construct-irrelevant variance*) yang dapat berdampak secara berbeda terhadap subkelompok tertentu.

Pedoman *Universal Design for Learning* (UDL) merekomendasikan penyediaan beragam cara keterlibatan (misalnya, menawarkan pilihan yang dipersonalisasi yang selaras dengan identitas dan motivasi pembelajar), beragam cara representasi (misalnya, menyajikan informasi dalam berbagai modalitas yang sesuai dengan konstruk yang diukur), serta beragam cara tindakan dan ekspresi (misalnya, fleksibilitas dalam moda respons yang memungkinkan mahasiswa menunjukkan pengetahuan, keterampilan, dan atribut lainnya, termasuk penyediaan sarana pendukung bagi mahasiswa dengan disabilitas perseptual dan motorik; CAST).

Baru-baru ini, Laitusis dan Karvonen mengembangkan karya sebelumnya menjadi kerangka kerja menyeluruh mengenai *Universal Design of Assessment Systems* (UDAS). Kerangka ini melampaui sekadar desain dan pelaksanaan penilaian berskala besar dengan mempertimbangkan bagaimana desain universal dapat diterapkan pada berbagai jenis penilaian (seperti penilaian di kelas dan penilaian formatif) serta mengakomodasi kebutuhan semua pemangku kepentingan terkait (seperti guru, orang tua, dan pengguna skor di luar interaksi antara mahasiswa dan butir soal). UDAS menekankan bahwa penilaian harus: (1) memenuhi kebutuhan psikologis, termasuk otonomi, kompetensi, dan keterhubungan (*relatedness*); (2) dirancang dengan memperhatikan fleksibilitas dan variabilitas pembelajar, termasuk inklusivitas dalam definisi konstruk, fleksibilitas dalam representasi, serta fleksibilitas dalam tindakan dan ekspresi; dan (3) menjamin penggunaan yang adil, bertanggung jawab, dan setara, termasuk melalui evaluasi validitas secara proaktif, mendorong interpretasi yang sesuai tujuan serta mencegah penggunaan yang tidak tepat, meningkatkan literasi penilaian, dan menetapkan konsekuensi atas penyalahgunaan skor. Penerapan UDAS dan prinsip-prinsip terkait dapat membantu memastikan bahwa penilaian yang dipersonalisasi memenuhi kebutuhan seluruh mahasiswa secara fleksibel dan bertanggung jawab.

Personalisasi dinamis berbasis bukti untuk perangkat pembelajaran dan penilaian (ED-PLAT). Berbagai bentuk personalisasi (yaitu, yang digerakkan oleh manusia, mesin, atau hibrida) memerlukan pengambilan keputusan yang didasarkan pada bukti. Saat ini, bukti yang tersedia masih belum memadai mengenai kapan, pada tingkat apa, dan pada fitur mana kendali sebaiknya diberikan kepada peserta tes dan/atau mesin. Untuk menjawab hal ini, Arslan. baru-baru ini memperkenalkan sebuah kerangka kerja, ED-PLAT, yang terdiri dari tiga langkah. Langkah pertama mencakup identifikasi fitur-fitur penting dalam lingkungan pembelajaran dan penilaian (misalnya, konteks butir soal, metode respons) serta peninjauan teori dan literatur yang ada untuk mengidentifikasi dampak fitur-fitur terpilih terhadap motivasi, aspek afektif, kognisi, serta karakteristik persepsi/motorik peserta didik, sehingga dampak yang diharapkan menjadi jelas.

Langkah kedua melibatkan pengumpulan bukti empiris (misalnya, melalui eksperimen terkontrol, uji A/B, atau umpan balik pengguna) untuk menentukan apakah kendali oleh peserta didik (*learner agency*) atau otomatisasi sistem yang paling efektif dalam meningkatkan hasil yang ditetapkan bagi populasi target termasuk peserta didik dengan keberagaman saraf (*neurodiverse*) yang paling berisiko terdampak oleh ketidaksesuaian model peserta didik dalam personalisasi berbasis mesin. Terakhir, langkah ketiga melibatkan penentuan pihak yang memegang kendali pada tingkat fitur apakah peserta didik, mesin, atau keduanya serta pembaruan dinamis atas pembagian kendali tersebut berdasarkan bukti empiris, sembari tetap menempatkan guru sebagai pemantau aktif yang dapat melakukan intervensi ketika data terbatas atau terjadi perubahan konteks.

5.3 BUKTI EFEKTIVITAS PERSONALISASI

Meskipun terdapat berbagai kemungkinan dasar pemikiran dan konsepsi teoretis yang dapat memandu penerapan personalisasi dalam konteks penilaian, bukti empiris yang menunjukkan bahwa personalisasi semacam itu berhasil mencapai tujuannya yaitu mengoptimalkan kinerja dan keterlibatan masih berada pada tahap awal perkembangan. Bukti mengenai efektivitas pendekatan pembelajaran yang dipersonalisasi telah dirangkum dalam berbagai tinjauan sistematis dan meta-analisis. Sebagai contoh, sebuah meta-analisis terhadap 34 publikasi yang mencakup 34 ukuran efek mengkaji dampak pembelajaran terpersonalisasi yang difasilitasi teknologi (dibandingkan dengan kelompok kontrol yang tidak menggunakan metode tersebut) dan melaporkan efek berskala sedang hingga besar (Hedges' g = 0,67) terhadap capaian pembelajaran serta efek berskala kecil (Hedges' g = 0,26) terhadap persepsi pembelajaran (misalnya, sikap dan motivasi); terdapat variasi ukuran efek berdasarkan faktor moderator seperti jumlah metode pembelajaran yang digunakan dan jenis perangkat lunak pembelajaran terpersonalisasi yang diterapkan (yaitu, *prompt* atau instruksi pemicu yang dipersonalisasi menunjukkan ukuran efek terbesar, diikuti oleh panduan terpersonalisasi, diagnosis terpersonalisasi, dan saran pembelajaran terpersonalisasi; Zheng)

Dalam ranah literasi membaca, sebuah meta-analisis terhadap 27 publikasi yang mencakup 27 ukuran efek mengkaji dampak teknologi pembelajaran terpersonalisasi dan adaptif (dibandingkan dengan perangkat lunak pembelajaran non-terpersonalisasi atau materi

pembelajaran tradisional yang tidak berbasis teknologi) bagi mahasiswa jenjang K–12; studi ini mengamati ukuran efek keseluruhan yang kecil namun positif (Hedges' g = 0,29) terhadap kemahiran keterampilan membaca, dengan variasi ukuran efek yang bergantung pada karakteristik studi (yaitu, peran guru, dosis intervensi, dan penggunaan pembelajaran campuran/*blended instruction*) serta karakteristik mahasiswa (misalnya, tingkat kelas dan status sosial-ekonomi; Alrawashdeh).

Melampaui hasil berbasis kinerja, meta-analisis oleh Lin. mengenai personalisasi berbasis minat yang mencakup 26 ukuran efek minat, delapan ukuran efek beban kognitif, 46 ukuran efek retensi, dan enam ukuran efek transfer dari 34 publikasi melaporkan bahwa dibandingkan dengan kondisi kontrol tanpa personalisasi, personalisasi menunjukkan efek sedang hingga besar terhadap minat (Hedges' g keseluruhan = 0,55; pada matematika Hedges' g = 0,64), efek sedang hingga besar terhadap beban kognitif (Hedges' g = 0,54), efek sedang terhadap retensi pembelajaran (Hedges' g = 0,48), dan efek sedang terhadap transfer pembelajaran (Hedges' g = 0,36).

Lin melaporkan adanya heterogenitas yang signifikan pada ukuran efek minat berdasarkan pendekatan diagnostik (efek signifikan untuk pendekatan personalisasi berbasis survei versus berbasis teknologi), tingkat perincian/skala (*grain size*; efek signifikan untuk personalisasi tingkat individu versus tingkat kelompok), dan domain konten (efek signifikan diperoleh untuk matematika versus domain lain), serta heterogenitas signifikan pada ukuran efek retensi berdasarkan benua (studi di Asia dan Amerika Utara menunjukkan efek signifikan, sedangkan studi di Eropa dan Afrika tidak) dan metodologi (kuasi-eksperimen menunjukkan ukuran efek yang lebih besar dibandingkan eksperimen, namun kedua efek tersebut signifikan secara statistik). Dengan demikian, secara umum bukti mendukung penerapan pembelajaran yang dipersonalisasi, dengan catatan bergantung pada jenis personalisasi yang diterapkan, cara penerapannya, dan kepada siapa personalisasi tersebut ditujukan.

Terkait bukti yang mendukung asesmen yang dipersonalisasi, sebagaimana telah disebutkan sebelumnya, saat ini masih sangat sedikit contoh operasional asesmen yang dipersonalisasi yang diterapkan dalam skala luas untuk menentukan dampak personalisasi terhadap kinerja asesmen mahasiswa; meskipun tidak diragukan lagi akan muncul lebih banyak contoh dan bukti akan segera tersedia seiring kemajuan AI yang memungkinkan lebih banyak penerapan personalisasi dalam skala luas. Namun, beberapa contoh terkini yang penting menunjukkan potensi penyelarasan fitur asesmen dengan karakteristik mahasiswa dengan cara-cara yang selaras dengan pendekatan asesmen yang penuh kepedulian (*caring assessment approach*).

Wang mengkaji dampak penilaian yang relevan secara sosiokultural dalam konteks penilaian berbasis skenario (SBA) untuk keterampilan pemahaman bacaan di lingkungan digital. Para penulis meneliti kinerja pemahaman bacaan dari sekitar 7.000 mahasiswa sekolah menengah atas yang mengerjakan salah satu dari tiga bentuk SBA dengan topik berbeda, yaitu ekosistem dan spesies invasif, imigrasi Eropa ke Amerika Serikat, serta gerakan budaya Harlem Renaissance pada awal abad ke-20. Mengingat Harlem Renaissance memiliki signifikansi sejarah dan budaya bagi warga Afrika-Amerika, diprediksi bahwa topik ini akan lebih familier—

dan penilaiannya lebih menarik serta mendukung kinerja bagi mahasiswa kulit hitam dibandingkan dengan topik-topik lainnya. Hasil penelitian memang menunjukkan bahwa mahasiswa kulit hitam memiliki pengetahuan topik yang jauh lebih rendah mengenai ekosistem dan imigrasi dibandingkan mahasiswa non-kulit hitam, namun memiliki tingkat pengetahuan yang setara mengenai topik Harlem Renaissance, berdasarkan pengukuran pengetahuan latar belakang topik yang dilakukan sebelum penilaian. Perbedaan dalam pengetahuan topik ini sejalan dengan perbedaan dalam tingkat keterlibatan saat mengerjakan tes; mahasiswa kulit hitam dan non-kulit hitam menunjukkan tingkat keterlibatan yang serupa (berdasarkan durasi pengerjaan tugas) pada topik Harlem Renaissance, sedangkan mahasiswa kulit hitam menunjukkan tingkat keterlibatan yang lebih rendah pada dua topik lainnya.

Terakhir, perbedaan skor antara kelompok mahasiswa kulit hitam dan non-kulit hitam lebih kecil pada bentuk tes Harlem Renaissance (kedua kelompok menunjukkan kinerja yang setara setelah memperhitungkan pencapaian sebelumnya berdasarkan skor tes standar negara bagian untuk mata pelajaran Bahasa Inggris), sementara mahasiswa kulit hitam memperoleh skor lebih rendah dibandingkan mahasiswa non-kulit hitam pada dua bentuk tes lainnya. Semua bentuk SBA menunjukkan reliabilitas dan kualitas psikometrik yang setara, yang mengindikasikan bahwa bentuk tes yang relevan secara budaya dapat diberikan kepada kelompok mahasiswa sasaran dengan cara yang memaksimalkan keterlibatan, meminimalkan perbedaan skor antar-kelompok, serta menjaga kualitas teknis.

Sparks meneliti lebih lanjut hubungan antara pengetahuan awal peserta tes dengan keterlibatan serta kinerja mereka dalam tes berbasis situasi (SBA) pemahaman bacaan. Dengan menggunakan data dari sekitar 1.300 mahasiswa sekolah menengah pertama, para penulis menerapkan pengukuran pengetahuan latar belakang topik, serta pengukuran baru berupa laporan mandiri mengenai *funds of knowledge* (pengetahuan dan pengalaman yang dimiliki Mahasiswa dari latar belakang budaya dan kehidupan sehari-hari) dan pengalaman mahasiswa, untuk mengkaji hubungan antara berbagai aspek pengetahuan awal dan kinerja mahasiswa pada SBA bertopik pertanian organik. Setelah mengontrol variabel demografis mahasiswa (yaitu ras/etnis, jenis kelamin, tingkat kelas, kelayakan program makan siang gratis/bersubsidi, status pendidikan luar biasa, dan status pembelajar bahasa Inggris), pengukuran kosakata terkait topik menjelaskan tambahan varians sebesar 27% dalam skor SBA.

Memasukkan pengukuran pengetahuan dan pengalaman berdasarkan laporan mandiri ke dalam model menjelaskan tambahan varians yang kecil namun signifikan secara statistik (2%) di luar faktor demografi dan pengetahuan topik. Penambahan variabel keterlibatan saat mengerjakan tes (berdasarkan waktu pengerjaan tugas) menjelaskan tambahan varians sebesar 18% dalam skor SBA. Para penulis menyimpulkan bahwa dengan menerapkan pengukuran laporan mandiri mengenai pengetahuan dan pengalaman di samping pengukuran pengetahuan kosakata topik yang umum digunakan dalam konteks SBA versi tes yang dipersonalisasi secara lebih nuansial atau dukungan dinamis dapat diterapkan dalam SBA untuk meningkatkan efektivitas akses Mahasiswa terhadap materi tes.

Sinharay meneliti dampak penyajian butir soal yang diadaptasi agar memuat kosakata, konteks, dan skenario yang relevan dengan budaya Hispanik (yaitu Meksiko, Dominika, dan Puerto Riko) yang umum dijumpai di AS. Adaptasi dilakukan terhadap butir soal matematika NAEP untuk kelas 8, dan soal-soal tersebut diujikan dalam versi adaptasi maupun versi asli kepada sekitar 950 mahasiswa dari berbagai negara bagian. Analisis menunjukkan bahwa butir soal yang diadaptasi menghasilkan pengurangan sebesar 25% pada selisih skor rata-rata antara mahasiswa Hispanik dan mahasiswa kulit putih, dan seperti yang diperkirakan tidak ada pengurangan berarti pada kesenjangan skor antara mahasiswa kulit putih dan mahasiswa Afrika-Amerika.

Meskipun dampak yang diperoleh untuk butir soal yang dikontekstualisasikan tersebut tergolong kecil, adaptasi yang dilakukan pun minimal dan sebagian besar hanya mencerminkan penyesuaian pada tingkat permukaan, bukan penyesuaian yang lebih mendalam. Para penulis menyarankan potensi arah pengembangan di masa depan, seperti memperluas tingkat adaptasi, mengeksplorasi adaptasi bagi subkelompok peserta tes lainnya (misalnya, mahasiswa Afrika-Amerika), mengkaji dampak adaptasi serupa di bidang lain (misalnya, literasi), serta mengeksplorasi sumber bukti tambahan (misalnya, data proses yang lebih kaya dari interaksi mahasiswa selama asesmen). Dengan demikian, temuan Sinharay menunjukkan bahwa adaptasi pada tingkat permukaan sekalipun berpotensi memengaruhi pengalaman asesmen mahasiswa dengan cara yang meningkatkan aspek keadilan.

5.4 VALIDASI DAN ISU TERKAIT UNTUK ASESMEN PERSONALISASI

Meskipun personalisasi dalam asesmen dapat memberikan berbagai manfaat, masih terdapat kekurangan teori dan penelitian empiris mengenai karakteristik pengukurannya, termasuk aspek validasi. Terdapat beberapa isu utama yang perlu dipertimbangkan dalam menghasilkan skor dari asesmen yang mungkin bervariasi dalam hal konten dan format antarindividu, namun tetap diharapkan menghasilkan interpretasi yang valid, adil, dan dapat diperbandingkan. Kunci bagi validitas dan keadilan terletak pada sejauh mana asesmen yang dipersonalisasi menghasilkan komparabilitas konstruk dan komparabilitas skor. Pada bagian berikut, kami pertama-tama membahas komparabilitas data versus konstruk serta komparabilitas skor, kemudian beralih ke pembahasan mengenai validasi dan keadilan.

Komparabilitas Data versus Konstruk

Terkait dengan dua kutub ekstrem yaitu standardisasi ketat dan personalisasi penuh—Mislevy memperkenalkan konsep komparabilitas data dan komparabilitas konstruk. Kedua aspek ini dimaksimalkan secara bersamaan dalam kondisi standardisasi ketat dan homogenitas populasi yang tinggi, di mana skor dapat dianggap setara (dapat dipertukarkan) antar-bentuk tes. Namun, dalam kondisi keragaman populasi yang tinggi, upaya memaksimalkan komparabilitas data melalui standardisasi mau tidak mau akan mengorbankan komparabilitas konstruk. Sebagai contoh, keharusan bagi seluruh peserta tes untuk menunjukkan pengetahuan dan kemampuan mereka dalam batasan penyajian dan respons yang sempit dapat menghalangi peserta yang tidak mampu bekerja dalam batasan tersebut untuk menunjukkan kompetensinya (misalnya, pada tes penalaran verbal yang disajikan secara

tertulis, kesulitan dalam *decoding* atau pengenalan kata dapat mengganggu proses pengukuran secara tidak relevan). Sebaliknya, dalam kondisi keragaman tinggi yang sama, penyimpangan dari standardisasi yang dimungkinkan oleh personalisasi penuh justru dapat meningkatkan komparabilitas konstruk, karena setiap peserta tes idealnya diukur dengan cara yang sesuai bagi mereka. Akan tetapi, akibatnya daya banding data menjadi berkurang karena setiap orang dinilai secara unik.

Komparabilitas konstruk merupakan klaim yang perlu didukung oleh argumen validitas. Idealnya, argumen tersebut didasarkan pada teori, logika, dan data. Elemen teoretis dan logis dapat mencakup penjelasan mengenai bagaimana definisi konstruk tingkat tinggi, rubrik penilaian, dan proses penilaian bekerja sama sehingga kinerja atau artefak yang unik dapat dipetakan ke dalam skala umum yang sama. Bukti empiris dapat mencakup hasil analisis faktor, analisis tugas kognitif, analisis *log file*, atau metodologi lain untuk menunjukkan bahwa skor-skor tersebut memiliki makna yang kurang lebih serupa.

Komparabilitas Skor pada Asesmen yang Dipersonalisasi

Menurut *Standards for Educational and Psychological Testing* (Standar Pengujian Pendidikan dan Psikologis), "skor yang dapat dibandingkan" (*comparable scores*) merujuk pada skor yang dapat dianggap dapat saling menggantikan (misalnya, AERA, APA & NCME). Dalam tes standar, prosedur psikometrik penyetaraan (*equating*) memungkinkan skor diperlakukan sebagai sesuatu yang dapat saling menggantikan. Beberapa ahli, termasuk Dorans dan Middleton serta Thissen, mencatat bahwa ketika sebuah tes diadaptasi untuk berbagai kelompok populasi guna mengakomodasi perbedaan bahasa, budaya, dan sebagainya, berbagai versi tes tersebut pada dasarnya merupakan "tes yang berbeda secara fundamental" (*disparate tests*).

Asesmen yang dipersonalisasi melibatkan adaptasi berdasarkan berbagai karakteristik individu dengan tingkat yang berbeda-beda, sehingga versi-versi berbeda dari asesmen ini merupakan tes yang berbeda secara fundamental. Thissen juga mencatat bahwa sangat sulit untuk menghasilkan skor yang disetarakan untuk tes-tes yang berbeda secara fundamental karena kurangnya butir soal yang sama dan peserta tes yang sama, yang menghalangi dilakukannya penyetaraan untuk tes-tes tersebut. Mislevy, Sireci, dan Winter juga mengisyaratkan kemungkinan berkurangnya komparabilitas pada asesmen yang dipersonalisasi. Selain itu, pengumpulan data yang memungkinkan penyetaraan skor atau pemeriksaan sejauh mana skor pada asesmen yang dipersonalisasi dapat saling menggantikan merupakan hal yang bermasalah. Namun, praktik operasional dari asesmen lain yang berbeda secara fundamental termasuk asesmen yang diterjemahkan (misalnya, PISA) dan asesmen yang melibatkan akomodasi dapat diterapkan dalam konteks asesmen yang dipersonalisasi untuk mengembangkan beberapa strategi yang memungkinkan klaim adanya tingkat komparabilitas skor tertentu.

Pertama, jika seseorang menilai kompetensi yang dapat diekspresikan dengan berbagai cara, ia dapat menetapkan struktur asesmen yang memungkinkan ekspresi yang luas tersebut, menentukan kriteria penilaian untuk mengakomodasi keluasan itu, dan kemudian melatih penilai secara ekstensif untuk mengembangkan standar bersama dalam memberikan skor yang

cukup sebanding terhadap respons mahasiswa atas tugas-tugas yang mungkin tampak sangat berbeda. Skor-skor ini kemudian dapat dianggap sebagai penilaian pencapaian yang sebanding antar-mahasiswa, dengan mempertimbangkan konteks serta materi, proses, dan gagasan yang mereka pilih untuk dikerjakan. Penilaian untuk asesmen Portofolio Seni dan Desain AP, yang merupakan asesmen yang dipersonalisasi, mengikuti prosedur berikut.

Kedua, penggunaan kerangka kerja berbasis teori untuk pengembangan butir soal dapat memungkinkan pembuatan klaim komparabilitas yang cukup kuat untuk asesmen yang dipersonalisasi. Kerangka kerja semacam itu akan mendeskripsikan fitur-fitur abstrak dari situasi, perilaku, atau kognisi yang dapat mewujudkan suatu konstruk sebagaimana ia muncul dalam berbagai konteks. Salah satu kerangka kerja berbasis teori tersebut adalah *learning progressions* atau alur perkembangan belajar, yaitu hipotesis berbasis penelitian mengenai bagaimana kemampuan mahasiswa dalam suatu bidang berkembang seiring waktu. Penerapan *evidence-centered design* yang memanfaatkan teori mapan dalam mendefinisikan model Mahasiswa, tugas, dan bukti yang mendasari desain asesmen juga dapat berguna dalam menetapkan butir soal yang sebanding.

Ketiga, Mislevy mencatat bahwa pengembangan dan penyusunan tes yang cermat dapat membantu dalam membuat klaim mengenai komparabilitas skor pada asesmen yang dipersonalisasi. Seseorang dapat memulai dari kumpulan butir soal dengan karakteristik tertentu atau prosedur skematis untuk membuat butir soal semacam itu. Pada pendekatan pertama, subset butir soal ditempatkan ke dalam bentuk tes yang berbeda, baik secara acak maupun diseimbangkan berdasarkan berbagai faktor seperti konten butir soal dan karakteristik peserta tes. Dalam pendekatan "prosedur skematis", butir soal dibuat berdasarkan spesifikasi mengenai jenis butir soal, konten, tuntutan kognitif, dan sebagainya, sehingga setiap bentuk tes baru disusun dengan komposisi butir soal yang serupa. Kedua pendekatan tersebut akan memperkuat klaim komparabilitas baik antar-bentuk tes maupun di dalam bentuk tes itu sendiri.

Keempat, dimungkinkan untuk mengandalkan penilaian sekelompok ahli guna memutuskan apakah berbagai versi dari butir soal yang sama memberikan bukti yang sebanding mengenai konstruk yang dituju dan, sebagai konsekuensinya, apakah berbagai bentuk asesmen yang dipersonalisasi dapat dianggap setara (*parallel*). Dalam beberapa kasus, penilaian ahli dapat diterima dengan mudah. Kesetaraan antara tes matematika dalam format Braille dan tes yang sama dalam format cetak tidak diperdebatkan secara serius. Tes yang merupakan hasil penerjemahan dari bentuk dasar umumnya diterima sebagai tes yang sebanding jika penerjemahan tersebut dilakukan dengan cermat. Salah satu contohnya adalah PISA, yang menyelenggarakan tes menggunakan terjemahan dari butir-butir soal yang sama dalam berbagai versi kombinasi negara dan bahasa, sehingga menghasilkan tes yang berbeda satu sama lain. Klaim mengenai keterbandingan skor dari versi-versi tes yang telah diterjemahkan ini didasarkan pada evaluasi oleh pakar serta studi penghubungan psikometris.

Validasi untuk Asesmen yang Dipersonalisasi

Asesmen yang dipersonalisasi dapat dianggap sebagai perluasan dari CAT (*Computerized Adaptive Testing*), yang menyesuaikan asesmen bagi setiap peserta tes

berdasarkan satu karakteristik kognitif (yaitu, konstruk yang sedang dinilai). Validasi untuk asesmen yang dipersonalisasi, dibandingkan dengan CAT tradisional, jauh lebih kompleks, terutama karena adanya pertimbangan mengenai karakteristik tambahan peserta tes dan variabel personalisasi. Sinharay mengusulkan sebuah kerangka kerja untuk validasi asesmen yang dipersonalisasi. Kerangka kerja ini didasarkan pada pendekatan ancaman terhadap validitas (*threats-to-validity*) dari Crooks, yang memandang asesmen sebagai proses yang terbagi ke dalam delapan tahapan yang saling terkait berikut ini:

1. Pemberian tugas asesmen kepada mahasiswa.
2. Pemberian skor atas kinerja mahasiswa pada tugas-tugas tersebut.
3. Agregasi skor dari tugas-tugas individual untuk menghasilkan satu atau lebih skor gabungan (skor total atau skor subskala).
4. Generalisasi dari tugas-tugas spesifik yang tercakup dalam skor gabungan ke seluruh domain tugas serupa (domain yang dinilai).
5. Ekstrapolasi dari domain yang dinilai ke domain target yang lebih luas yang mencakup semua tugas yang relevan dengan interpretasi yang diusulkan.
6. Evaluasi kinerja mahasiswa, yang menghasilkan penilaian atau pertimbangan.
7. Keputusan mengenai tindakan yang akan diambil berdasarkan penilaian tersebut.
8. Dampak terhadap mahasiswa dan partisipan lain yang timbul dari proses asesmen, interpretasi, dan keputusan tersebut.

Validasi dilakukan melalui pemeriksaan cermat terhadap setiap tahapan. Pemeriksaan setiap mata rantai dengan tujuan menemukan potensi kelemahan dalam rangkaian inferensi—memberikan pendekatan sistematis terhadap validasi. Validator sebaiknya memulai dengan mempertimbangkan mata rantai administrasi, kemudian melanjutkan ke tahapan berikutnya dalam daftar di atas, sembari mengevaluasi ancaman terhadap validitas yang terkait dengan setiap mata rantai. Langkah terakhir dalam validasi adalah mengidentifikasi mata rantai terlemah dalam rangkaian tersebut. Mata rantai terlemah inilah yang menjadi titik fokus utama saat berupaya menyempurnakan asesmen.

Sinharay mengkaji ancaman yang terkait dengan setiap mata rantai di atas untuk asesmen yang dipersonalisasi dan membahas apakah tingkat keparahan ancaman tersebut berkurang atau justru meningkat dibandingkan dengan asesmen terstandar (hipotetis) yang bertujuan mengukur pengetahuan, keterampilan, dan kemampuan serupa pada populasi yang serupa pula. Mereka juga mengkaji ancaman terhadap validitas untuk dua asesmen yang dipersonalisasi, termasuk ujian Portofolio Seni dan Desain 3D AP serta asesmen eksperimental yang relevan secara budaya dan dikontekstualisasikan bagi mahasiswa Pribumi. Sinharay menyimpulkan bahwa beberapa ancaman terhadap validitas tampaknya dapat dimitigasi dalam asesmen yang dipersonalisasi. Ancaman-ancaman tersebut meliputi:

- motivasi rendah serta tugas atau respons yang tidak dikomunikasikan (di bawah aspek “Administrasi”);
- pemahaman yang kurang memadai mengenai informasi asesmen dan keterbatasannya, interpretasi konstruk yang kurang didukung secara memadai, serta interpretasi atau penjelasan yang bias (aspek “Evaluasi”);

- keputusan pedagogis yang kurang tepat (aspek “Keputusan”); dan
- konsekuensi positif tidak tercapai, serta timbulnya dampak negatif yang serius (aspek “Dampak”).

Namun, Sinharay mencatat bahwa ancaman utama terhadap validitas asesmen yang dipersonalisasi berkaitan dengan tahapan "Penilaian Skor" (*Scoring*) akibat adanya pertanyaan mengenai komparabilitas skor. Mereka juga mencatat bahwa skor mungkin dapat diperbandingkan untuk beberapa jenis inferensi (misalnya, interpretasi konstruk pada tingkat abstrak) dan penggunaan skor tertentu, namun tidak untuk yang lainnya. Tingkat ancaman terhadap validitas ini bergantung pada penggunaan skor tes dan konsep komparabilitas yang dianut. Untuk ujian lisensi atau sertifikasi, tingkat ancamannya bisa menjadi serius jika kemampuan untuk saling menggantikan (*interchangeability*) menjadi syarat bagi penggunaan tertentu. Akan tetapi, jika penggunaan tersebut dapat mengakomodasi variasi kondisi yang menghalangi kemampuan saling menggantikan namun dianggap dapat meningkatkan validitas (misalnya, menggunakan mobil sendiri untuk tes mengemudi), maka ancamannya mungkin jauh lebih ringan. Tingkat ancaman yang lebih rendah juga dapat terjadi pada penggunaan di mana konsekuensi dari keputusan yang keliru tergolong ringan dan mudah dipulihkan, seperti dalam asesmen yang dipersonalisasi untuk tujuan penyesuaian pembelajaran di kelas. Sebagai alternatif, jika seseorang dapat menggunakan salah satu dari empat strategi yang dibahas pada bagian sebelumnya untuk menyatakan adanya komparabilitas, maka ancaman terhadap validitas yang terkait dengan tahapan "Penilaian Skor" akan berkurang secara signifikan.

Pendekatan validasi lain juga dapat diterapkan, termasuk kerangka kerja terpadu Messick atau kerangka kerja Embretson, di mana validitas dikonseptualisasikan sebagai sistem yang universal dan interaktif. Namun, isu mengenai komparabilitas skor kemungkinan besar akan tetap menjadi persoalan penting dalam pendekatan apa pun. Sebagai contoh, kerangka kerja Embretson mencakup 11 kategori bukti mengenai makna dan signifikansi tes, termasuk kategori "properti psikometrik" dan "model penilaian skor." Akibatnya, potensi kurangnya komparabilitas akan melemahkan bukti validitas dalam kedua kategori tersebut.

5.5 KEADILAN DALAM INTERPRETASI SKOR DARI ASESMEN PERSONALISASI

Kane menekankan dua aspek utama keadilan: (1) konsistensi makna skor dan (2) perlakuan yang adil terhadap berbagai kelompok. Konsistensi makna skor mencakup upaya memastikan bahwa kesimpulan dan inferensi yang sama dapat ditarik dari hasil tes untuk berbagai kelompok. Aspek keadilan ini telah menjadi fokus utama dalam penelitian mengenai keadilan. Aspek ini berarti bahwa penilaian memberikan peluang yang setara untuk terlibat dalam proses penilaian, menjamin kesetaraan konstruk yang diukur, serta menghasilkan skor yang dapat diperbandingkan bagi individu dari berbagai latar belakang. Bagi populasi yang sangat heterogen, upaya untuk mengutamakan keadilan antarkelompok menghadirkan tantangan yang signifikan. Kelompok demografis yang umumnya menjadi fokus dalam penelitian mengenai keadilan seperti kelompok berdasarkan jenis kelamin, ras/etnis, atau bahasa ibu biasanya memiliki tingkat heterogenitas yang tinggi. Faktanya, personalisasi menantang gagasan bahwa prinsip pengukuran utama, seperti keadilan, harus berfokus pada

bukti konsistensi makna skor di tingkat kelompok. Secara khusus, personalisasi yang berlandaskan perspektif sosiokultural mempertanyakan asumsi homogenitas di dalam kelompok (*within-group homogeneity*) yang digunakan dalam analisis statistik untuk mendeteksi *differential item functioning* (DIF) metode yang lazim digunakan untuk mengevaluasi keadilan dalam penilaian.

Penelitian terdahulu menunjukkan bahwa peningkatan heterogenitas dalam sampel kelompok fokus (*focal-group*) mengurangi kemampuan deteksi terhadap butir soal yang bias. Temuan ini mengindikasikan bahwa pelanggaran terhadap asumsi homogenitas memiliki implikasi signifikan bagi analisis keadilan dan bias yang umum dilakukan. Keterbatasan metodologi DIF untuk populasi yang sangat heterogen menegaskan bahwa konsep keadilan kita saat ini yang didefinisikan pada tingkat kelompok tidak dapat diterapkan pada tingkat individu. Penilaian yang dipersonalisasi membuka peluang untuk meninjau kembali konsep keadilan dari perspektif individu.Perlakuan yang setara terhadap berbagai kelompok mengacu pada keseragaman kondisi tes bagi para peserta tes. Secara tradisional, perlakuan setara ini dicapai melalui standardisasi. Konsep ini berasumsi bahwa meskipun aspek-aspek tertentu dalam suatu asesmen mungkin menguntungkan sebagian mahasiswa, aspek-aspek lain akan menguntungkan mahasiswa lainnya, sehingga secara rata-rata menyeimbangkan berbagai keuntungan dan kerugian yang ada. Namun, sebagaimana telah dibahas sebelumnya, standardisasi tidak selalu menciptakan kondisi tes yang setara bagi individu yang berasal dari latar belakang sosiokultural yang beragam.

Akibatnya, pendekatan ini mungkin tidak secara konsisten memberikan bukti yang dapat diperbandingkan mengenai pengetahuan, keterampilan, dan kemampuan mahasiswa. Individu dengan pengalaman tes yang beragam seperti tingkat keakraban dengan format tes dan lingkungan yang kaya teknologi serta mereka yang berasal dari budaya dan konteks yang tidak terbiasa melakukan tugas kognitif dalam batasan waktu tertentu atau memiliki cara pandang berbeda terhadap tugas asesmen, mungkin berinteraksi dengan tes melalui cara-cara khas yang menghasilkan skor dengan tingkat validitas lebih rendah.

Mengembangkan konsep "keadilan marginal" (*marginal fairness*) yang diperkenalkan oleh Hansen terkait aksesibilitas tes, Mislevy mengusulkan konsep "keadilan bersyarat" (*conditional sense of fairness*). Konsep ini berarti mempertimbangkan informasi spesifik pada tingkat individu untuk tujuan penarikan kesimpulan—alih-alih merata-ratakan variasi yang ada. Pertimbangan ini menyiratkan bahwa aspek-aspek asesmen divariasikan secara sengaja agar setiap mahasiswa dapat mengakses, berinteraksi dengan, dan merespons tugas dengan hambatan minimal, sehingga tantangan utama yang dihadapi benar-benar berfokus pada konstruk yang hendak diukur.

Pendekatan ini berbeda dari standardisasi konvensional dan lebih menekankan pada variasi sistematis antarindividu dalam pelaksanaan maupun kondisi asesmen demi memperoleh bukti yang setara mengenai konstruk target. Contoh dari variasi semacam ini meliputi pengujian mahasiswa penyandang disabilitas dengan menggunakan akomodasi, adaptasi tes ke dalam berbagai bahasa, serta asesmen yang dipersonalisasi sebagaimana telah dibahas sebelumnya dalam bab ini.

Dalam asesmen yang dipersonalisasi, gagasan mengenai keadilan, konsistensi makna skor, dan perlakuan yang setara berkaitan dengan sejauh mana personalisasi mencapai tujuannya bagi individu: yaitu mengoptimalkan kinerja dan keterlibatan, serta mengukur kompetensi individu secara akurat. Oleh karena itu, aspek kunci keadilan dalam asesmen yang dipersonalisasi adalah sejauh mana tujuan personalisasi tersebut tercapai. Dalam evaluasi keadilan, diperlukan bukti empiris untuk memastikan apakah personalisasi:

- Lebih selaras dengan minat, karakteristik, dan latar belakang pembelajar; serta
- Meningkatkan keterlibatan, kinerja, dan ketepatan pengukuran, idealnya bagi semua individu.

5.6 TANTANGAN DAN ARAH PENGEMBANGAN PERSONALISASI

Personalisasi yang efektif membutuhkan pendekatan yang bijaksana dan berbasis bukti yang mempertimbangkan baik adaptivitas yang didorong oleh mesin maupun adaptabilitas yang didorong oleh manusia. Pendekatan tersebut harus memperlakukan peserta didik secara holistik dengan memanfaatkan kinerja, pengetahuan dan keterampilan, upaya, minat, latar belakang bahasa, dan identitas budaya mereka sebelumnya untuk memandu personalisasi konteks, tugas, dan item penilaian. Meskipun AI dapat meningkatkan skala dan mengoptimalkan personalisasi, peran aktif peserta didik tetap penting untuk memastikan pengalaman penilaian tetap inklusif, menarik, dan selaras dengan kebutuhan dan potensi setiap peserta didik. Memanfaatkan kerangka kerja seperti penilaian yang penuh perhatian, UDAS, dan ED-PLAT dapat membantu pengembang penilaian memandu kapan, di mana, dan bagaimana mereka menerapkan personalisasi.

Salah satu tantangan signifikan dalam penilaian yang dipersonalisasi adalah memastikan validitas, keadilan, dan keterbandingan klaim penilaian yang dihasilkan, tantangan yang menjadi lebih kompleks seiring meningkatnya konsekuensi penilaian. Tantangan ini termasuk mengatasi kompatibilitas pengukuran di berbagai tugas dan di berbagai individu dari latar belakang yang berbeda. Selain itu, diperlukan bukti empiris untuk mendukung personalisasi dalam berbagai konteks dan tujuan, untuk menetapkan bagaimana dan kapan personalisasi dilakukan berdasarkan karakteristik dan perilaku tertentu, dan untuk mengidentifikasi kualitas pembelajar mana yang tidak perlu dipertimbangkan. Yang terpenting, penelitian harus berfokus pada validasi, keterbandingan, dan keadilan dari bentuk penilaian personalisasi ini, serta pada dampak berbagai jenis personalisasi terhadap keterlibatan pembelajar, kinerja, dan ketepatan pengukuran.

BAB 6
ALUR KERJA NLP MODERN DALAM PENGUKURAN PENDIDIKAN

Kemajuan teknologi yang pesat telah memperluas cakupan data penilaian hingga mencakup butir soal respons konstruksi (*constructed-response* atau CR), di mana peserta ujian menyajikan jawaban mereka dalam bentuk teks tertulis yang terdiri dari satu atau lebih paragraf, ataupun respons lisan dengan beragam panjang durasi. Butir soal semacam ini menawarkan berbagai manfaat, seperti menilai kemampuan berpikir tingkat tinggi dan organisasi semantik mahasiswa, serta mengidentifikasi miskonsepsi. Sebagian besar upaya penelitian terkait butir soal CR berfokus pada pengembangan teknik untuk memberikan penilaian secara otomatis, meminimalkan ketergantungan pada penilaian manual, mengurangi waktu yang dibutuhkan untuk penilaian, serta meningkatkan konsistensi skor. Namun, berbeda dengan skor pada butir soal pilihan ganda, respons tertulis maupun lisan tidak mengikuti format yang baku. Bahasa merupakan sistem kompleks yang memiliki aturan tata bahasa, struktur semantik, dan konvensi penulisan, di mana makna kata sering kali berubah bergantung pada konteksnya. Akibatnya, hubungan yang kompleks antar-kata tersebut tidak mudah ditangkap melalui nilai kategorikal atau numerik semata.

Pemrosesan bahasa alami (NLP) telah mengatasi kompleksitas tersebut melalui pendekatan berbasis alur kerja (*pipeline*) yang mengubah teks tidak terstruktur—yang selanjutnya disebut sebagai dokumen (misalnya, respons peserta tes terhadap butir soal *constructed-response* atau CR)—menjadi representasi numerik terstruktur guna meningkatkan keterpahaman dan keandalan analisis data tahap lanjutan. Pendekatan awal terhadap analisis teks muncul pada tahun 1960-an seiring dengan pengembangan representasi *bag-of-words* dan matriks dokumen-istilah (*document-term matrices* atau DTM), yang mengukur frekuensi kemunculan kata di dalam setiap dokumen. Representasi-representasi ini meletakkan dasar bagi apa yang kemudian menjadi alur kerja NLP standar yang menekankan pada pembersihan teks (misalnya, menghilangkan tanda baca dan kata-kata yang, meskipun sering muncul, tidak menambah makna interpretatif), lematisasi (misalnya, mengubah berbagai bentuk kala/ *tense* dari kata *begin* menjadi bentuk dasarnya/ *infinitive*), tokenisasi (yaitu, memisahkan dokumen menjadi token-token seperti kata), dan ekstraksi fitur (yaitu, memetakan dokumen dan kata ke dalam DTM).

Urutan langkah ini, yang selanjutnya disebut sebagai alur kerja (*pipeline*) NLP tradisional, telah banyak digunakan dalam penelitian pendidikan, termasuk pada penerapan awal sistem penilaian otomatis atau *automated scoring systems* (AES) melalui analisis semantik laten (*latent semantic analysis*; Deerwester, Dumais, Furnas, Landauer, & Harshman) serta dalam menelaah penalaran mahasiswa saat menjawab butir soal respons konstruksi (*constructed-response* atau CR) menggunakan alokasi Dirichlet laten (*latent Dirichlet allocation* atau LDA; Blei, Ng, & Jordan).

Penelitian yang lebih baru telah mengintegrasikan teknik NLP untuk mendukung analisis dan pemberian umpan balik pada tulisan, serta sistem AES. Meskipun representasi

dokumen dalam bentuk matriks dokumen-istilah (*document-term matrix* atau DTM) memberikan gambaran terstruktur atas jawaban peserta tes, pendekatan tradisional ini mengasumsikan bahwa kata-kata merupakan fitur yang independen, sehingga mengabaikan urutan dan hubungan kontekstualnya. Keterbatasan ini dapat menjadi kontraproduktif saat menganalisis respons yang kompleks dan bergantung pada konteks, karena metode ekstraksi fitur tradisional mungkin tidak memadai. Kemajuan terkini dalam kecerdasan buatan telah mengatasi berbagai tantangan yang dihadapi oleh alur kerja NLP tradisional dengan memanfaatkan model bahasa *deep learning*, yang umumnya dikenal sebagai model bahasa berskala besar atau *large language models* (LLM; Arisoy, Sainath, Kingsbury, & Ramabhadran). Model-model ini disebut "mendalam" (*deep*) karena menggunakan berbagai lapisan jaringan saraf yang bekerja secara bersamaan untuk memahami pola Bahasa termasuk semantik dan sintaksis—alih-alih mengandalkan fitur yang dirancang secara manual. LLM dilatih menggunakan kumpulan data eksternal berskala masif (seperti buku, artikel, dan halaman web), yang memungkinkan model tersebut mempelajari makna kata berdasarkan konteksnya. Dengan memodelkan konteks, NLP modern menghasilkan representasi numerik kontinu untuk kata dan dokumen—yang umum disebut sebagai *embedding*—yang secara otomatis menangkap makna semantik serta hubungan antarunsur, termasuk dalam kasus di mana kata yang sama dapat memiliki makna berbeda bergantung pada konteks linguistiknya. Akibatnya, model-model ini memerlukan tahap pra-pemrosesan teks yang lebih sedikit dibandingkan dengan alur kerja NLP tradisional.

Paradigma ini, yang selanjutnya disebut sebagai *pipeline* NLP modern, menandai pergeseran dari representasi berbasis penghitungan (*count-based*) menuju representasi bahasa yang peka terhadap konteks (*context-aware*). Salah satu LLM yang paling serbaguna adalah *Bidirectional Encoder Representations from Transformers* (BERT; Devlin, Chang, & Lee). BERT merupakan model fleksibel yang menjalankan dua peran berbeda dalam analisis teks. Pertama, BERT dapat digunakan sebagai ekstraktor fitur yang mengubah respons tertulis menjadi *embedding*; *embedding* ini kemudian dapat berfungsi sebagai input bagi model hilir (*downstream models*), seperti pemodelan topik atau analisis berbasis regresi. Dalam peran ini, BERT bertindak sebagai komponen modular yang menyediakan representasi dokumen yang peka terhadap konteks, sembari mengandalkan model statistik atau *machine learning* terpisah untuk keperluan eksplorasi atau inferensi. Kedua, BERT dapat digunakan sebagai kerangka kerja *end-to-end*, di mana *embedding* disesuaikan (*fine-tuned*) untuk memprediksi atau mengklasifikasikan hasil secara langsung dari respons mentah peserta tes misalnya, untuk tugas penilaian otomatis. Dalam skenario ini, ekstraksi fitur dan prediksi dipelajari secara bersamaan dalam satu model tunggal, alih-alih dilakukan sebagai langkah-langkah yang terpisah.

Bagi para peneliti dan praktisi di bidang pengukuran pendidikan, perubahan ini mungkin terasa asing. Alih-alih hanya mengandalkan skor butir soal yang dapat diamati dan asumsi parametrik yang eksplisit, model NLP modern mempelajari pola-pola kompleks secara otomatis. Model-model ini dapat berfungsi layaknya "kotak hitam" (*black box*), yang memunculkan pertanyaan penting mengenai keterpahaman (*interpretability*), validasi, dan

keselarasan dengan teori pengukuran. Tantangan tambahan lainnya adalah karena model seperti BERT dapat digunakan untuk berbagai tujuan, alur kerja (*pipeline*) NLP tidaklah seragam untuk setiap tugas; sebaliknya, alur tersebut harus disesuaikan dengan tujuan penelitian tertentu.

Bab ini menyajikan tinjauan komprehensif mengenai alur kerja NLP modern untuk tugas pengukuran pendidikan yang melibatkan respons tertulis mahasiswa terhadap butir soal respons terbuka (*constructed-response* atau CR) dengan menggunakan BERT. Bab ini terutama ditujukan bagi pembaca di bidang pengukuran pendidikan yang memiliki sedikit pengalaman sebelumnya dengan NLP. Kami meminimalkan penggunaan notasi matematika yang rumit, mendefinisikan istilah teknis saat pertama kali digunakan, serta mengaitkan pilihan pemodelan dengan konsep-konsep yang umum dalam pengukuran pendidikan. Bab ini membahas langkah-langkah utama dalam alur kerja tersebut, serta mengilustrasikan bagaimana teks tidak terstruktur dapat diubah menjadi representasi bermakna untuk analisis dan interpretasi data. Untuk memperjelas peran ganda BERT dan perbedaan alur kerja yang dihasilkannya, kami menyajikan dua contoh penerapan dalam bidang pendidikan.

Pertama, untuk tugas penilaian otomatis, BERT digunakan sebagai kerangka kerja *end-to-end* di mana *embedding* dan prediksi skor dipelajari secara bersamaan. Kedua, untuk analisis eksploratif yang bertujuan mengungkap pola penalaran dan pemikiran peserta tes, BERT digunakan sebagai ekstraktor fitur untuk menghasilkan *embedding* yang berfungsi sebagai input bagi pemodelan topik; hal ini menghasilkan pendekatan terintegrasi yang umum disebut sebagai BERTopic.

6.1 ALUR KERJA NLP BERBASIS BERT SECARA UMUM

Bagian ini menguraikan langkah-langkah utama dalam alur kerja NLP berbasis BERT, yang mencakup (1) pra-pemrosesan teks, (2) ekstraksi fitur, serta (3) pemodelan dan evaluasi. Meskipun langkah-langkah ini umum ditemukan dalam berbagai tugas, penerapannya bervariasi tergantung pada tujuan tugas pengukuran pendidikan tersebut.

Pra-pemrosesan Teks

Karena respons tertulis mentah dari peserta tes sering kali mengandung banyak gangguan (*noisy*), tidak terstruktur, dan tidak konsisten, alur kerja NLP biasanya dimulai dengan mengubah respons tertulis tersebut menjadi urutan terstruktur yang terdiri dari unit-unit numerik bermakna secara linguistik. Tujuannya adalah untuk mengurangi variasi bahasa sekaligus mempertahankan informasi yang relevan untuk interpretasi dan/atau penilaian respons tersebut. Namun, dalam alur kerja berbasis BERT, meskipun lematisasi dan penghapusan *stopword* sering kali tidak diperlukan saat menggunakan model bahasa, koreksi ejaan tetap dapat meningkatkan kinerja model, terutama jika kesalahan ejaan cukup parah hingga mengganggu interpretasi model. Karena bergantung pada model yang telah dilatih sebelumnya (*pretrained models*), tokenisasi tidak termasuk dalam tahap pra-pemrosesan teks; sebaliknya, proses ini ditangani secara internal selama ekstraksi fitur.

Ekstraksi Fitur

Setelah tahap pra-pemrosesan teks, metode ekstraksi fitur mengubah data tekstual menjadi representasi numerik yang sesuai untuk analisis data tahap selanjutnya. Metode *embedding* awal, seperti Word2Vec, mengandalkan model bahasa jaringan saraf sederhana yang mempelajari representasi tetap untuk setiap kata, alih-alih memperbarui maknanya berdasarkan konteks. Sebagai contoh, kata *bank* akan memiliki representasi yang sama baik saat merujuk pada lembaga keuangan maupun tepi sungai. Namun, representasi semacam ini menjadi bermasalah dalam penilaian pendidikan karena peserta ujian mungkin menggunakan kata-kata secara kreatif, dan perbedaan makna yang halus bisa jadi krusial untuk penilaian yang akurat.

BERT mengatasi keterbatasan ini dengan menggunakan arsitektur *transformer* yang menyertakan *positional encoding* untuk mempertahankan urutan kata serta menggunakan lapisan *multi-head self-attention* untuk menyesuaikan representasi kata secara dinamis berdasarkan konteks, sehingga mengatasi keterbatasan metode *word embedding* awal. BERT menerapkan pendekatan ini menggunakan *encoder transformer* dua arah (*bidirectional*) berlapis banyak. Dalam arsitektur ini, representasi kata disempurnakan melalui lapisan-lapisan yang berurutan, dengan menggabungkan informasi dari kata-kata sebelum dan sesudahnya dalam respons tersebut. Pada setiap lapisan, BERT pertama-tama mengidentifikasi kata-kata lain dalam respons yang paling relevan untuk menafsirkan suatu kata menggunakan mekanisme *self-attention*. Kemudian, model ini menggabungkan informasi kontekstual tersebut melalui jaringan saraf *feed-forward* untuk memperbarui representasi kata itu. Sebagai contoh, kata *bank* dipengaruhi lebih kuat oleh kata *loan* (pinjaman) dalam konteks keuangan dan oleh kata *river* (sungai) dalam konteks geografis. Pengulangan proses ini di berbagai lapisan memungkinkan BERT menyusun representasi yang kaya dan peka-konteks atas respons peserta ujian.

Bergantung pada tugas pengukuran pendidikan yang dilakukan, model BERT memiliki variasi dalam kompleksitas arsitektur (misalnya, jumlah lapisan, *attention head*, dan parameter) serta jenis *embedding* yang dihasilkannya. Model tujuan umum yang umum digunakan meliputi *bert-base-uncased*, *bert-base-cased*, dan model yang lebih ringan yaitu *DistilBERT-base-uncased*. Model-model khusus domain, seperti BioBERT dan ClinicalBERT, telah menunjukkan peningkatan kinerja dalam konteks ilmiah dan biomedis. Namun, belum ada konsensus mengenai model yang optimal untuk pengukuran pendidikan, sehingga mendorong para peneliti untuk mengevaluasi berbagai arsitektur guna melihat kinerjanya pada tugas-tugas spesifik. Selain itu, bergantung pada tugas pendidikan yang dilakukan, BERT dapat digunakan untuk mengekstrak *embedding* tingkat kata yang menangkap informasi semantik kata atau *embedding* tingkat dokumen yang merangkum makna keseluruhan dari jawaban tertulis peserta tes.

Pemodelan dan Evaluasi

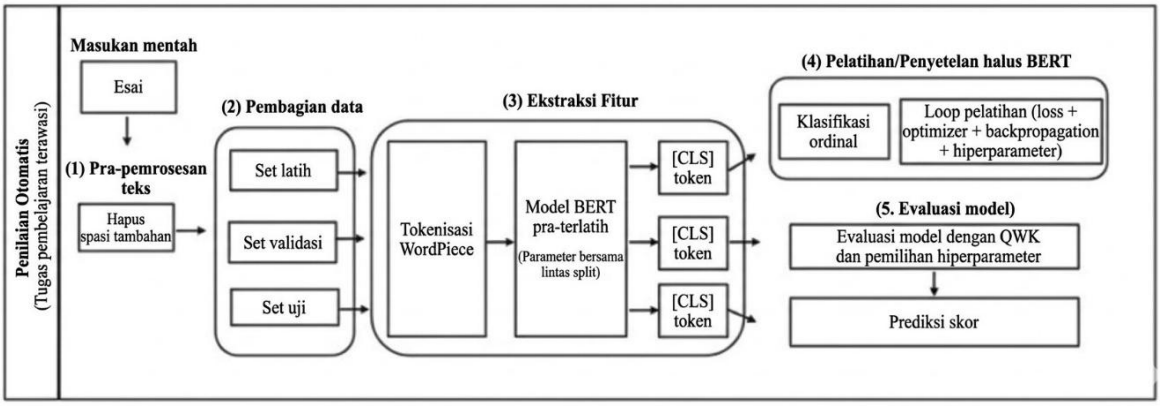
BERT dapat diintegrasikan ke dalam analisis pengukuran pendidikan melalui dua cara berbeda, bergantung pada tujuan penelitian. Pertama, jika tujuannya adalah memprediksi skor untuk sampel peserta tes yang baru seperti dalam tugas penilaian otomatis respons yang

sebelumnya telah dinilai oleh para ahli dibagi menjadi set pelatihan dan set evaluasi, kemudian BERT disesuaikan (*fine-tuned*) secara langsung untuk tugas prediksi tersebut. Dengan kata lain, *embedding* dan parameter penilaian dipelajari secara bersamaan untuk mengoptimalkan akurasi prediksi skor bagi respons peserta tes yang belum pernah dilihat sebelumnya. Kedua, jika tujuannya adalah mengungkap struktur laten yang mendasari respons tertulis peserta tes, kita menggunakan pendekatan pemodelan topik. Dalam hal ini, *embedding* tingkat dokumen yang telah dipelajari digunakan untuk mengekstrak pola yang berkaitan dengan penalaran, proses berpikir, atau konstruk laten lainnya yang muncul dari data respons. Karena pendekatan ini mengandalkan *embedding* alih-alih DTM (*Document-Term Matrix*), model topik yang dihasilkan disebut BERTopic. Masing-masing pendekatan ini akan dijelaskan lebih lanjut secara berturut-turut pada Bagian 6.2 dan 6.3.

Setelah menerapkan teknik pembelajaran mesin dan memperoleh hasilnya, diperlukan mekanisme evaluasi untuk menilai kualitas kinerja model, yang disesuaikan dengan tugas pengukuran pendidikan tertentu. Sebagai contoh, ketika tujuannya adalah memprediksi skor peserta tes, peneliti biasanya menilai tingkat kesesuaian antara skor prediksi berbasis BERT dan skor acuan yang diberikan oleh penilai manusia. Sebaliknya, jika tujuannya adalah mengeksplorasi atau meringkas jawaban tertulis peserta tes ke dalam klaster atau topik, proses evaluasi berfokus pada penilaian koherensi semantik kata-kata dalam setiap topik serta tingkat perbedaan kata antar-topik.

6.2 ALUR KERJA NLP BERBASIS BERT UNTUK PENILAIAN OTOMATIS

Bagian ini menguraikan bagaimana alur kerja (*pipeline*) berbasis BERT dapat diadaptasi untuk penilaian otomatis. Gambaran umum mengenai alur kerja tersebut disajikan dalam Gambar 6.1, dan setiap komponen dijelaskan secara rinci pada sub-bagian berikutnya. Terakhir, bagian ini ditutup dengan contoh empiris yang mengilustrasikan bagaimana alur kerja tersebut diterapkan dalam praktik. Alur kerja berbasis BERT untuk penilaian otomatis biasanya terdiri dari empat tahap utama: (1) pra-pemrosesan teks; (2) pembagian data yang telah diproses sebelumnya menjadi set pelatihan, validasi, dan pengujian; (3) ekstraksi fitur; serta (4) penyesuaian halus (*fine-tuning*) dan evaluasi model menggunakan data yang disisihkan (*held-out data*).



Gambar 6.1 Alur kerja NLP untuk penilaian otomatis menggunakan BERT.

Pra-pemrosesan Teks

Melanjutkan Bagian 6.1, tahap pertama dalam alur kerja NLP adalah mengubah esai mentah peserta ujian menjadi format terstruktur yang dapat diproses oleh model. Secara umum, satu-satunya tugas pra-pemrosesan dalam pengaturan ini melibatkan normalisasi spasi, termasuk menghapus spasi di awal dan akhir teks serta memperbaiki kesalahan spasi.

Pembagian Data

Dengan teks yang telah melalui pra-pemrosesan, data dibagi menjadi tiga subset: (1) pelatihan (*training*), (2) validasi, dan (3) pengujian (*testing*), guna mendukung pengembangan dan evaluasi model yang andal. Set pelatihan digunakan untuk menyempurnakan (*fine-tune*) parameter BERT dengan mempelajari pola dari teks yang telah diubah menjadi representasi vektor (*embedded text*). Set validasi digunakan untuk memantau kinerja di berbagai iterasi dan memilih konfigurasi hiperparameter yang membantu mencegah *overfitting*.

Setelah konfigurasi model akhir dipilih, kinerja dievaluasi menggunakan set pengujian; set ini sama sekali tidak digunakan selama proses penyempurnaan maupun pemilihan model, melainkan hanya untuk memberikan evaluasi yang tidak bias mengenai seberapa baik model yang telah dilatih dapat melakukan generalisasi terhadap data yang belum pernah dilihat sebelumnya. Singkatnya, set pelatihan digunakan untuk mempelajari parameter model, set validasi digunakan untuk memilih konfigurasi model yang optimal, dan set pengujian digunakan semata-mata untuk mengevaluasi kesesuaian akhir antara prediksi model dan penilaian penilai manusia. Pembagian ini biasanya dilakukan secara acak, atau dengan stratifikasi untuk menjaga keseimbangan kelas, guna memastikan bahwa setiap subset merepresentasikan keseluruhan dataset.

Ekstraksi Fitur

Ekstraksi fitur dalam aplikasi ini berfokus pada pengubahan respons tertulis peserta ujian menjadi representasi numerik untuk keperluan prediksi skor. Proses ini bergantung pada model yang digunakan dan dimulai dengan pemilihan arsitektur BERT yang telah dilatih sebelumnya (*pretrained*); arsitektur ini menentukan bagaimana respons tertulis dikonversi menjadi unit yang kompatibel dengan model serta menetapkan tingkat granularitas dari representasi yang dipelajari.

Setelah model *pretrained* dipilih, respons peserta ujian diproses menggunakan *tokenizer* yang sesuai dengan model tersebut melalui teknik tokenisasi WordPiece. WordPiece memecah kata menjadi sub-kata (*subwords*) berdasarkan frekuensinya dan memberikan pengenal unik bagi setiap sub-kata dari kosakata yang telah ditentukan sebelumnya. Hasilnya, kata-kata umum seperti *those* dan *kids* tetap utuh tanpa dipecah. Sebaliknya, kata-kata yang jarang muncul seperti *bioluminescence* dipecah menjadi bagian-bagian subkata (misalnya: *bio*, ##*lum*, ##*ine*, ##*scence*), di mana awalan ## menunjukkan bahwa bagian kata tersebut menyambung dengan bagian sebelumnya. Pendekatan ini memungkinkan model untuk menggunakan kembali representasi yang telah dipelajari sebelumnya guna menafsirkan kata-kata yang asing atau jarang ditemui, alih-alih menganggapnya sebagai kata yang tidak dikenal.

Selama proses tokenisasi, BERT juga menyisipkan token khusus untuk mendukung tugas penilaian otomatis. [CLS] adalah token yang ditambahkan di awal setiap dokumen, yang

berfungsi sebagai representasi gabungan untuk keseluruhan urutan teks. [SEP] adalah token yang menandai batas antara dua kalimat. Untuk meningkatkan efisiensi pemrosesan, BERT memproses teks dalam bentuk *batch* (kelompok), artinya model menangani beberapa respons secara paralel. Karena tidak semua jawaban memiliki panjang yang sama, BERT menambahkan token [PAD] yang tidak memiliki makna pada dokumen-dokumen yang lebih pendek. Untuk contoh ilustratif mengenai proses ini, silakan lihat Contoh 6.1.

Setelah tokenisasi, respons peserta ujian diteruskan melalui lapisan *embedding* yang menghasilkan representasi numerik untuk setiap token (yaitu setiap kata atau subkata, termasuk token [CLS]), serta *embedding* posisi yang mengodekan urutan kata (Adoma, Henry, Chen, & Andre). Melalui lapisan-lapisan yang berurutan, token-token tersebut berinteraksi menggunakan mekanisme *self-attention*, yang memungkinkan representasi mereka diperbarui berdasarkan konteks di sekitarnya. Secara khusus, token [CLS] mengumpulkan informasi kontekstual dari semua token lainnya dan umum digunakan sebagai representasi dengan panjang tetap untuk keseluruhan dokumen. Dalam penilaian esai otomatis, *embedding* [CLS] akhir ini biasanya berfungsi sebagai fitur masukan bagi model prediksi. Pada tahap ini, *embedding* [CLS] merupakan representasi tujuan umum yang dipelajari selama proses pra-pelatihan BERT. *Embedding* tersebut kemudian disesuaikan (*fine-tuned*) selama pelatihan model untuk menangkap pola-pola yang spesifik bagi tugas penilaian otomatis. Contoh ilustratif mengenai proses ini disajikan dalam Contoh 6.2 dan 6.3.

Contoh 6.1: Tokenisasi WordPiece dan Padding pada BERT

Token Khusus

- [CLS] ditempatkan di awal jawaban peserta (untuk tugas klasifikasi).
- [SEP] menandai akhir jawaban peserta atau memisahkan beberapa kalimat dalam jawaban peserta.
- [PAD] ditambahkan ke urutan yang lebih pendek agar semua input dalam satu *batch* memiliki panjang yang sama.

Contoh. Misalkan seorang peneliti memproses jawaban tertulis dari empat peserta secara bersamaan (*batch size* = 4). Tiga jawaban berisi 13 token, namun jawaban keempat hanya berisi 11 token. Untuk memastikan panjang input yang konsisten, BERT menambahkan dua token *padding* ke urutan yang lebih pendek tersebut.

- **Urutan sebelum *padding***: [CLS] those kids like eating ice ##cream every single day [SEP].
- **Urutan setelah *padding***: [CLS] those kids like eating ice ##cream every single day [SEP] [PAD] [PAD].

BERT juga membuat *attention mask* (atau *input mask*) yang menunjukkan token mana yang merupakan data asli (1) dan mana yang merupakan *padding* (0). Untuk contoh ini, *attention mask* bagi peserta keempat adalah: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0].

Contoh 6.2: Penyematan Kata di BERT

1 **Tokenisasi**:
[CLS] anak-anak itu suka makan es ## krim [SEP] [PAD] [PAD]

2 **Penyematan**: Setiap token diubah menjadi vektor dimensi tinggi dengan panjang 768 (dalam BERT_base) atau 1,024 (dalam BERT_large).

ice → [0.31, -0.12, 0.45, ..., 0.27] ## cream → [0.12, -0.34, 0.56, ..., -0.09]

3 **Kontekstualisasi melalui perhatian diri**: Pada setiap lapisan transformator, setiap vektor token diperbarui berdasarkan hubungannya dengan semua token lain dalam kalimat, sehingga dimensi vektor mewakili fitur laten yang menangkap informasi semantik, sintaksis, dan kontekstual.

Layer1: ## cream = [0.2, -0.3, 0.5, ..., 0.1]
Layer6: ## cream = [0.5, -0.1, 0.7, ..., 0.4]
Layer12: ## cream = [0.8, 0.3, 0.1, ..., 0.6]

4 **Keluaran**: Setelah lapisan terakhir, vektor 768 dimensi harus mengkodekan makna token dalam konteks spesifiknya. Misalnya, representasi akhir dari "##krim" harus berbeda di setiap kalimat:

ice ## cream ≠ whipped ## cream

Token khusus [CLS], yang ditampilkan di awal urutan, berfungsi sebagai representasi agregat dari keseluruhan respons dan biasanya digunakan sebagai masukan untuk tugas penilaian otomatis.

Contoh 6.3: *Self-Attention* + Jaringan Feedforward
Contoh ini menggunakan kalimat yang sama dengan Contoh 6.2, namun berfokus secara lebih rinci pada proses yang terjadi di dalam satu lapisan encoder transformer.

1 **Input ke Lapisan l**: Misalkan lapisan sebelumnya menghasilkan vektor kontekstual untuk setiap token:

ice → [0.31, -0.12, 0.45, ..., 0.27] ## cream → [0.12, -0.34, 0.56, ..., -0.09]
kids → [0.08, 0.41, -0.10, ..., 0.05] eating → [-0.02, 0.22, 0.18, ..., 0.11]

2 ***Self-attention:*** Di dalam lapisan ini, ##cream membentuk bobot perhatian (attention weights) pada semua token. (Bobot yang ditunjukkan di bawah ini bersifat ilustratif dan disederhanakan untuk tujuan pemaparan.)

$$\alpha_{[CLS]} = 0.05, \alpha_{kids} = 0.10, \alpha_{like} = 0.07, \alpha_{eating} = 0.09, \alpha_{ice} = 0.30, \alpha_{\#\#cream} = 0.35$$

Output perhatian merupakan jumlah berbobot dari vektor nilai (*value vectors*):

$$att_{\#\#cream} = \sum_j \alpha_j \, x_j$$

Sebagai contoh, dua dimensi pertamanya adalah:

$$att_{\{\#\#\ cream\}[1]} = 0.05x_{\{[CLS]\}[1]} + \cdots + 0.30x_{\{ice\}[1]} + 0.35x_{\{\#\#\ cream\}[1]} \approx 0.20,$$
$$att_{\{\#\#\ cream\}[2]} = 0.05x_{\{[CLS]\}[2]} + \cdots + 0.30x_{\{ice\}[2]} + 0.35x_{\{\#\#\ cream\}[2]} \approx -0.30.$$

Dengan demikian, vektor yang telah diperbarui setelah *self-attention* adalah:

$$att_{\{\#\#\ cream\}} \rightarrow [0.20, -0.30, 0.50, \ldots, 0.10]$$

3 **Jaringan saraf *feedforward***: Selanjutnya, jaringan saraf *feedforward* (FFN) berbasis posisi—yang dilengkapi dengan bobot, bias, dan fungsi aktivasi—diterapkan pada setiap token secara independen:

$$FFN(h) = W_{\{2\}\sigma(W_{\{1\}h} + b_{\{1\}})} + b_{\{2\}}$$

Jika diterapkan pada keluaran *attention* untuk ##cream, hasilnya adalah:

$$ffn_{\{\#\#\ cream\}} = FFN\left(att_{\{\#\#\ cream\}}\right) \rightarrow [0.26, -0.18, 0.63, \ldots, 0.15]$$

4 **Keluaran Lapisan l**: Dengan adanya koneksi residual dan normalisasi lapisan, lapisan tersebut menghasilkan representasi berikutnya:

$$x^{\{(l)\}}_{\{\#\#\ cream\}} \rightarrow [0.24, -0.25, 0.58, \ldots, 0.12]$$

Interpretasi: *Self-attention* memperbarui ##cream dengan menggabungkan informasi dari token-token terkait, dan FFN melakukan penyandian ulang secara non-linear terhadap konteks gabungan tersebut. Pengulangan proses pembaruan dua tahap ini di berbagai lapisan menghasilkan vektor-vektor yang terus berkembang seperti yang ditunjukkan pada Contoh 6.3.

Pemodelan dan Evaluasi

Setelah ekstraksi fitur, tugas penilaian otomatis dilakukan menggunakan arsitektur *deep learning* ujung-ke-ujung (*end-to-end*). Arsitektur ini terdiri dari dua komponen utama: (1) *encoder* dan (2) *prediction head* (komponen prediksi). *Encoder* tersebut menggunakan model BERT yang telah dilatih sebelumnya (*pretrained*) sebagaimana dijelaskan pada sub-bagian sebelumnya, dan memproses respons yang telah di-*tokenize* melalui beberapa lapisan *transformer* dua arah (*bidirectional*). Melalui mekanisme *self-attention*, *embedding* kata awal yang sebelumnya dipelajari dari korpus umum berskala besar diberikan konteks secara dinamis. Keluaran akhir dari *encoder* ini adalah vektor berdimensi tinggi yang terkait dengan token [CLS], yang berfungsi sebagai representasi semantik agregat dan tetap (*fixed*) dari keseluruhan respons. Representasi ini kemudian diteruskan ke *prediction head* yang diimplementasikan sebagai jaringan saraf *feed-forward*; komponen ini memproyeksikan *embedding* berdimensi tinggi tersebut menjadi nilai skalar tunggal yang merepresentasikan prediksi skor esai.

Fine-Tuning

Proses pelatihan yang digunakan pada tahap ini selanjutnya disebut sebagai *fine-tuning*. Secara umum, pelatihan biasanya menyiratkan pembelajaran parameter khusus tugas dengan bobot yang diinisialisasi secara acak. Namun, *fine-tuning* dimulai dengan model yang telah memiliki pemahaman mendalam mengenai struktur bahasa hasil dari tahap pra-pelatihan (*pre-training*). Selama tahap ini, model disesuaikan dengan karakteristik spesifik tugas penilaian otomatis dan *dataset* yang digunakan dengan cara memperbarui parameternya menggunakan skor esai berlabel dari *set* pelatihan.

Prosedur Pembelajaran

Proses *fine-tuning* dijalankan melalui siklus pembelajaran iteratif, di mana BERT secara bertahap mempelajari cara memberikan skor yang selaras dengan penilaian manusia. Pada setiap iterasi, model melakukan tiga langkah: (1) *forward pass* (proses maju), (2) penghitungan *loss* (kerugian/kesalahan), serta (3) *backpropagation* (perambatan balik) dan optimasi. Pertama, selama *forward pass*, model memproses sekumpulan esai (*batch*) dan menghitung prediksi skor. Kedua, fungsi *loss* dihitung dengan membandingkan prediksi skor tersebut terhadap skor yang diberikan oleh penilai manusia. Perbandingan ini biasanya dilakukan menggunakan *mean squared error* (MSE) karena metode ini menyediakan ukuran kesalahan prediksi yang dapat diturunkan secara matematis (*differentiable*).

Terakhir, sebuah *optimizer* menggunakan sinyal kesalahan ini untuk menyesuaikan parameter model, yang mencakup bobot pada *prediction head*, mekanisme *attention* internal pada lapisan *transformer*, serta *embedding* kata awal. Dengan demikian, melalui pembaruan *word embedding*, model dapat menggeser representasi semantik kata agar lebih selaras dengan rubrik penilaian spesifik. Optimalisasi ini meminimalkan *loss* seiring berjalannya waktu, serta secara bertahap menyempurnakan kemampuan model dalam memetakan pola linguistik ke skor kualitas yang holistik.

Konfigurasi Hiperparameter

Meskipun model BERT menyediakan landasan yang kokoh, keberhasilan penerapannya untuk tugas penilaian otomatis sangat bergantung pada pemilihan hiperparameter. Ini adalah pengaturan yang ditentukan pengguna yang mengontrol bagaimana model belajar selama proses *fine-tuning* (penyesuaian halus). Berbeda dengan parameter internal yang dipelajari dari data, *hyperparameter* harus ditetapkan sebelum *fine-tuning* dimulai. Secara umum, terdapat enam *hyperparameter* yang krusial bagi kinerja model. Masing-masing dijelaskan sebagai berikut:

1. ***Batch size* (Ukuran*batch*)**: Jumlah esai peserta ujian yang diproses sebelum parameter model diperbarui. Ukuran *batch* yang lebih kecil menghasilkan pembaruan bobot yang lebih sering, sehingga membantu mencapai konvergensi lebih cepat. Sebaliknya, ukuran yang lebih besar menghasilkan pembaruan yang lebih jarang, yang berujung pada laju konvergensi yang lebih lambat.
2. ***Epoch***: Jumlah berapa kali model telah memproses keseluruhan dataset pelatihan. Dengan kata lain, satu *epoch* selesai setelah model memproses semua *batch*. Melatih model dalam beberapa *epoch* saja dapat menyebabkan *underfitting*. Sebaliknya,

melatih dalam terlalu banyak *epoch* dapat menyebabkan *overfitting*, di mana model menghafal pola spesifik dalam data pelatihan termasuk *noise* (derau) alih-alih mempelajari pola objektif yang bersifat umum.

3. ***Learning rate* (Laju pembelajaran)**: Menentukan besaran penyesuaian parameter selama pelatihan. Laju yang rendah dapat menyebabkan konvergensi yang sangat lambat, sedangkan laju yang tinggi dapat membuat proses pelatihan menjadi tidak stabil.
4. ***Dropout rate* (Tingkat *dropout*)**: Teknik regularisasi yang secara acak menetapkan keluaran dari sebagian neuron menjadi nol pada setiap langkah pelatihan untuk mencegah *overfitting*.
5. ***Optimizer***: Algoritma yang bertanggung jawab memperbarui parameter model untuk meminimalkan fungsi kerugian (*loss function*). Pilihan yang umum digunakan adalah Adam serta penerusnya, AdamW. Meskipun keduanya merupakan algoritma dengan laju pembelajaran adaptif, AdamW menyempurnakan Adam dengan menerapkan mekanisme *weight decay* (peluruhan bobot) yang lebih efektif, sehingga menawarkan kemampuan generalisasi yang lebih baik.
6. ***Activation function* (Fungsi aktivasi)**: Fungsi yang memungkinkan model mempelajari pola non-linear yang kompleks. Fungsi ini diterapkan pada keluaran setiap lapisan, mengubah nilai-nilai tersebut sebelum diteruskan ke lapisan berikutnya. Beberapa fungsi aktivasi yang paling banyak digunakan adalah *hyperbolic tangent* (Tanh) dan *Rectified Linear Unit* (ReLU)

Interaksi antar-hiperparameter ini bersifat kompleks, karena perubahan pada satu konfigurasi hiperparameter sering kali memengaruhi perilaku hiperparameter lainnya. Sebagai contoh, perubahan ukuran *batch* biasanya menuntut penyesuaian yang sepadan pada laju pembelajaran (*learning rate*). Oleh karena itu, alih-alih mengandalkan pengaturan bawaan, studi empiris sering kali menggunakan metode *grid search* untuk menentukan konfigurasi yang menghasilkan tingkat keandalan tertinggi bagi topik esai tertentu.

Evaluasi

Setelah parameter model dioptimalkan, langkah penting berikutnya adalah memvalidasi kinerja sistem menggunakan metrik yang mencerminkan kesepakatan yang bermakna dengan penilai manusia. Penting untuk membedakan tahap evaluasi ini dari tahap pelatihan yang dijelaskan pada Bagian 6.2.4.1. Selama pelatihan, model meminimalkan fungsi kerugian yang disebut *mean squared error* (MSE), yang menghasilkan gradien matematis yang halus dan diperlukan untuk penyesuaian bobot. Namun, nilai MSE sulit diinterpretasikan secara intuitif dan tidak memperhitungkan sifat ordinal dari penilaian esai. Oleh karena itu, untuk evaluasi, kami menggunakan metrik yang berbeda, yaitu *quadratic weighted kappa* (QWK). QWK adalah metrik kesepakatan yang memberikan penalti lebih berat pada ketidaksesuaian penilaian yang signifikan dibandingkan yang kecil; hal ini berbeda dengan ukuran akurasi sederhana yang menganggap semua kesalahan memiliki bobot yang sama.

QWK didefinisikan dalam Persamaan (6.1), di mana k adalah jumlah kategori penilaian yang mungkin (misalnya, untuk rentang skor 0–3, $k = 4$), dan i serta j masing-masing adalah

penilaian spesifik dari penilai manusia dan model. O adalah matriks kebingungan (*confusion matrix*), dan elemennya O_{ij} berisi jumlah item yang menerima skor i dari penilai manusia dan skor j dari model. w adalah matriks yang mendefinisikan penalti untuk ketidaksesuaian. Dalam QWK, bobotnya adalah selisih kuadrat antara penilaian: $w_{ij} = (i - j)^2$. E adalah matriks yang mewakili kesepakatan yang diharapkan secara kebetulan. Ini adalah hasil kali luar (*outer product*) dari distribusi skor kedua penilai:

$$k = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} O_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} E_{ij}} \tag{6.1}$$

Nilai QWK berkisar antara -1 hingga 1. Nilai QWK yang lebih tinggi menunjukkan kesesuaian yang lebih kuat antara prediksi dan skor sebenarnya. Skor sempurna sebesar 1 menunjukkan bahwa model sepenuhnya selaras dengan penilai manusia. Sebaliknya, nilai 0 menunjukkan kesesuaian yang buruk, di mana prediksi tersebut tidak lebih baik daripada hasil acak. Skor negatif, meskipun jarang terjadi dalam praktik, menunjukkan bahwa tingkat kesesuaian model dengan penilai manusia lebih rendah daripada yang diharapkan secara kebetulan.

Studi Pendidikan 1: Penilaian Esai Otomatis

Bagian-bagian berikut menyajikan rincian studi kasus empiris yang menunjukkan cara menerapkan alur kerja (*pipeline*) berbasis BERT untuk tugas penilaian otomatis.

Data

Studi ini menggunakan dataset *Automated Student Assessment Prize*, yang mencakup delapan butir soal esai unik, masing-masing disertai dengan kumpulan respons mahasiswa yang terkait. Untuk tujuan demonstrasi, kami berfokus pada Butir Soal 1 sebagai contoh yang digunakan secara berkelanjutan. Tabel 6.1 merangkum karakteristik dasar dari Butir Soal 1. Esai-esai tersebut bervariasi dalam hal panjang dan kualitas, mencerminkan beragamnya kemampuan menulis mahasiswa. Variabel target untuk mesin penilaian adalah skor holistik, yang berkisar antara 2 hingga 12. Kumpulan data dibagi menjadi tiga subset berbeda: 1.426 esai dalam set pelatihan, 357 dalam set validasi, dan 589 dalam set pengujian.

Tabel 6.1 Ringkasan Dataset ASAP (Set 1)

Butir	Latar Belakang Prompt	Pelatihan	Validasi	Tes	Rentang Skor
1	Tulislah sebuah esai yang menyatakan pendapat Anda mengenai dampak komputer terhadap manusia.	1.426	357	589	(2, 12)

Desain Studi

Sejalan dengan alur kerja NLP modern yang dijelaskan pada bagian sebelumnya, kami menerapkan strategi pra-pemrosesan minimal terhadap esai mentah. Penerapannya terbatas pada normalisasi spasi (*whitespace normalization*) untuk menghilangkan masalah pemformatan. Kami kemudian menggunakan tokenisasi WordPiece untuk memperoleh *word embedding* awal, guna memastikan representasi semantik esai tetap terjaga. Pencarian kisi terbatas (*constrained grid search*) dilakukan menggunakan Python (versi 3.11.4) terhadap tiga hiperparameter: (1) *learning rate*, (2) *batch size*, dan (3) *dropout rate*. Desain ini bertujuan memberikan bukti berbasis data mengenai efektivitas strategi pemodelan untuk tugas penilaian otomatis dengan cara mengisolasi pengaruh hiperparameter tersebut sembari menjaga komponen lainnya tetap konstan. Setiap hiperparameter dimanipulasi pada tiga tingkat nilai.

Learning rate divariasikan pada nilai 1×10^{-5}, 2×10^{-5}, dan 3×10^{-5}. Nilai-nilai ini dipilih karena berada di batas bawah rentang *fine-tuning* standar, sehingga meminimalkan risiko hilangnya pengetahuan yang telah dipelajari sebelumnya (*pretrained knowledge*). *Batch size* ditetapkan pada angka 8, 16, dan 32, yang memungkinkan kami mengevaluasi *trade-off* antara pembaruan yang sering namun fluktuatif (*noisy*) dari *batch* kecil dengan gradien yang stabil dan lebih halus dari *batch* besar. Terakhir, *dropout rate* divariasikan pada nilai 0,10, 0,20, dan 0,30, di mana nilai yang lebih tinggi memberikan regularisasi lebih kuat untuk mencegah *overfitting*. Penyesuaian tersebut menghasilkan 27 konfigurasi hiperparameter yang berbeda.

Untuk mengisolasi dampaknya, arsitektur model dasar, *optimizer*, dan fungsi aktivasi dijaga tetap konstan di semua kondisi eksperimen. Secara khusus, kami menggunakan DistilBERT-base-uncased sebagai arsitektur *transformer* yang telah dilatih sebelumnya (*pretrained*); arsitektur ini dipilih karena efisiensi komputasinya dan kinerja ekstraksi fitur yang tinggi dalam konteks tersebut. DistilBERT adalah model terkompresi enam lapisan yang telah dilatih sebelumnya (*pre-trained*) menggunakan BooksCorpus dan Wikipedia bahasa Inggris. Untuk optimasi, kami memilih AdamW dengan *weight deca* sebesar 0,01; ini merupakan varian modern dari Adam yang memiliki kemampuan generalisasi lebih baik. Kami juga mengimplementasikan *regression head* tiga lapisan dengan dimensi tersembunyi [512, 256, 128] dan aktivasi ReLU. Strategi *pooling* yang digunakan adalah pada token [CLS].

Setiap konfigurasi menjalani proses *fine-tuning* hingga 50 *epoch*. Rekomendasi umum sebanyak dua hingga empat *epoch* untuk model BERT didasarkan pada tolok ukur klasifikasi yang ekstensif dan stabil. Namun, AES merupakan tugas regresi yang spesifik terhadap *prompt*, dengan ukuran sampel yang relatif lebih kecil dan pengawasan (*supervision*) yang berpotensi lebih banyak mengandung *noise*. Oleh karena itu, penggunaan jumlah *epoch* yang lebih banyak diharapkan dapat memastikan bahwa konfigurasi yang menjanjikan tidak dihentikan terlalu dini akibat fluktuasi metrik jangka pendek. Akan tetapi, untuk memastikan diperolehnya model optimal sekaligus mencegah *overfitting*, kami menerapkan mekanisme *early stopping* dengan batas toleransi (*patience*) sepuluh *epoch*. Konsekuensi dari pengaturan ini adalah jika nilai QWK pada set validasi tidak mengalami peningkatan selama sepuluh *epoch* berturut-turut, proses *fine-tuning* akan dihentikan. Selama *fine-tuning*, seluruh parameter internal termasuk

word embedding diperbarui pada setiap *epoch*. Model akhir untuk setiap konfigurasi dipilih berdasarkan nilai QWK validasi tertinggi, kemudian dievaluasi menggunakan set uji (*test set*) yang telah dipisahkan sebelumnya untuk melaporkan hasil akhir.

Hasil

Hasil pencarian kisi (*grid search*) hiperparameter dirangkum dalam Tabel 6.2, yang menyoroti bahwa kinerja BERT sensitif terhadap interaksi antar-hiperparameter. Di antara 27 konfigurasi tersebut, nilai QWK validasi berkisar antara 0,78 hingga 0,86, sedangkan QWK pengujian berkisar antara 0,61 hingga 0,85. Hiperparameter menunjukkan dampak yang jelas. Konfigurasi yang menggunakan laju pembelajaran (*learning rate*) tertinggi (3×10^{-5}) yang dikombinasikan dengan tingkat *dropout* tinggi (0,30) menunjukkan ketidakstabilan yang nyata; sebagai contoh, Konfigurasi 24 menghasilkan kinerja validasi terendah dalam kelompok laju pembelajaran tinggi dengan nilai QWK sebesar 0,78. Hasil ini mengindikasikan bahwa ketika bobot model diperbarui secara terlalu agresif sembari kehilangan terlalu banyak informasi akibat *dropout*, proses *fine-tuning* mengalami kesulitan untuk mencapai solusi yang dapat digeneralisasi. Sebaliknya, kinerja paling stabil diamati pada laju pembelajaran terendah (1×10^{-5}). Untuk semua ukuran *batch* dan tingkat *dropout* pada kategori ini (ID Konfigurasi 1–9), model secara konsisten mencapai nilai QWK validasi antara 0,8388 dan 0,8574. Hasil tersebut menunjukkan bahwa strategi optimasi yang konservatif secara efektif memitigasi risiko *catastrophic forgetting* (hilangnya pengetahuan lama secara drastis), sehingga mempertahankan fitur linguistik hasil pra-pelatihan yang diperlukan untuk penilaian.

Tabel 6.2 Hasil Pencarian Grid

ID Konfigurasi	Tingkat Pembelajaran	Ukuran Batas	Dropout	QWK Validasi	QWK Uji
1	1 x 10^-5	8	0.1	0.85	0.82
2	1 x 10^-5	8	0.2	0.84	0.80
3	1 x 10^-5	8	0.3	0.85	0.72
4	1 x 10^-5	16	0.1	0.86	0.79
5	1 x 10^-5	16	0.2	0.84	0.80
6	1 x 10^-5	16	0.3	0.84	0.74
7	1 x 10^-5	32	0.1	0.85	0.83
8	1 x 10^-5	32	0.2	0.86	0.85
9	1 x 10^-5	32	0.3	0.85	0.81
10	2 x 10^-5	8	0.1	0.84	0.82
11	2 x 10^-5	8	0.2	0.85	0.69
12	2 x 10^-5	8	0.3	0.85	0.80
13	2 x 10^-5	16	0.1	0.85	0.81
14	2 x 10^-5	16	0.2	0.83	0.74
15	2 x 10^-5	16	0.3	0.81	0.65
16	2 x 10^-5	32	0.1	0.86	0.74
17	2 x 10^-5	32	0.2	0.85	0.81
18	2 x 10^-5	32	0.3	0.84	0.80

19	3 x 10^-5	8	0.1	0.84	0.84
20	3 x 10^-5	8	0.2	0.83	0.82
21	3 x 10^-5	8	0.3	0.84	0.79
22	3 x 10^-5	16	0.1	0.84	0.69
23	3 x 10^-5	16	0.2	0.84	0.81
24	3 x 10^-5	16	0.3	0.78	0.69
25	3 x 10^-5	32	0.1	0.85	0.74
26	3 x 10^-5	32	0.2	0.85	0.79
27	3 x 10^-5	32	0.3	0.84	0.61

Pemilihan model didasarkan pada nilai QWK validasi, karena bagian data validasi dikhususkan untuk penyesuaian hiperparameter, sedangkan bagian data pengujian disisihkan untuk evaluasi akhir. Konfigurasi terbaik adalah Konfigurasi 16, yang menggunakan laju pembelajaran 2×10^{-5}, ukuran *batch* 32, dan tingkat *dropout* 0,10. Model ini mencapai kinerja puncak tertinggi selama pelatihan dengan nilai QWK validasi sebesar 0,86. Saat diterapkan pada set pengujian yang disisihkan, model ini menghasilkan nilai QWK sebesar 0,74, yang mengindikasikan adanya *overfitting* terhadap sampel validasi. Perlu dicatat bahwa konfigurasi dengan nilai QWK pengujian tertinggi menggunakan pengaturan yang sedikit berbeda (laju pembelajaran 1×10^{-5}, ukuran *batch* 32, *dropout* 0,20) dan mencapai nilai QWK validasi sebesar 0,86.

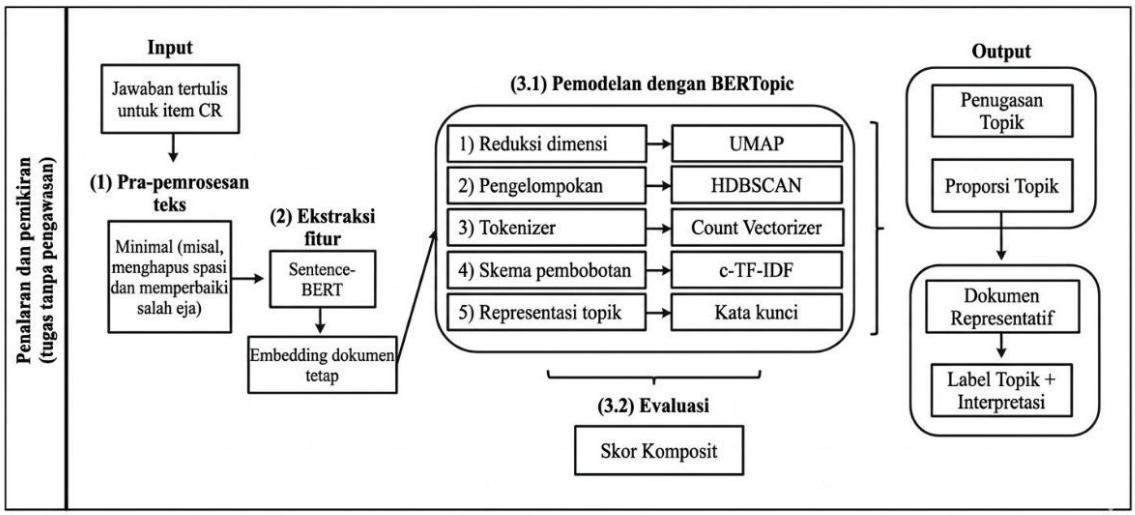
Pola ini menyoroti bahwa kinerja didorong oleh interaksi antar-hiperparameter, bukan oleh pengaturan tunggal yang berdiri sendiri, dan bahwa konfigurasi yang memaksimalkan kinerja validasi tidak menjamin kinerja terbaik pada set pengujian (*test set*). Meskipun prosesnya mengikuti protokol seleksi yang ketat, set validasi digunakan berulang kali selama pemilihan model di antara berbagai konfigurasi kandidat. Akibatnya, nilai QWK validasi maksimum dari sekumpulan konfigurasi bisa menjadi estimasi yang terlalu optimis terhadap kemampuan generalisasi yang sebenarnya, karena nilai tersebut dipilih berdasarkan pengukuran validasi yang mengandung *noise* (derau) sebuah efek yang dikenal sebagai *winner's curse* (kutukan pemenang). Untuk mengurangi sensitivitas terhadap *noise* pada validasi, beberapa pendekatan lanjutan dapat diterapkan guna memilih hiperparameter dari sekumpulan kandidat yang kuat, alih-alih hanya mengandalkan satu estimasi titik terbaik. Secara khusus, kami menyarankan pembuatan sekumpulan konfigurasi kandidat yang berada dalam rentang margin kecil dari nilai QWK validasi terbaik. Dari kumpulan ini, pilihlah konfigurasi yang memiliki nilai rata-rata QWK validasi tertinggi berdasarkan sampel berulang, seperti melalui pembagian *train*/*validation* yang diulang atau validasi silang *K-fold*.

6.3 ALUR KERJA NLP BERBASIS BERT UNTUK PENALARAN DAN PEMIKIRAN

Bagian ini menguraikan bagaimana alur kerja (*pipeline*) NLP berbasis BERT dapat diadaptasi untuk mengeksplorasi penalaran dan pemikiran peserta tes, melampaui sekadar pemberian skor pada respons mereka terhadap butir soal *Constructed-Response* (CR). Gambar 6.2 menyajikan ringkasan visual alur kerja NLP tersebut menggunakan BERTopic. Sub-bagian berikutnya menjelaskan setiap langkah dalam alur kerja ini, yang diakhiri dengan contoh

empiris untuk mengilustrasikan penerapannya dalam tugas pengukuran pendidikan ini. Alur kerja berbasis BERT untuk mengeksplorasi penalaran dan pemikiran ini diimplementasikan menggunakan kerangka kerja pemodelan topik BERTopic. Berbeda dengan penilaian otomatis berbasis BERT yang melakukan *fine-tuning* pada model untuk memprediksi skor BERTopic mengadopsi pendekatan modular dan *unsupervised* (yaitu, eksploratif) untuk mengungkap topik laten dari respons peserta tes.

Dengan demikian, alih-alih memprioritaskan akurasi prediksi, BERTopic berfokus pada identifikasi dan interpretasi tema-tema umum yang selanjutnya disebut sebagai topik yang mencerminkan cara peserta tes mengartikulasikan penalaran mereka. Karena berfokus pada fitur tingkat dokumen, respons direpresentasikan menggunakan *embedding* semantik padat (*dense semantic embeddings*), yang kemudian dikelompokkan ke dalam topik-topik dan didefinisikan menggunakan kata-kata yang paling dominan dalam setiap topik. Alur kerja ini terdiri dari tiga langkah utama: (1) pra-pemrosesan teks; (2) ekstraksi fitur melalui *embedding* dokumen; dan (3) pemodelan dan evaluasi, mencakup reduksi dimensi, klasterisasi, representasi topik, dan evaluasi model.



Gambar 6.2 Alur kerja NLP untuk mengeksplorasi penalaran dan pemikiran peserta ujian menggunakan BERTopic.

Pra-pemrosesan Teks

Pada dataset yang terdiri dari teks singkat dan spesifik secara konten, seperti respons singkat terhadap butir soal *constructed-response* (CR), pra-pemrosesan minimal sering kali lebih disukai untuk mempertahankan keragaman linguistik dan variasi halus dalam ekspresi mahasiswa. Sejalan dengan strategi pra-pemrosesan yang digunakan dalam tugas penilaian otomatis (*automated scoring*), pra-pemrosesan teks dibatasi pada normalisasi spasi, seperti menghapus spasi di awal dan akhir teks serta memperbaiki kesalahan spasi.

Ekstraksi Fitur

Untuk mengeksplorasi penalaran dan pemikiran peserta tes, BERTopic menggunakan *embedding* tingkat dokumen yang menghasilkan satu representasi vektor untuk setiap respons

peserta tes, sehingga memungkinkan respons dikelompokkan berdasarkan kemiripan semantiknya. *Embedding* ini biasanya diperoleh menggunakan Sentence-BERT, yang juga dikenal sebagai kerangka kerja *sentence-transforme*. Berbeda dengan tugas pendidikan prediktif seperti AES, Sentence-BERT menggunakan lapisan *pooling* tambahan biasanya disebut sebagai *mean pooling* yang menggabungkan *embedding* kata yang telah dikontekstualisasikan menjadi satu representasi tingkat dokumen, yang dikenal sebagai *embedding* dokumen.

Model Sentence-BERT telah dilatih sebelumnya (*pretrained*) dan disesuaikan (*fine-tuned*) pada tugas kemiripan semantik eksternal untuk memetakan kalimat yang memiliki kemiripan semantik agar berada lebih dekat satu sama lain dalam ruang *embedding*. Sentence-BERT versi awal disesuaikan menggunakan dataset eksternal seperti (1) korpus *Stanford Natural Language Inference*, sebuah koleksi berisi 570.000 pasangan kalimat bahasa Inggris yang diberi label kategori *entailment* (implikasi logis), kontradiksi, dan independensi semantik; serta (2) korpus kemiripan tekstual semantik (*semantic textual similarity*), yang memuat pasangan kalimat dengan anotasi skor dari manusia. Saat ini, tersedia banyak model yang telah dilatih sebelumnya, yang sering kali telah disesuaikan pada korpus yang lebih luas dan dioptimalkan untuk berbagai tugas, termasuk pengelompokan (*clustering*) dan pencarian semantik. Perlu dicatat bahwa proses penyesuaian (*fine-tuning*) ini terjadi sebelum tahap BERTopic dan bukan merupakan bagian dari proses pemodelan topik; oleh karena itu, *embedding* dokumen digunakan sebagai representasi yang tetap (*fixed*).

Pemodelan dan Evaluasi

Setelah jawaban peserta tes diubah menjadi *embedding* dokumen, BERTopic memulai tahap pemodelannya, yang mencakup reduksi dimensi, pengelompokan (*clustering*), tokenisasi, skema pembobotan, dan representasi topik. Karena struktur dan kualitas topik yang diekstraksi bergantung pada pilihan pemodelan serta pengaturan hiperparameter yang terkait, evaluasi dilakukan bersamaan dengan proses pemodelan untuk memastikan topik-topik tersebut dapat diinterpretasikan dan memiliki karakteristik yang berbeda satu sama lain.

Pemodelan

1 **Reduksi dimensi**: Karena model klaster mengalami kesulitan dalam menangani data berdimensi tinggi, *embedding* dokumen perlu direduksi untuk menghasilkan topik yang lebih koheren. Secara *default*, BERTopic menggunakan *uniform manifold approximation and projection* (UMAP), sebuah teknik non-linear yang dirancang untuk mempertahankan struktur data lokal maupun global. UMAP menghitung jarak antar-pasangan dalam ruang *embedding* asli dan memproyeksikan data ke ruang berdimensi lebih rendah sembari mempertahankan hubungan-hubungan tersebut melalui optimasi stokastik. Tiga hiperparameter perlu ditentukan oleh pengguna untuk memulai reduksi dimensi dengan UMAP: (1) metrik jarak, yang ditentukan dalam Python sebagai `*metric*`, yang menetapkan fungsi jarak untuk menghitung kemiripan dalam ruang berdimensi tinggi (biasanya menggunakan *cosine similarity*); (2) jumlah tetangga, yang ditentukan sebagai `*n_neighbors*`, yang mendefinisikan berapa banyak titik sampel yang akan digunakan saat melakukan pendekatan *manifold*; dan (3) jumlah

komponen, yang ditentukan sebagai `*n_components*`, yang menunjukkan dimensi *embedding* yang diharapkan setelah reduksi dimensi.

2 **Klasterisasi**: Setelah mereduksi dimensi *embedding*, BERTopic menggunakan algoritma klasterisasi untuk mengelompokkan respons yang memiliki keterkaitan semantik ke dalam topik-topik yang koheren. Proses ini biasanya dilakukan melalui *hierarchical density-based spatial clustering of applications with noise* (HDBSCAN) atau K-means, dengan HDBSCAN sebagai algoritma *default* di BERTopic. HDBSCAN adalah metode non-parametrik yang mengidentifikasi klaster dengan berbagai bentuk dan kepadatan, sembari memperlakukan titik data yang tersebar jarang (*sparse*) sebagai *noise*. Metode ini mengestimasi kepadatan lokal titik data dalam ruang *embedding* yang telah direduksi dan secara otomatis menentukan jumlah klaster berdasarkan struktur data yang mendasarinya. Keunggulan utama HDBSCAN, khususnya dibandingkan dengan model probabilistik seperti LDA, adalah kemampuannya dalam mengidentifikasi dokumen pencilan (*outlier*). Dalam Python, dokumen yang diberi label −1 menunjukkan jawaban peserta tes yang dianggap sebagai *noise* dan tidak termasuk dalam klaster topik koheren mana pun. Beberapa hiperparameter yang dapat dioptimalkan meliputi ukuran klaster minimum yang dalam Python disebut sebagai `*min_cluster_size*` yang mengatur jumlah minimum respons peserta tes untuk membentuk sebuah klaster, serta sampel minimum yang disebut sebagai `*min_samples*` yang mengatur batas jumlah pencilan (*outlier*) yang diperbolehkan. Sebaliknya, metode K-means mempartisi data ke dalam jumlah klaster yang telah ditentukan sebelumnya dengan cara meminimalkan varians di dalam klaster. Meskipun kurang fleksibel dibandingkan HDBSCAN, metode ini memungkinkan kontrol yang lebih langsung terhadap granularitas topik.

3 **Tokenizer**: Berbeda dengan AES, tokenisasi dalam BERTopic bukan merupakan bagian dari ekstraksi fitur dan tidak memengaruhi *embedding* dokumen ataupun hasil klasterisasi. Sebaliknya, tokenisasi diterapkan setelah proses klasterisasi dan berfungsi secara khusus untuk mendukung representasi topik. Secara spesifik, BERTopic menerapkan prosedur tokenisasi tradisional berbasis unigram untuk mengekstrak kata kunci dari dokumen-dokumen dalam setiap topik. Secara *default*, metode ini menggunakan *tokenizer* bawaan dari CountVectorizer, yang memisahkan kata berdasarkan spasi dan tanda baca. Meskipun prosedur ini menyerupai tokenisasi yang digunakan dalam model topik tradisional seperti LDA, penggunaannya semata-mata ditujukan untuk mengidentifikasi kata kunci yang dapat diinterpretasikan dan tidak berkontribusi pada proses pembelajaran struktur topik.

4 **Skema pembobotan**: Token yang diperoleh pada langkah sebelumnya digunakan bersama skema pembobotan c*lass-based term frequency–inverse document frequency* (c-TF–IDF) untuk menentukan tingkat kepentingan kata dalam setiap topik. Berbeda dengan TF–IDF tradisional yang menghitung tingkat kepentingan istilah pada tingkat dokumen, c-TF–IDF mengagregasi frekuensi istilah di seluruh dokumen yang termasuk dalam suatu klaster dan membandingkannya dengan frekuensi istilah tersebut di

seluruh korpus. Pendekatan ini menonjolkan istilah-istilah yang menjadi ciri khas topik tertentu sekaligus mengurangi bobot istilah umum yang muncul di berbagai topik. Silakan merujuk pada Mardones-Segovia untuk rincian lebih lanjut mengenai pendekatan TF–IDF tradisional.

5 **Representasi topik**: Representasi topik dalam BERTopic disusun dengan mengekstrak kata-kata peringkat teratas yang memiliki skor c-TF–IDF tertinggi dalam setiap topik. Kata-kata kunci ini berfungsi sebagai ringkasan padat mengenai gagasan dan proses penalaran yang mendasari setiap topik tersebut.

Setelah tahap pengelompokan (*clustering*) dan representasi topik, BERTopic menghasilkan penugasan topik, yang menunjukkan label topik bagi respons setiap peserta ujian berdasarkan keanggotaan klasternya. Ketika model dijalankan dengan pengaturan `*calculate_probabilities = True*`, model tersebut juga menghasilkan proporsi topik, yang menunjukkan distribusi probabilitas topik untuk setiap dokumen. Proporsi ini memungkinkan interpretasi keanggotaan topik secara *soft-clustering* (pengelompokan lunak), serupa dengan LDA, sehingga memungkinkan identifikasi dokumen yang paling kuat kaitannya dengan setiap topik. Dokumen-dokumen tersebut kemudian dapat digunakan untuk memandu pelabelan dan interpretasi topik oleh LLM (misalnya, GPT, Gemini). Namun, interpretasi berbasis LLM ini harus ditinjau secara cermat oleh pakar di bidang terkait untuk memastikan validitas dan akurasinya.

Evaluasi

Model topik dapat dievaluasi menggunakan dua kriteria yang saling melengkapi: (1) koherensi topik dan (2) ketidakmiripan (*dissimilarity*) topik. Koherensi topik mengukur rata-rata kemiripan semantik di antara kata kunci utama dalam setiap topik, sehingga mencerminkan sejauh mana topik tersebut konsisten secara internal dan bermakna secara semantik. Sebaliknya, ketidakmiripan topik dihitung sebagai rata-rata jarak berpasangan (*pairwise distance*) antara semua *embedding* topik, dengan jarak didefinisikan sebagai 1 − kemiripan kosinus (*cosine similarity*). Ukuran ini berfungsi sebagai proksi untuk keterpisahan topik, guna memastikan bahwa topik-topik tersebut mencerminkan aspek penalaran peserta ujian yang berbeda secara semantik, dan bukan konten yang redundan atau tumpang tindih.

Nilai ketidakmiripan berkisar antara 0 hingga 1; nilai yang lebih tinggi menunjukkan topik yang terpisah dengan lebih baik, sedangkan nilai yang lebih rendah menunjukkan tingkat tumpang tindih yang lebih besar. Koherensi berbasis *embedding* berkisar antara −1 hingga 1, di mana nilai yang lebih tinggi mencerminkan keterkaitan semantik yang lebih kuat di antara kata-kata utama dalam suatu topik. Untuk menyeimbangkan koherensi dan ketidakmiripan, kami mengembangkan skor gabungan (*composite score*). Karena metrik-metrik ini beroperasi pada skala yang berbeda, standardisasi diperlukan sebelum menggabungkannya menjadi indeks kualitas terpadu:

$$composite\ score = 0.5z\_dissimilarity + 0.5z\_coherence. \quad (6.2)$$

Karena skor-z distandardisasi di seluruh konfigurasi hiperparameter yang dievaluasi dalam *grid search*, skor gabungan ini ditujukan untuk pemilihan model dalam lingkup studi itu sendiri, bukan untuk perbandingan antar-studi. Distribusi skor dan dengan demikian makna dari nilai yang tinggi bersifat spesifik terhadap ruang pencarian yang digunakan dalam studi ini. Skor yang lebih tinggi menunjukkan konfigurasi yang menghasilkan topik yang lebih koheren (konsistensi internal) dan pemisahan topik yang lebih baik (tidak redundan), sedangkan skor yang lebih rendah mencerminkan koherensi yang berkurang, tumpang tindih antar-topik yang lebih besar, atau keduanya.

Dalam praktiknya, konfigurasi hiperparameter dengan skor gabungan tertinggi mewakili solusi topik yang paling mudah diinterpretasikan dan paling khas. Namun, untuk memastikan bahwa topik yang diekstraksi koheren secara semantik dan selaras dengan konten pendidikan dalam dataset, penilai manusia juga perlu melakukan tinjauan kualitatif terhadap kata kunci peringkat teratas dan respons peserta ujian yang representatif dalam setiap struktur topik.

Studi Pendidikan 2: Analisis Penalaran dan Pemikiran Menggunakan BERTopic

Bagian berikut ini menunjukkan bagaimana BERTopic dapat digunakan untuk menganalisis penalaran peserta ujian dalam konteks sains dan melakukan *grid search* guna mengidentifikasi konfigurasi hiperparameter yang menghasilkan struktur topik yang paling koheren dan beragam.

Data

Dataset ini mencakup respons dari 870 mahasiswa kelas lima yang menyelesaikan tes yang terdiri dari 23 butir soal pilihan ganda dan tiga butir soal respons terbuka (c*onstructed-response* atau CR; butir soal 23–25). Studi kasus ini berfokus pada Butir Soal 25, yang mencakup tiga sub-butir soal CR mengenai penggunaan listrik dalam kehidupan sehari-hari. Respons-respons tersebut singkat dan spesifik pada domain tertentu, serta mencerminkan penalaran ilmiah dan pemahaman konseptual tingkat sekolah dasar. Dataset ini sebelumnya telah dianalisis menggunakan metode NLP tradisional, dan analisis LDA menunjukkan bahwa model empat topik menghasilkan struktur yang paling mudah diinterpretasikan. Struktur topik yang dihasilkan mengekstraksi berbagai aspek penalaran mahasiswa mengenai listrik, termasuk (1) Sifat listrik dinamis versus listrik statis, (2) Menghubungkan fenomena alam dengan listrik buatan manusia, (3) Sistem distribusi listrik, dan (4) Penggunaan listrik sehari-hari.

Desain Studi

Kami melakukan pencarian kisi (*grid search*) menggunakan Python 3.11.4 dengan desain eksperimen faktorial penuh, yang secara sistematis memvariasikan strategi pra-pemrosesan teks, konfigurasi hiperparameter BERTopic, dan strategi pemodelan topik (yaitu struktur topik bebas versus tetap). Paragraf-paragraf berikut menguraikan tingkat (*level*) yang dipilih untuk setiap faktor serta memberikan dasar pemikiran bagi setiap keputusan tersebut.

Pada dataset yang terdiri dari teks dengan konten spesifik, seperti respons bidang sains yang diteliti di sini, pra-pemrosesan minimal sering kali lebih disukai. Namun, saat ini masih terbatas bukti empiris mengenai bagaimana pilihan pra-pemrosesan memengaruhi BERTopic

dalam konteks pendidikan. Berdasarkan pertimbangan tersebut, kami menguji empat strategi pra-pemrosesan yang diterapkan secara kumulatif: (1) pra-pemrosesan minimal, yang mencakup konversi huruf kecil dan normalisasi spasi; (2) perbaikan kesalahan ejaan, yang mengembangkan strategi pertama dengan mengoreksi kesalahan penulisan kata; (3) penghapusan *stopword* (kata umum yang tidak bermakna khusus), yang menggabungkan langkah-langkah sebelumnya serta menghilangkan *stopword* umum; dan (4) lematisasi, yang mencakup semua langkah sebelumnya dan menstandarisasi kata ke bentuk dasar atau bentuk kamusnya. Untuk ekstraksi fitur, kami menggunakan model bahasa pra-latih (*pretrained*) bawaan bernama all-MiniLM-L6-v2.

Model tersebut dipilih berdasarkan dua pertimbangan utama: model ini beradaptasi dengan baik pada dataset spesifik domain yang berukuran terbatas dan mendukung koherensi topik yang kuat dalam pengaturan tanpa pengawasan (*unsupervised*). Untuk mengurangi dimensi dari *embedding* kalimat yang berdimensi tinggi, kami menerapkan UMAP dengan memvariasikan dua hiperparameter utama: jumlah tetangga (*n_neighbors* = 10 atau 15) dan jumlah komponen keluaran (*n_components* = 5 atau 10). Dua strategi digunakan untuk menentukan struktur topik: (1) struktur topik bebas menggunakan HDBSCAN, yang memungkinkan jumlah topik terbentuk secara alami, dan (2) struktur topik tetap, di mana topik-topik serupa digabungkan secara *post hoc* untuk memaksakan jumlah topik yang telah ditentukan sebelumnya. Dalam pengaturan ini, parameter `*candidate_n_topics*` divariasikan dari 3 hingga 10 untuk mengamati bagaimana granularitas topik memengaruhi keterpahaman dan redundansi dalam struktur topik yang dihasilkan.

Untuk kedua strategi tersebut, parameter `*min_samples*` divariasikan pada dua nilai (5 atau 10); nilai yang lebih rendah mampu mendeteksi klaster dengan granularitas lebih halus, namun dengan konsekuensi meningkatnya *noise*. Sebaliknya, nilai yang lebih tinggi menghasilkan klaster yang lebih sedikit namun lebih kokoh. Terakhir, untuk menyusun representasi topik yang mudah dipahami, kami menggunakan c-TF–IDF guna memprioritaskan kata-kata yang sering muncul dalam suatu topik namun jarang muncul di topik lain sehingga meningkatkan distingtivitas topik.

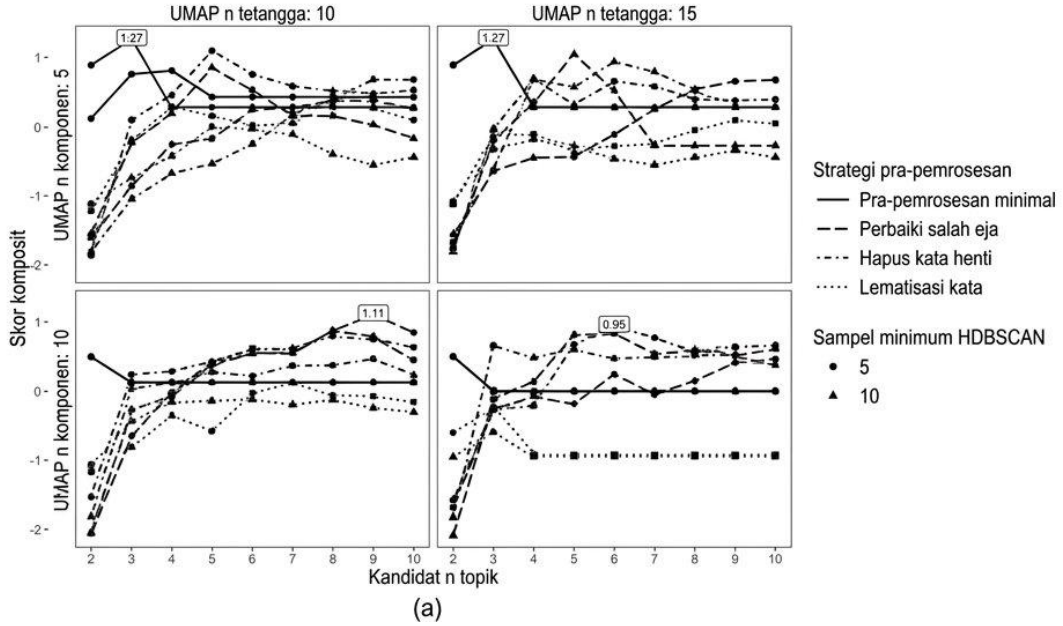
Evaluasi dan Interpretasi

Untuk setiap konfigurasi hiperparameter, kami menghitung ketidakmiripan topik (*topic dissimilarity*) dan koherensi topik (*topic coherence*), lalu memeringkat model-model tersebut menggunakan skor gabungan yang didefinisikan dalam Persamaan (6.2). Meskipun analisis kualitatif sangat penting untuk menilai keterinterpretasian topik secara menyeluruh, tujuan studi empiris ini bersifat metodologis dan ilustratif. Oleh karena itu, kami tidak menyajikan interpretasi topik secara manual, melainkan berfokus pada evaluasi kuantitatif yang diperoleh dari *grid search*.

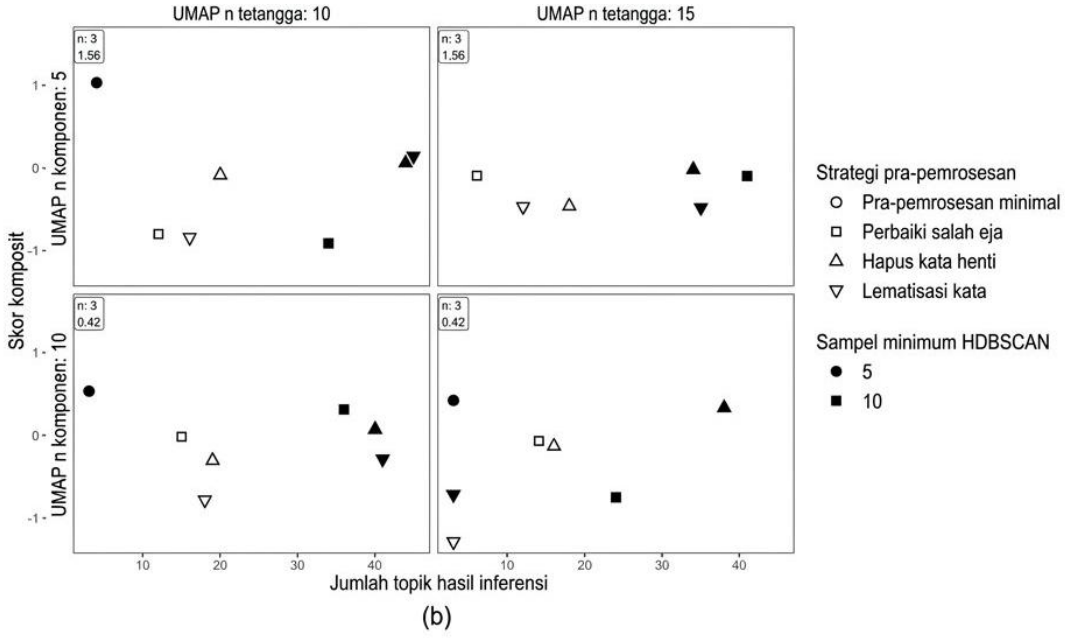
Hasil

Gambar 6.3a dan 6.3b menyajikan hasil *grid search* untuk kedua strategi pemodelan: (1) topik tetap (*fixed topic*) dan (2) topik bebas (*free topic*). Pada strategi topik tetap, skor gabungan bervariasi di berbagai konfigurasi UMAP. Secara keseluruhan, skor yang lebih tinggi diamati pada `*n_components* = 5` dibandingkan dengan `*n_components* = 10`. Untuk

`*n_components* = 5`, skor gabungan cenderung meningkat seiring bertambahnya jumlah kandidat topik dari dua menjadi empat atau lima. Namun, setelah titik tersebut, peningkatan skor gabungan menjadi stagnan, yang mengindikasikan adanya penurunan hasil tambahan (*diminishing returns*) dari penambahan topik lebih lanjut. Konfigurasi dengan skor tertinggi dalam strategi ini (skor gabungan = 1,27) diperoleh menggunakan pra-pemrosesan minimal dengan struktur tiga topik (sebelum penghapusan topik pencilan/ *outlier*), `*n_neighbors* = 10`, `*n_components* = 5`, dan `*min_samples* = 10`. Pra-pemrosesan minimal juga menghasilkan skor akhir 1,27 pada konfigurasi berbeda tetap dengan struktur tiga topik (sebelum penghapusan topik pencilan) namun dengan `*n_neighbors* = 15`, `*n_components* = 5`, dan `*min_samples* = 5`; hal ini menunjukkan ketahanan *robustness*) terhadap perubahan ukuran lingkungan (*neighborhood size*) dan kepadatan klaster saat strategi pra-pemrosesan teks dijaga tetap konstan.



(a)



Gambar 6.3 Perbandingan strategi pemodelan topik tetap dan bebas.

Catatan. Dalam strategi topik tetap, kandidat jumlah topik mungkin menyertakan topik pencilan. Skor akhir = $0{,}5_{z\text{dissimilarity}} + 0{,}5_{z\text{coherence}}$. Untuk konfigurasi yang disorot, nilai metrik mentahnya kira-kira: topik tetap: koherensi = 0,35, ketidakmiripan (*dissimilarity*) = 0,36; topik bebas: koherensi = 0,34, ketidakmiripan = 0,35.

Untuk `*n_components* = 10`, pra-pemrosesan minimal tidak menghasilkan skor gabungan tertinggi; sebaliknya, nilai-nilai tersebut cenderung stabil di sekitar angka nol pada berbagai model kandidat topik, mulai dari tiga hingga 10 topik. Pada konfigurasi ini, skor gabungan tertinggi di berbagai ukuran lingkungan UMAP dicapai ketika kesalahan ejaan diperbaiki atau *stopword* dihapus, khususnya dengan `*min_samples = 5*`. Di sisi lain, lematisasi secara konsisten menghasilkan skor gabungan terendah. Dalam strategi topik bebas (Gambar 6.3b), model secara otomatis menentukan jumlah topik berdasarkan struktur klasterisasi yang mendasarinya.

Model dengan kinerja terbaik di antara semua konfigurasi (skor gabungan = 1,56) diperoleh dengan pra-pemrosesan minimal, model tiga topik (setelah menghapus topik pencilan/*outlier*), *n_neighbors = 15*, *n_components = 5*, dan *min_samples = 5*. Pra-pemrosesan minimal juga menghasilkan skor gabungan yang sama (1,56) pada konfigurasi berbeda, juga dengan struktur tiga topik (setelah menghapus topik pencilan), namun dengan *n_neighbors = 15*, *n_components = 5*, dan *min_samples = 10*. Hasil ini menunjukkan bahwa kinerja BERTopic tetap stabil di berbagai pengaturan HDBSCAN ketika penyematan (*embedding*) UMAP ditetapkan secara tetap. Berbeda dengan strategi topik tetap, pra-pemrosesan minimal secara konsisten menghasilkan konfigurasi dengan skor tertinggi, sedangkan perbaikan kesalahan ejaan atau penghapusan kata henti (*stopword*) tidak muncul dalam daftar konfigurasi teratas.

Tabel 6.3 Glosarium Istilah NLP yang Digunakan dalam Bab Ini

Istilah	Definisi
Alur kerja NLP	Alur kerja menyeluruh (*end-to-end*) untuk analisis teks: pra-pemrosesan, ekstraksi fitur (*embedding*), pemodelan, dan evaluasi.
Token	Unit teks terkecil yang diproses oleh model (misalnya, kata, subkata, atau simbol khusus).
Tokenisasi	Prosedur yang memecah teks menjadi token; bisa berbasis kata, karakter, atau subkata.
Tokenisasi sub-kata	Strategi yang menguraikan kata-kata yang jarang muncul menjadi sub-unit yang sering muncul untuk menyeimbangkan ukuran kosakata dan cakupan.
WordPiece	Algoritma subkata berbasis data yang digunakan oleh BERT untuk membangun kosakata ringkas yang terdiri dari token dan sub-token.
Matriks dokumen-term (DTM)	Matriks dengan dokumen sebagai baris dan istilah sebagai kolom; entri matriks berupa hitungan atau bobot (misalnya, TF–IDF).
n-gram	Urutan *n* token yang berurutan (misalnya, bigram, trigram) yang digunakan sebagai fitur konteks sederhana.
Embedding	Representasi vektor bernilai riil dan padat (*dense*) dari unit linguistik (token, kalimat, dokumen).
Embedding kata (statis)	Vektor yang tidak bergantung pada konteks untuk suatu kata (misalnya, Word2Vec); vektor yang sama digunakan di semua konteks.
Embedding kontekstual	Vektor yang nilainya bergantung pada konteks di sekitarnya (misalnya, *embedding* token atau kalimat BERT).
Embedding kalimat	Vektor tunggal yang merangkum kalimat atau teks pendek (misalnya, dari Sentence-BERT).
Transformer	Arsitektur saraf yang dibangun dari tumpukan sub-lapisan *self-attention* dan *feedforward* berbasis posisi.
Pengodean posisi	Pengodean numerik yang ditambahkan ke vektor token agar model *transformer* dapat memanfaatkan urutan kata.
Self-attention	Mekanisme yang menghitung representasi bergantung-konteks dengan memberikan bobot pada hubungan antar-token melalui *query*, *key*, dan *value*.
Multi-head attention	"Kepala" (*head*) *self-attention* paralel yang menangkap pola relasional yang saling melengkapi.
Jaringan saraf *feedforward* (FFN)	Pemetaan non-linear berbasis posisi (biasanya diimplementasikan sebagai dua lapisan afin dengan fungsi aktivasi) yang diterapkan pada setiap token secara independen di dalam lapisan *transformer*.
Koneksi residual	Koneksi *skip* yang menambahkan input lapisan ke output-nya untuk menstabilkan optimasi dan mempertahankan informasi.

Normalisasi lapisan	Normalisasi yang diterapkan pada fitur tersembunyi di dalam suatu lapisan untuk meningkatkan stabilitas pelatihan.
Encoder (transformer)	Tumpukan blok *attention* dan FFN yang memetakan urutan input menjadi representasi kontekstual (digunakan oleh BERT).
Decoder (transformer)	Tumpukan dengan *masked self-attention* (dan *cross-attention* opsional) yang digunakan untuk tugas pembangkitan teks (*generation*).
BERT	*Bidirectional Encoder Representations from Transformers*; model khusus-encoder yang digunakan untuk ekstraksi fitur atau *fine-tuning*.
[CLS], [SEP], [PAD]	Token khusus masing-masing untuk representasi tingkat urutan (*sequence-level*), pemisahan segmen, dan padding.
Masker atensi	Vektor biner yang menandai token asli (1) dan padding (0) untuk membatasi perhatian (*attention*) hanya pada input yang valid.
Dimensi *embedding* pra-pelatihan	Panjang representasi vektor (misalnya, 768 untuk BERT_base). Pelatihan swa-awasi (*self-supervised*) berskala besar (misalnya, *masked language modeling*) yang mempelajari representasi linguistik umum.
Fine-tuning atau pelatihan	Adaptasi terbimbing (*supervised*) yang spesifik untuk tugas tertentu terhadap parameter model pra-latih menggunakan data berlabel.
Fungsi kerugian (*loss function*)	Fungsi objektif yang diminimalkan selama pelatihan (misalnya, *mean squared error* untuk regresi, cross-entropy untuk klasifikasi).
Optimizer	Algoritma yang memperbarui parameter menggunakan gradien (misalnya, Adam, AdamW).
Laju pembelajaran (*learning rate*)	Ukuran langkah yang mengontrol besarnya pembaruan parameter selama optimasi.
Ukuran batch	Jumlah contoh data yang diproses sebelum pembaruan parameter dilakukan.
Epoch	Satu putaran lengkap melalui seluruh set data pelatihan.
Dropout	Teknik regularisasi yang secara acak menetapkan nilai aktivasi menjadi nol selama pelatihan untuk mengurangi overfitting.
Fungsi aktivasi	Pemetaan non-linear di dalam jaringan saraf (misalnya, ReLU, GELU, Tanh).
Hiperparameter	Pengaturan yang ditentukan pengguna untuk mengontrol arsitektur model atau proses pelatihan (misalnya, learning rate, batch size).
Grid/*random search*	Prosedur sistematis (*grid*) atau stokastik (acak) untuk mengeksplorasi konfigurasi hiperparameter.
Quadratic weighted kappa (QWK)	Indeks kesepakatan untuk skor ordinal yang memberikan penalti lebih berat pada ketidaksepakatan yang lebih besar.

LSA	*Latent Semantic Analysis*; metode faktorisasi matriks yang diterapkan pada DTM untuk mengungkap struktur semantik.
LDA	Latent Dirichlet Allocation; model topik probabilistik pada DTM yang menghasilkan distribusi topik dan kata.
BERTopic	Kerangka kerja pemodelan topik yang menggunakan *embedding* kalimat, reduksi dimensi, klasterisasi, dan c-TF–IDF untuk representasi topik.
UMAP	*Uniform Manifold Approximation and Projection*; teknik reduksi dimensi non-linear yang mempertahankan struktur lokal maupun global.
HDBSCAN	Klasterisasi berbasis kepadatan yang menemukan klaster dengan kepadatan bervariasi dan melabeli titik-titik yang tersebar jarang sebagai noise.
c-TF–IDF	*Class-based* TF–IDF yang menggabungkan dokumen dalam satu klaster untuk menyoroti token yang spesifik bagi topik tersebut.
Koherensi topik	Ukuran keterkaitan semantik di antara kata-kata teratas dalam suatu topik (nilai lebih tinggi menunjukkan topik yang lebih mudah diinterpretasikan).
Ketidakmiripan topik	Pemisahan antar-representasi topik (misalnya, 1 − cosine similarity), yang menunjukkan tidak adanya redundansi.
Head attention	Salah satu mekanisme atensi paralel dalam *multi-head attention.*
Head klasifikasi	Lapisan khusus tugas yang ditempatkan di atas *embedding* (misalnya, menggunakan vektor [CLS]) untuk keperluan prediksi.
Penghentian dini	Kriteria penghentian pelatihan yang dipicu saat kinerja validasi tidak lagi meningkat, guna mencegah *overfitting*.

6.4 REFLEKSI ATAS ALUR KERJA NLP BERBASIS BERT

Bab ini menyajikan tinjauan komprehensif mengenai alur kerja (*pipeline*) NLP berbasis BERT untuk mengevaluasi jawaban tertulis peserta ujian terhadap butir soal *constructed-response* (CR). Bab ini menguraikan alur kerja umum pada Bagian 6.2 dan mengilustrasikan bagaimana setiap komponen alur kerja dapat disesuaikan melalui dua tugas pengukuran pendidikan. Istilah-istilah teknis utama yang digunakan dalam bab ini dirangkum dalam Tabel 6.3. Tugas pengukuran pertama berfokus pada AES dan mengilustrasikan bagaimana model BERT yang telah dilatih sebelumnya (*pretrained*) dapat disesuaikan untuk penilaian otomatis melalui berbagai konfigurasi hiperparameter. Dengan pra-pemrosesan minimal dan pembagian data secara bertingkat (*stratified split*), kami melakukan *grid search* terbatas pada parameter *learning rate*, *batch size*, dan*dropout*, sembari menjaga arsitektur model dan komponen pelatihan lainnya tetap konstan.

Hasil *grid search* menunjukkan bahwa kinerja penilaian DistilBERT terutama dipengaruhi oleh interaksi antar-hiperparameter, bukan oleh satu pengaturan tunggal secara terpisah. Dari 27 konfigurasi yang diuji, nilai QWK validasi berkisar antara 0,78 hingga 0,86, dan

nilai QWK pengujian berkisar antara 0,61 hingga 0,85; perilaku tidak stabil muncul ketika *learning rate* tinggi dipadukan dengan *dropout* tinggi, yang mengindikasikan adanya kesulitan dalam mencapai konvergensi menuju solusi yang dapat digeneralisasi akibat pembaruan yang terlalu agresif dan regularisasi yang berat. Perbedaan yang tidak terlalu mencolok pada nilai parameter individu masih dapat diamati. Sebagai contoh, tingkat *learning rate* terendah secara konsisten menghasilkan kinerja validasi yang kuat di berbagai pengaturan *batch size* dan *dropout*; hal ini menunjukkan bahwa efek parameter tunggal memang ada, namun paling tepat ditafsirkan dalam konteksnya. Yang penting, konfigurasi yang menghasilkan QWK validasi tertinggi tidak menghasilkan kinerja pengujian tertinggi.

Sebaliknya, nilai QWK pengujian terbaik dicapai oleh pengaturan yang berbeda namun berdekatan, yang menegaskan bahwa pilihan optimal berdasarkan validasi bisa jadi mencerminkan *overfitting* atau *noise* ketika banyak kandidat konfigurasi dibandingkan. Secara keseluruhan, hasil penyesuaian parameter (*tuning*) ini memberikan dasar praktis untuk memilih konfigurasi yang kuat bagi tugas ini. Selain itu, hal ini juga menyiratkan bahwa jika anggaran komputasi memungkinkan, mengonfirmasi pilihan di antara sejumlah kecil pengaturan berkinerja terbaik menggunakan validasi silang (*cross-validation*) dapat semakin memperkuat keyakinan terhadap kemampuan generalisasi model. Tugas pengukuran kedua mengilustrasikan penggunaan BERTopic untuk mengeksplorasi penalaran dan pola pikir dalam jawaban peserta ujian terhadap butir soal CR. *Grid search* dilakukan untuk mengidentifikasi kombinasi hiperparameter yang menghasilkan representasi topik paling koheren dan perbedaan paling nyata antar-topik. Peningkatan skor akhir yang konsisten yang dicapai dengan pra-pemrosesan minimal di berbagai strategi pencarian menyoroti keuntungan mempertahankan informasi linguistik yang lebih kaya saat menggunakan BERTopic. Hasil ini menunjukkan manfaat praktis dibandingkan dengan pedoman alur kerja NLP tradisional; dalam pendekatan tradisional, pra-pemrosesan teks yang intensif yang sering kali bersifat iteratif, memakan waktu, dan bergantung pada penilaian ahli kerap dianggap sebagai aspek krusial untuk memperoleh hasil yang dapat diinterpretasikan.

Saat membandingkan kinerja BERTopic di berbagai strategi pencarian, hasilnya menunjukkan bahwa penetapan topik secara *post hoc* tidak mengungguli skor akhir yang diperoleh dengan struktur topik bebas. Dengan demikian, hasil ini menyoroti bahwa penggunaan pendekatan otomatis untuk mengidentifikasi struktur topik terbaik dapat menghasilkan representasi topik yang lebih koheren dan beragam, terutama ketika pra-pemrosesan yang diterapkan minimal. Pola ini konsisten dengan bukti sebelumnya yang dilaporkan dalam konteks ilmu computer. Serupa dengan hasil tugas penilaian otomatis, kinerja BERTopic tampak bervariasi tergantung pada pengaturan hiperparameter, yang mengisyaratkan bahwa konfigurasi optimal mungkin bergantung pada korpus yang digunakan. Oleh karena itu, penerapan BERTopic di masa mendatang dalam konteks pendidikan mungkin memerlukan pengujian berbagai rangkaian hiperparameter untuk mengidentifikasi struktur topik yang paling baik dalam merangkum pemikiran dan penalaran peserta tes.

Dibandingkan dengan LDA tradisional yang diterapkan pada dataset ini oleh Mardones-Segovia, BERTopic mengidentifikasi lebih sedikit topik; hal ini kemungkinan disebabkan oleh

pengaruh gabungan antara klasterisasi berbasis *embedding* dan skor akhir komposit, yang secara eksplisit menyeimbangkan koherensi dan ketidakmiripan topik. Namun, perlu dicatat bahwa karena studi ini berfokus pada ilustrasi dan evaluasi metodologis, aspek keterinterpretasian dari struktur topik dengan kinerja terbaik tidak dinilai. Oleh karena itu, penelitian selanjutnya perlu membandingkan secara langsung hasil LDA dan BERTopic terkait keterinterpretasian dan koherensi topik. Singkatnya, kedua studi tersebut menunjukkan cara menggunakan teknik NLP berbasis BERT untuk tugas pengukuran pendidikan serta pentingnya memilih rangkaian hiperparameter yang tepat dalam arsitektur semacam itu. Akan tetapi, bab ini juga mengakui adanya keterbatasan. Kedua studi tersebut mengandalkan pendekatan kuantitatif untuk mengevaluasi kinerja model berbasis BERT, sehingga berpotensi mengabaikan factor-faktor lain yang memengaruhi validitas interpretasinya. Demikian pula, meskipun hasil empiris menunjukkan pola-pola penting, penelitian mendatang sebaiknya mengkaji kinerja model berbasis BERT pada berbagai konfigurasi hiperparameter, ukuran sampel, panjang teks, dan rentang skor guna memberikan panduan bagi peneliti dan praktisi agar selaras dengan standar pengukuran pendidikan.

BAB 7
AI GENERATIF DAN PEMBUATAN SOAL CEPAT

7.1 PENGANTAR PENGEMBANGAN BUTIR SOAL TRADISIONAL

Butir soal berfungsi sebagai unit analisis dasar dalam pengujian pendidikan dan psikologis, di mana setiap butir soal menyediakan elemen penyusun yang diperlukan untuk membentuk format tes akhir. Reliabilitas skor tes dan bukti validitas untuk penggunaan skor tes berkaitan langsung dengan kinerja butir soal, dan konsep penilaian "berkualitas tinggi" didasarkan pada karakteristik psikometrik butir soal tersebut (Kane, 1992, 2013; *American Educational Research Association, American Psychological Association, dan National Council on Measurement in Education*). Oleh karena itu, tidak mengherankan jika pengembangan butir soal sering kali dianggap sebagai aktivitas terpenting dalam proses pengujian. Pendekatan tradisional dalam pembuatan butir soal yang di sini didefinisikan sebagai pertanyaan pilihan ganda dengan satu jawaban terbaik (*single best option multiple-choice questions* atau MCQ) bergantung pada pakar materi pelajaran (*Subject Matter Experts* atau SME) yang menulis setiap butir soal secara individual. Penulisan butir soal merupakan proses yang bersifat iteratif: SME menggunakan penilaian dan keahlian mereka untuk menyusun draf butir soal, yang kemudian disunting, ditinjau, dan direvisi hingga memenuhi standar kualitas.

SME memainkan peran krusial dengan cara mengidentifikasi, mengorganisasi, dan mengevaluasi konten. Setiap butir soal mencerminkan konseptualisasi unik SME mengenai pengetahuan dan keterampilan; hal ini memasukkan unsur subjektivitas ke dalam proses pengembangan butir soal dan menjadikan setiap butir soal unik. Sifat individual ini menjadikan penulisan butir soal sebagai sebuah "seni" yang berlandaskan pada pengetahuan dan wawasan SME, yang sekaligus membawa risiko munculnya bias subjektif dalam MCQ. Akibat adanya subjektivitas ini, pedoman dan proses peninjauan digunakan untuk meningkatkan standardisasi hasil akhir butir soal. Pedoman penulisan butir soal digunakan untuk menjamin keadilan dan kesetaraan. Pedoman ini merangkum praktik-praktik terbaik, kesalahan umum, serta berbagai ekspektasi yang ditetapkan. Haladyna dan Downing memperkenalkan salah satu pedoman komprehensif pertama untuk soal pilihan ganda (MCQ), yaitu sebuah taksonomi yang terdiri dari 46 aturan yang disusun berdasarkan referensi otoritatif dan masukan para ahli, dengan tujuan menyatukan standar pembuatan butir soal yang jelas, adil, dan valid. Pada tahun 2002, Haladyna, Downing, dan Rodriguez menyempurnakan taksonomi tersebut menjadi pedoman ringkas berisi 31 aturan dengan memanfaatkan bukti empiris dan konsensus ahli. Pedoman yang sering dikutip ini telah memberikan pengaruh besar terhadap penulisan butir soal di bidang ilmu sosial, kedokteran, dan psikologi. Proses penyempurnaan berulang melalui peninjauan butir soal digunakan untuk mengevaluasi apakah butir soal tersebut telah memenuhi kriteria penting yang diuraikan dalam pedoman. Peninjauan biasanya dilakukan oleh komite yang terdiri dari para ahli materi pelajaran (SME) yang tidak terlibat dalam kegiatan awal penulisan butir soal. Peninjauan ini menghasilkan umpan balik

terperinci mengenai berbagai standar kualitas butir soal yang, pada gilirannya, dapat digunakan untuk merevisi dan menyempurnakan butir soal aslinya.

7.2 KETERBATASAN PENULISAN BUTIR SOAL TRADISIONAL

Pendekatan tradisional dalam penulisan butir soal memiliki dua keterbatasan. Pertama, pendekatan ini memakan waktu dan biaya karena setiap butir soal diperlakukan sebagai unit unik yang memerlukan penulisan, penyuntingan, dan peninjauan secara individual. Berbagai komponen kualitas seperti keselarasan dengan kurikulum, tingkat kesulitan, kejelasan, dan format harus dinilai untuk setiap butir soal, sering kali oleh para ahli materi pelajaran (SME) yang waktu kerjanya bernilai tinggi. Rudner memperkirakan bahwa biaya pengembangan satu butir soal operasional untuk tes berskala besar menggunakan pendekatan tradisional berkisar antara 15JT hingga 25JT. Sebagaimana dicatat oleh Drasgow:

> Kebutuhan akan sejumlah besar butir soal sulit dipenuhi karena pendekatan tradisional dalam pengembangan tes menggunakan butir soal sebagai satuan dasar. Artinya, setiap butir soal dibuat secara manual satu per satu ditulis, ditinjau, direvisi, disunting, dimasukkan ke dalam komputer, dan dikalibrasi seolah-olah belum pernah ada butir soal serupa yang dibuat sebelumnya.

Kedua, pendekatan tradisional dalam penulisan butir soal memakan biaya besar dan mungkin sulit untuk ditingkatkan skalanya. Ketika dibutuhkan 500 butir soal, maka 500 butir soal tersebut harus ditulis dan ditinjau secara individual. Produksi berskala besar tidak mempersingkat atau menyederhanakan proses untuk setiap butir soal; meskipun hasil akhirnya meningkat secara keseluruhan, prosesnya tetap padat karya karena besarnya upaya manusia yang diperlukan untuk menyusun draf, meninjau, dan merevisi butir soal.

Menuju Pembuatan Butir Soal Otomatis

Pembuatan butir soal otomatis atau *Automatic Item Generation* (AIG) adalah pendekatan alternatif dalam penulisan butir soal yang dapat menjawab kebutuhan untuk menghasilkan sejumlah besar butir soal tes baru secara efisien dan ekonomis. AIG sebagai metode pemanfaatan teknologi telah dijelaskan oleh Bormuth lebih dari 50 tahun yang lalu; metode ini belakangan kembali mendapat perhatian karena organisasi penyelenggara tes membutuhkan sejumlah besar butir soal tes berkualitas tinggi yang spesifik terhadap materi tertentu. Salah satu faktor penyebab meningkatnya permintaan ini adalah dipopulerkannya penggunaan berbagai bentuk tes untuk mengatasi kekhawatiran terkait keterbukaan materi soal (*item exposure*) dan keamanan tes. Penggunaan satu bentuk tes yang telah berlangsung lama menyebabkan materi soal mudah diketahui pihak luar karena semua peserta tes melihat rangkaian soal yang sama; hal ini memunculkan konsep pengembangan bentuk tes parallel yang berisi soal-soal berbeda namun mengukur materi dan konstruk yang sama untuk dapat diselenggarakan secara bersamaan. Karena bentuk-bentuk paralel ini digunakan secara bergantian, soal-soal pada setiap bentuk harus memiliki kesetaraan dalam hal materi dan spesifikasi statistik; konsekuensinya, diperlukan jumlah soal yang lebih banyak serta proses untuk memastikan kesepadanan karakteristik materi dan statistik.

7.3 PERKEMBANGAN AWAL DALAM PEMBUATAN SOAL OTOMATIS (AIG)

Kebutuhan akan lebih banyak soal pada awalnya memunculkan cikal bakal AIG sebagaimana yang kita kenal saat ini, yaitu penulisan soal secara sistematis. Guttman mengusulkan teori facet (*facet theory*), yang mengorganisasikan pengetahuan terkait ke dalam komponen-komponen yang disebut "facet". Pendekatan ini memungkinkan pembuatan soal melalui substitusi pengetahuan, sekaligus mengurangi subjektivitas dengan membatasi struktur soal. Berlandaskan logika tersebut, Bormuth memperkenalkan metode berbasis aturan yang menggunakan transformasi prosa dan ekspresi algoritmik. Meskipun inovatif, pendekatan Bormuth terbatas pada soal berbasis teks dan bergantung pada tata bahasa transformasional, yang kala itu belum banyak dipahami oleh para pengembang tes.

Kerangka kerja sistematis pertama yang diadopsi secara luas adalah pengembangan *item shell* atau kerangka dasar soal. Sebuah *shell* memuat garis besar struktur soal; struktur ini kemudian direplikasi untuk menghasilkan soal baru dengan cara memasukkan materi secara manual ke dalam *shell* tersebut guna menciptakan soal-soal yang serupa. Berbeda dengan metode Bormuth, penggunaan *shell* bersifat konkret, dapat diterapkan pada materi teks maupun numerik, serta praktis bagi para ahli materi pelajaran atau *Subject Matter Experts* (SME). Pendekatan ini yang terkadang disebut sebagai pembuatan templat (*templating*) atau cloning menjamin kesepadanan sifat statistik dengan cara mengontrol materi yang disubstitusikan.

Pemodelan Butir Soal Sebagai Pendekatan Modern Pertama

Dengan diperkenalkannya pemodelan butir soal sebuah pendekatan induktif untuk menghasilkan butir soal yang memanfaatkan struktur pengetahuan unit analisisnya beralih dari butir soal individual menjadi kelompok atau rangkaian butir soal. Model butir soal berbeda dari kerangka butir soal (*item shell*) karena memiliki struktur modular yang terdiri dari berbagai elemen. Sebuah model butir soal harus memuat setidaknya satu elemen; elemen sendiri merupakan komponen dalam butir soal yang mencakup konten variabel serta aturan yang menentukan bagaimana konten tersebut dapat bervariasi. Elemen dapat berupa berbagai hal, seperti bagian dari kalimat, satu kalimat utuh, istilah kunci, ataupun nilai numerik. Ketika model soal yang memiliki elemen-elemen tertentu dimanipulasi, dua jenis soal dapat dihasilkan: varian atau isomorf. Berbeda dengan klon atau soal berbasis templat, varian diharapkan memiliki karakteristik psikometris yang berbeda satu sama lain (misalnya, nilai tingkat kesulitan yang berbeda). Karena soal yang dihasilkan hanya berbeda pada elemen-elemen yang dimodifikasi sesuai definisi dalam model soal, elemen yang memengaruhi sifat psikometris tersebut dikenal sebagai *radicals*. Namun, model soal juga dapat dirancang untuk menghasilkan isomorf yang diharapkan memiliki karakteristik psikometris yang setara (misalnya, nilai tingkat kesulitan yang serupa); dalam model semacam ini, elemen yang memengaruhi sifat psikometris disebut sebagai *incidentals*. Singkatnya, model dengan elemen *radical* dapat digunakan untuk menghasilkan varian soal pada berbagai tingkat kesulitan, sedangkan model dengan elemen *incidental* dapat digunakan untuk menghasilkan soal isomorf dalam rentang tingkat kesulitan tertentu.

Pengaturan elemen *incidental* dan *radical* dalam AIG (*Automatic Item Generation*) memerlukan kerangka kerja teoretis yang dikenal sebagai teori kuat (*strong theory*) dan teori lemah (*weak theory*). Teori kuat mewakili pendekatan deterministik di mana soal dihasilkan secara presisi untuk mengukur keterampilan dan pengetahuan spesifik pada tingkat kesulitan yang dapat diprediksi. Model kognitif yang berlandaskan pada kerangka kerja teoretis menentukan bagaimana pengetahuan dan keterampilan diorganisasikan dalam suatu bidang materi, sehingga memandu pengembangan model soal untuk menghasilkan isomorf yang sifat psikometrisnya dapat diestimasi untuk setiap komponen yang dimanipulasi. Jika berhasil, pendekatan ini memperkuat hubungan inferensial antara kinerja pada soal dan interpretasi skor karena fitur-fitur model kognitif dapat memprediksi kinerja tes. Jika tidak tersedia model kognitif yang terdefinisi dengan baik, pedoman penulisan soal menjadi dasar bagi teori lemah; dalam hal ini, para ahli materi pelajaran (*Subject Matter Experts* atau SME) mengembangkan elemen *radical* berdasarkan penilaian dan pengalaman, serta mengidentifikasi fitur-fitur yang diperkirakan memengaruhi tingkat kesulitan tanpa landasan teoretis yang kuat. Meskipun kurang presisi dibandingkan teori kuat, teori lemah menawarkan solusi praktis ketika model kognitif atau struktur pengetahuan tidak tersedia.

Keterbatasan Pemodelan Butir Soal

Pemodelan butir soal menghadapi keterbatasan yang signifikan, dan hanya sedikit organisasi yang segera mengadopsinya untuk membuat butir soal ujian yang sesungguhnya. Metode ini paling berhasil pada bidang materi yang dapat diprediksi seperti matematika, namun kurang berhasil pada bidang seni bahasa. Metode seperti tata bahasa transformasional dan aturan desain butir soal diusulkan untuk analogi verbal dan pemahaman bacaan, namun hanya menghasilkan kumpulan butir soal yang terbatas dan tidak isomorfik. Bidang yang kaya akan teks menghadirkan tantangan tersendiri karena butir soal seni bahasa sering kali bergantung pada fakta-fakta yang baku, sehingga membatasi variasi sistematis.

Faktor pembatas lainnya adalah tidak adanya panduan untuk menerapkan pemodelan butir soal. Penelitian sering kali menekankan aspek psikometris namun mengabaikan isu-isu terkait materi; berfokus pada materi tanpa membahas aspek teknologi; atau mengeksplorasi teknologi tanpa mempertimbangkan aspek psikometris. Belum ada kerangka kerja komprehensif yang dapat mengoordinasikan berbagai tugas, mulai dari konseptualisasi materi hingga pembuatan butir soal operasional.

Evaluasi kualitas butir soal juga menjadi tantangan tersendiri. Penulisan butir soal secara tradisional bergantung pada pedoman yang sudah mapan, namun tidak ada standar serupa untuk model butir soal yang mampu menghasilkan ratusan butir soal. Meskipun teori yang kuat yaitu *Automatic Item Generation* (AIG) sangat ideal untuk menangkap keahlian materi, contoh penerapannya terbatas pada bidang khusus seperti penalaran spasial, rotasi mental, dan kelancaran kata. Demonstrasi pada bidang yang lebih luas seperti matematika atau kedokteran jarang ditemukan karena kurangnya model kognitif yang terdefinisi dengan baik untuk menerapkan teori yang kuat tersebut. Pada akhirnya, pemodelan butir soal menyediakan proses sistematis yang sering kali menghasilkan butir-butir soal yang terlalu

mirip satu sama lain serta tidak memiliki kerangka kerja pemersatu untuk memandu pengembangannya dari awal hingga akhir.

7.4 PEMODELAN KOGNITIF SEBAGAI PENDEKATAN MODERN KEDUA

Gierl dan Lai memperkenalkan alur kerja tiga tahap untuk menghasilkan butir soal tes yang menggabungkan keahlian manusia dengan perakitan berbasis komputer (untuk representasi diagram, lihat Gierl, Lai, & Tanygin, Gambar 1.1). Pendekatan AIG yang menggunakan pemodelan kognitif ini menyajikan representasi terstruktur mengenai bagaimana peserta tes berpikir dan memecahkan masalah, sehingga memastikan bahwa pembuatan butir soal didasarkan pada penalaran spesifik-domain.

Pada Tahap 1, para ahli materi pelajaran (SME) mengidentifikasi konten menggunakan model kognitif sebagai kerangka kerja yang mengorganisasikan pengetahuan, keterampilan, dan kemampuan yang diperlukan untuk memecahkan masalah dalam domain tertentu. Pada Tahap 2, SME membuat model butir soal yang berfungsi sebagai templat untuk tugas penilaian. Model butir soal mendefinisikan komponen-komponen yang dapat dimanipulasi, seperti pokok soal (*stem*), pilihan jawaban, dan materi pendukung seperti grafik atau tabel. Untuk soal pilihan ganda (MCQ) standar, pokok soal, pilihan jawaban yang benar, dan pengecoh (*distractor*) dapat dimanipulasi untuk memungkinkan variasi sistematis komponen butir soal berdasarkan model kognitif tersebut. Pada Tahap 3, butir soal dihasilkan menggunakan sistem berbasis komputer untuk mengelola tugas kombinatorial kompleks yang diperlukan. Program seperti IGOR menerapkan pemrograman batasan logis (*logical-constraint programming*) untuk menggabungkan konten model kognitif dengan struktur model butir soal. Sistem melakukan iterasi melalui semua kemungkinan kombinasi elemen, menerapkan batasan yang ditetapkan oleh SME untuk memastikan hasil yang bermakna, sehingga memungkinkan pembuatan ratusan atau ribuan butir soal.

Model kognitif itu sendiri mengikuti pendekatan abstraksi-dekomposisi dan terdiri dari tiga komponen: masalah dan skenario, sumber informasi, dan fitur. Masalah mendefinisikan fokus penilaian; skenario menyediakan konten spesifik; sumber informasi mengorganisasikan petunjuk yang relevan dengan pemecahan masalah; dan fitur merinci elemen yang dapat dimanipulasi serta batasan-batasannya. Batasan berfungsi sebagai aturan logis untuk memastikan perakitan butir soal yang valid. SME menyusun model-model ini dengan mendefinisikan hubungan antara skenario dan fitur untuk memandu pembuatan butir soal secara sistematis. Dengan menguraikan pengetahuan domain dalam model kognitif dan menyematkannya ke dalam model butir soal, pendekatan ini menjamin transparansi dan kemampuan untuk direplikasi. Berbeda dengan metode AIG sebelumnya yang hanya berfokus pada psikometrika atau teknologi semata, pemodelan kognitif menekankan keterlibatan ahli materi pelajaran (*Subject Matter Experts* atau SME) di sepanjang prosesnya. Para SME menentukan konten dan struktur sebelum teknologi diterapkan, sehingga memperkuat peran penting mereka dalam menghasilkan butir soal berkualitas tinggi. Integrasi antara keahlian manusia dan otomatisasi ini merupakan metode praktis dan dapat ditingkatkan skalanya (*scalable*) untuk menghasilkan butir soal yang berlandaskan teori serta validitas konten.

Manfaat dan Keterbatasan Pemodelan Kognitif

AIG berbasis model kognitif menawarkan manfaat signifikan untuk menghasilkan sejumlah besar butir soal yang beragam dan berkualitas tinggi. Pendekatan ini mudah dipelajari oleh SME serta dapat direplikasi, dan mencakup kerangka kerja validasi untuk mengevaluasi kualitas butir soal. Sebagaimana disimpulkan oleh Falcao:

> AIG adalah metode penilaian generasi baru yang menggabungkan keahlian manusia dengan algoritma komputer. Melalui tinjauan ini, kami mengumpulkan data yang memungkinkan kami menyimpulkan bahwa AIG merupakan metode pengembangan butir soal yang layak dan valid, serta mampu menghasilkan butir soal berkualitas.

AIG dengan pemodelan kognitif juga mengatasi keterbatasan metode penulisan butir soal tradisional terkait skalabilitas dan biaya. Model butir soal berfungsi sebagai templat yang dapat digunakan kembali dan disesuaikan di berbagai bidang materi, sehingga mengurangi kebutuhan untuk membuat setiap butir soal secara individual. AIG telah digunakan untuk menghasilkan ratusan butir soal dalam berbagai bidang seperti sains, matematika, kedokteran, farmasi, dan kedokteran gigi. Analisis menunjukkan bahwa butir soal AIG mengukur kinerja pada rentang kemampuan yang luas dan memiliki daya pembeda yang kuat, baik berdasarkan indeks teori tes klasik maupun teori respons butir.

Statistik butir soal untuk pilihan jawaban yang benar sebanding dengan statistik butir soal yang dikembangkan secara tradisional. AIG yang disertai pemodelan kognitif juga mendukung pemrosesan lanjutan untuk bank soal melalui pengodean konten atau metadata yang tepat, yang dapat ditetapkan pada tingkat model maupun tingkat butir soal. Metadata memfasilitasi pengelolaan bank soal, penyusunan perangkat tes, dan pelacakan pasca-pelaksanaan tes proses-proses yang bisa menjadi jauh lebih menantang ketika menangani butir soal yang dihasilkan dalam jumlah besar. Pemodelan kognitif memang menghadirkan sejumlah keterbatasan. Pertama, tidak ada definisi atau deskripsi eksplisit untuk mengarakterisasi kemiripan butir soal. Oleh karena itu, kemiripan butir soal tetap menjadi aspek penting yang harus didefinisikan oleh pengembang tes dan ditangani melalui penyempurnaan model atau tinjauan serta penyuntingan pascaproduksi.

Selain itu, butir soal yang dihasilkan tidak memiliki statistik yang telah ditetapkan sebelumnya, yang berarti hambatan berupa perlunya uji coba awal (*pretesting*) masih tetap ada. AIG berbasis teori kuat (*strong theory AIG*) bertujuan memprediksi karakteristik butir soal tanpa uji coba awal, namun terbatas pada ranah sempit seperti penalaran analogis, rotasi mental, dan kelancaran kata. AIG berbasis teori lemah (*weak theory AIG*), yang dipandu oleh penilaian ahli materi pelajaran (SME) alih-alih model formal, tetap menjadi alternatif praktis ketika kerangka kerja teoretis tidak tersedia, namun tetap memerlukan uji coba awal untuk mengonfirmasi sifat-sifat psikometriknya.

Generasi Berikutnya dalam Pengembangan Butir Soal

Proses pengembangan butir soal tradisional sebagaimana dijelaskan di sini menegaskan bahwa ahli materi pelajaran (SME) mengendalikan sebagian besar proses melalui upaya cermat yang mendukung kualitas butir soal yang tinggi dan argumen validitas yang kuat,

namun pada dasarnya memakan waktu, berbiaya tinggi, dan sulit untuk ditingkatkan skalanya. Untuk mengatasi keterbatasan ini, bidang psikometrik secara bertahap beralih ke pendekatan yang lebih otomatis, yang puncaknya terjadi pada tahun 2010-an dengan munculnya AIG berbasis templat yang menggunakan pemodelan kognitif. Pendekatan ini menghadirkan tantangan baru, seperti kebutuhan akan ukuran keunikan butir soal yang lebih presisi, peningkatan tekanan pada hambatan uji coba awal, serta terbatasnya penerapan AIG berbasis teori kuat. Seiring berkembangnya praktik penilaian, tahap selanjutnya setelah penulisan butir soal tradisional dan AIG hampir pasti akan melibatkan integrasi kemampuan kecerdasan buatan (AI). AI berkembang dengan kecepatan luar biasa dan, hanya dalam beberapa tahun, telah menunjukkan potensi yang mengesankan untuk mengatasi tantangan lama dalam pembuatan butir soal. Pada bagian selanjutnya dari bab ini, kita akan mengkaji bagaimana kemampuan AI yang sedang berkembang dapat memungkinkan kita memitigasi keterbatasan yang diketahui dalam otomatisasi pengembangan butir soal, sekaligus menyajikan potensi risiko yang spesifik terkait AI.

Kemampuan AI yang Sedang Berkembang dalam Penilaian

AI di sini didefinisikan sebagai istilah umum bagi teknologi yang memungkinkan komputer melakukan tugas-tugas yang biasanya memerlukan kognisi, penilaian, dan pengambilan keputusan manusia. Pembelajaran mesin (*machine learning* atau ML), yang merupakan bagian dari AI, menggunakan teknik statistik untuk melakukan klasifikasi dan prediksi berdasarkan kumpulan data yang besar, dengan mempertahankan variabel demi kekuatan prediksi alih-alih penjelasan teoretis. Pembelajaran mendalam (*deep learning*), yang merupakan bagian dari ML, menggunakan jaringan saraf tiruan dan kumpulan data masif untuk memungkinkan mesin berlatih dan melakukan prediksi secara mandiri. Pemrosesan bahasa alami (NLP) menjembatani bahasa manusia dan komputer, serta memungkinkan interpretasi, analisis, dan pembuatan teks atau ucapan.

Di ranah NLP-lah potensi AI untuk mendukung penilaian berkualitas tinggi pertama kali muncul, dengan fokus pada interpretasi, kategorisasi, dan analisis teks. Pada awal tahun 1990-an, pencocokan yang didukung NLP terhadap rubrik penilaian yang telah ditetapkan memungkinkan penilaian otomatis untuk jawaban ujian yang banyak memuat teks, seperti esai. Sistem seperti eRater dari ETS mengotomatiskan proses penilaian esai dengan akurasi tinggi sekaligus secara efektif mengurangi ketergantungan pada penilai manusia. Menjelang akhir tahun 2010-an, penelitian NLP dalam bidang penilaian berkembang melampaui analisis teks yang sudah ada hingga ke pembuatan teks baru melalui penggunaan korpus data pelatihan terstruktur berskala besar yang disusun oleh peneliti. Duolingo menggabungkan sumber daya bahasa publik dan kumpulan data berlabel untuk membuat konten bagi *Duolingo English Test*.

Pendekatan ini membutuhkan keahlian manusia yang khusus dan signifikan, serta investasi teknis dan sumber daya yang besar. Munculnya model bahasa besar (LLM) yang mengintegrasikan NLP dan pembelajaran mendalam (*deep learning*) membuat pembuatan teks menjadi layak dilakukan dengan sumber daya internal yang lebih sedikit (ringkasan tingkat tinggi yang bermanfaat mengenai LLM disajikan oleh Wang). NLP menyediakan kerangka kerja

linguistic seperti sintaksis, semantik, dan konteks sementara komponen pembelajaran mendalam memungkinkan LLM mempelajari hubungan dan representasi yang kompleks di antara kata, kalimat, dan dokumen. Model seperti BERT memperkenalkan arsitektur *encoder-only* untuk input dan klasifikasi teks, sedangkan model *decoder-only* seperti GPT yang memproses teks dari kiri ke kanan memungkinkan pembuatan teks prediktif.

Potensi LLM untuk mendukung pembuatan butir soal secara utuh mulai terlihat jelas seiring dengan peningkatan kemampuan model dalam aspek-aspek seperti pemahaman kontekstual. Publikasi von Davier di jurnal *Psychometrika* merupakan salah satu karya tulis pertama di bidang asesmen yang mengeksplorasi potensi model bahasa berbasis *deep learning*, dengan memanfaatkan basis data publik yang berisi butir-butir soal inventori kepribadian. Studi tersebut menyimpulkan bahwa pembuatan butir soal secara utuh dapat dilakukan, namun juga mencatat keterbatasan kegunaan akibat ukuran set pelatihan yang kecil.

Akan tetapi, besarnya kebutuhan teknis dan sumber daya untuk eksplorasi semacam itu membatasi penggunaan awal NLP pada analisis butir soal yang sudah ada, yang dianggap lebih layak untuk dilakukan. NLP dinilai sebagai solusi potensial untuk tugas-tugas krusial berskala besar dalam tinjauan dan pengodean bank soal seperti pencocokan butir soal demi menjaga keamanan tes di mana penggunaan perangkat lunak pendeteksi plagiarisme eksternal tidak memungkinkan untuk diterapkan. Sebuah LLM sumber terbuka (*open-source*) khusus yang dilatih menggunakan teks ilmiah, yaitu SciBERT, digunakan bersama algoritma klasterisasi untuk mengidentifikasi pola tersembunyi dalam kumpulan butir soal yang berpotensi mengindikasikan penggunaan bahasa yang stereotipikal atau bias pada butir soal yang sudah ada.

Meskipun menjanjikan pada awalnya, penggunaan NLP untuk mentransformasi pembuatan konten melalui penghasilan butir soal yang lengkap, akurat, dan berkualitas tinggi dalam skala besar sebagian besar masih terbatas pada ranah penelitian hingga akhir tahun 2010-an. Secara umum, para ahli psikometri belum memiliki pemahaman mendalam mengenai NLP. Akses terhadap jenis dan jumlah data yang diperlukan untuk menyempurnakan keluaran LLM juga masih terbatas di dalam organisasi. Selain itu, ketergantungan pada pakar materi pelajaran (*Subject Matter Experts* atau SME) menuntut adanya proses pembuatan butir soal yang dapat dijelaskan dan intuitif, sementara metodologi NLP saat itu tidak transparan maupun mudah dijelaskan kepada pihak di luar bidang tersebut. Tanpa sepengetahuan sebagian besar praktisi di bidang asesmen kala itu, sebuah aplikasi berbasis LLM sedang menanti untuk diluncurkan; aplikasi ini mampu menjawab aspek aksesibilitas dan transparansi dengan begitu baik sehingga memberikan dampak transformatif, tidak hanya bagi dunia teknologi pendidikan, tetapi juga bagi seluruh dunia.

7.5 TITIK BALIK ADOPSI AI DALAM PEMBUATAN BUTIR SOAL

Meskipun kerangka pemikiran kami mungkin menyederhanakan evolusi yang kompleks, kami berpendapat bahwa membagi lanskap teknologi pendidikan menjadi era "pra-ChatGPT" dan "pasca-ChatGPT" dengan Desember 2022 (saat OpenAI merilis ChatGPT)

sebagai titik baliknya tetap bermanfaat untuk memahami dampak AI terhadap pengembangan butir soal. Sebelum adanya ChatGPT, akses terhadap LLM (*Large Language Models*) sebagian besar terbatas pada individu dengan keahlian teknis dan matematika tingkat lanjut, sehingga membatasi penerapan praktisnya. Kehadiran ChatGPT, dengan antarmuka yang intuitif dan bersifat percakapan, mengubah dinamika ini secara mendasar. Dalam waktu singkat, masyarakat umum mulai antusias menjajal alat canggih ini yang mampu menghasilkan teks yang koheren dan relevan secara kontekstual untuk hampir semua topik. Terobosan ini juga memicu kesadaran luas mengenai potensi transformatif sistem-sistem tersebut bagi pengembangan butir soal.

Di dunia pendidikan, topik penelitian pertama dan yang paling menonjol adalah, "Seberapa baik ChatGPT dapat menjawab serangkaian butir soal tes tertentu?" Sebagai contoh, di bidang lisensi medis, sejumlah penelitian menguji kinerja ChatGPT pada ujian lisensi seperti *United States Medical Licensing Examination* (USMLE). Sebuah publikasi dari NBME mengenai topik ini mengevaluasi kinerja ChatGPT menggunakan kumpulan lengkap butir soal USMLE® yang sudah tidak digunakan lagi (*retired items*) untuk ketiga tahap ujian tersebut. Para penulis menyimpulkan bahwa kinerja aplikasi tersebut jauh di bawah ekspektasi untuk peserta ujian USMLE pada umumnya dan berada di bawah perkiraan nilai ambang batas kelulusan untuk dua dari tiga ujian tersebut. Penelitian mengenai kemampuan ChatGPT dalam menjawab soal latihan ujian lisensi pengacara (*bar exam*) tingkat negara bagian menunjukkan tren serupa: kinerja aplikasi jauh melampaui hasil tebakan acak, namun belum mencapai tingkat kelulusan yang diperkirakan untuk ujian-ujian tersebut.

Banyak penelitian yang mengevaluasi kinerja ChatGPT pada berbagai ujian penerimaan dan lisensi bermunculan dengan cepat; namun, sebagian besar memiliki relevansi yang terbatas karena bergantung pada soal latihan publik yang terlalu sederhana (misalnya, hanya berbasis teks) dan tidak setara dengan butir soal yang digunakan dalam ujian yang sesungguhnya. Terlepas dari keterbatasan-keterbatasan tersebut, kesimpulan dari studi awal mengenai ChatGPT (di mana versi spesifik GPT yang digunakan bervariasi tergantung pada waktu analisis data) menunjukkan hasil yang menjanjikan; seiring dirilisnya setiap versi terbaru GPT, kinerjanya dalam menjawab butir soal yang ada pun semakin membaik. Namun, para peneliti juga segera menyoroti berbagai potensi keterbatasannya. Sebuah ringkasan mengenai kinerja ChatGPT pada beragam butir soal selama tiga bulan pertama peluncurannya mencatat bahwa sistem ini sering kali memberikan jawaban yang terdengar masuk akal namun sepenuhnya keliru sebuah fenomena yang dikenal sebagai "halusinasi." Ketergantungan aplikasi ini pada pola statistik bahasa yang sudah ada berarti tidak adanya penerapan penalaran mendalam ataupun logika khusus bidang ilmu, sehingga kinerjanya menjadi tidak konsisten di berbagai bidang.

Tidak adanya transparansi dalam penalaran atau dasar pemikiran di balik jawaban yang dihasilkan membatasi kegunaan ChatGPT sebagai sarana pembelajaran formatif; hal serupa juga terjadi akibat adanya bias bahasa Inggris dalam data pelatihannya. Mengingat banyak butir soal yang digunakan untuk menguji kinerja ChatGPT berasal dari materi latihan yang tersedia untuk umum, ada kemungkinan soal-soal tersebut telah tercakup dalam data

pelatihan, sehingga memudahkan ChatGPT untuk menjawabnya. Para peneliti sering kali mencatat bahwa kemampuan ChatGPT dalam menjawab soal pilihan ganda (MCQ) tidak serta-merta memengaruhi validitas penilaian yang berbasis pada jenis soal tersebut. Kebutuhan bagi manusia untuk menunjukkan penguasaan materi tes tetap ada, dan argumen yang kuat untuk menggunakan soal pilihan ganda sebagai alat ukur penguasaan pengetahuan tersebut pun tetap berlaku.

Menggunakan AI untuk Membuat Soal Pilihan Ganda (MCQ)

Terlepas dari keterbatasan yang mulai muncul, para peneliti dengan cepat beralih ke pertanyaan logis berikutnya mengenai potensi ChatGPT: "Bisakah ChatGPT digunakan untuk menghasilkan butir soal yang akurat dan berkualitas tinggi secara andal?" Terkait pertanyaan ini, bidang asesmen telah menyaksikan lonjakan volume penelitian yang dipublikasikan sejak tahun 2022, terutama karena munculnya berbagai aplikasi berbasis LLM dengan antarmuka pengguna yang menarik. Industri pengujian bahkan telah menerbitkan edisi khusus yang membahas potensi AI generatif [misalnya, *Journal of Educational Measurement*, volume 59(3)], serta pedoman tingkat tinggi untuk menangani aspek logistik dan keputusan bisnis yang diperlukan guna mengintegrasikan AI generatif ke dalam berbagai proses pengembangan tes. Oleh karena itu, narasi perkembangan kronologis sebagaimana digunakan dalam bab ini sejauh ini mungkin tidak lagi relevan.

Sebagai gantinya, kami mengalihkan fokus pada wawasan dan rekomendasi terkini mengenai penggunaan model AI generatif dari perspektif yang dapat diterapkan pada berbagai LLM. Rekomendasi kami dikelompokkan ke dalam bidang-bidang krusial bagi pengembang tes, psikometrian, ahli materi pelajaran, dan pihak lain yang bertugas mengambil keputusan mengenai penggunaan terbaik model-model tersebut: memilih LLM, menyusun *prompt* (instruksi) yang efektif, menggunakan pembuatan kode dan rasionalisasi otomatis, serta meninjau butir soal yang dihasilkan AI. Rekomendasi ini dimaksudkan agar dapat diterapkan baik saat organisasi membangun solusi sendiri maupun saat bekerja sama dengan vendor.

Strategi Pemanfaatan LLM Dalam Pembuatan Butir Soal

Langkah awal dalam memanfaatkan LLM untuk pembuatan soal pilihan ganda (MCQ) adalah membandingkan dan memilih model yang paling sesuai dengan kebutuhan pengujian spesifik organisasi. Proses ini dimulai dengan menetapkan persyaratan yang jelas, seperti kebutuhan akan pengetahuan spesifik bidang ilmu, peninjauan dokumentasi mengenai akurasi faktual yang tinggi atau tingkat halusinasi yang rendah pada LLM tersebut, serta kemampuan yang terbukti dalam mengikuti *prompt* yang terstruktur. Setelah kriteria ini ditetapkan, pengembang tes sebaiknya mengevaluasi model-model terkemuka (misalnya, GPT-5, Claude, Gemini) dengan melakukan uji coba skala kecil yang menyimulasikan tugas pengembangan butir soal. Uji coba ini membantu menilai kekuatan dan keterbatasan masing-masing model.

Selain akurasi dan efisiensi, aspek keadilan dan bias harus menjadi pertimbangan utama dalam pemilihan model. LLM publik dilatih menggunakan kumpulan data masif yang sering kali mencakup sumber internet yang tidak terkontrol; hal ini dapat memasukkan bias sosial terkait ras, gender, status sosial-ekonomi, dan karakteristik yang dilindungi lainnya. Berbagai studi menunjukkan bahwa model yang berbeda memiliki tingkat bias yang bervariasi,

dengan beberapa di antaranya menunjukkan preferensi sistematis dalam konteks pendidikan. Oleh karena itu, organisasi sebaiknya meninjau dokumentasi yang tersedia mengenai strategi mitigasi bias untuk LLM tersebut serta memeriksa audit independen atau tolok ukur yang mengukur aspek keadilan.

Setelah mengumpulkan data kinerja dan keadilan, proses seleksi harus berfokus pada pemilihan model yang paling mampu menyeimbangkan kualitas, efisiensi, skalabilitas, dan kesetaraan. Hal ini mencakup evaluasi terhadap aspek-aspek seperti seberapa baik model menangani *prompt* pendidikan yang bernuansa tanpa memunculkan stereotip, kemampuannya menghasilkan pilihan pengecoh (*distractor*) yang masuk akal tanpa bias budaya, serta responsivitasnya terhadap penyempurnaan secara iteratif. Organisasi juga perlu mempertimbangkan apakah model tersebut mendukung fitur transparansi seperti alat untuk menjelaskan cara kerja model (*explainability tools*) atau jejak audit yang sangat penting untuk menjaga kepercayaan dalam lingkungan pengujian yang memiliki konsekuensi tinggi (*high-stakes testing*).

Terakhir, LLM yang dipilih harus selaras dengan batasan organisasi, khususnya terkait prinsip keamanan data dan AI yang bertanggung jawab. Pengembang tes harus mendokumentasikan tingkat keamanan yang diinginkan untuk konten yang dihasilkan, termasuk bagaimana model menangani data masukan, apakah *prompt* dan keluaran disimpan atau dicatat dalam log, serta apakah ada bagian dari proses pembuatan konten yang melibatkan peninjauan oleh manusia dari pihak vendor. Untuk penilaian berisiko tinggi, organisasi sebaiknya memprioritaskan struktur di mana LLM beroperasi dalam lingkungan aman tingkat perusahaan (*enterprise-grade*) yang menjamin isolasi data serta tidak menyimpan masukan atau keluaran pengguna. Jika bekerja sama dengan vendor, organisasi harus mensyaratkan dokumentasi yang jelas mengenai kebijakan penanganan data, termasuk apakah konten digunakan untuk pelatihan model atau jaminan kualitas, serta apakah vendor akan memiliki akses terhadap materi yang dihasilkan.

Membuat *Prompt* yang Efektif untuk Pembuatan Soal

Rekayasa *prompt* yang efektif sangat penting untuk memandu LLM dalam menghasilkan soal pilihan ganda (MCQ) berkualitas tinggi. Namun, berkomunikasi dengan LLM tidak sama dengan melatih atau mengevaluasi penulis soal manusia, dan menentukan tingkat kompleksitas yang tepat untuk sebuah *prompt* ternyata lebih menantang daripada yang terlihat pada awalnya. Oleh karena itu, pengembang perlu meluangkan waktu untuk memahami praktik terbaik dalam pemodelan *prompt* dan harus merencanakan pengembangan *prompt* sebagai proses yang berulang dan memakan waktu. Rekayasa *prompt* yang efektif bukanlah tugas sekali jalan, melainkan proses berkelanjutan yang berkembang seiring dengan adanya masukan dari peninjau dan data kinerja. Praktik terbaik untuk rekayasa *prompt* dalam pembuatan soal pilihan ganda terus berkembang, namun beberapa rekomendasi telah muncul di bidang-bidang berikut.

- Kejelasan dan spesifisitas – Nyatakan dengan jelas tugas, batasan, dan format keluaran yang diinginkan.

- Penjangkaran kontekstual. Sediakan materi referensi latar belakang yang relevan—yang sudah dimiliki dan digunakan oleh organisasi Anda—untuk memandu respons. Untuk pembuatan soal pilihan ganda, hal ini dapat mencakup cetak biru ujian (*exam blueprints*), panduan penulisan soal atau panduan gaya penulisan khusus organisasi, serta kurikulum yang berlaku.
- Penetapan peran. Berikan peran tertentu kepada model (misalnya, "Anda adalah praktisi bersertifikat di bidang materi ini") untuk membentuk perspektifnya.
- Contoh dan pembelajaran *few-shot*. Sertakan contoh soal berkualitas tinggi untuk menetapkan ekspektasi mengenai nada, kompleksitas, dan struktur. Contoh soal berkualitas rendah juga dapat disertakan, meskipun biasanya lebih efektif untuk menginstruksikan LLM tentang apa yang harus dilakukan daripada apa yang tidak boleh dilakukan.
- Penyempurnaan berulang. Gunakan pemberian *prompt* bertahap atau siklus umpan balik untuk meningkatkan akurasi dan konsistensi.
- Pengendalian bias dan keadilan. Instruksikan model secara eksplisit untuk menghindari stereotip atau bahasa yang bias.

Contoh penerapan awal pembuatan soal berbasis AI dapat ditemukan di beberapa bidang yang telah memiliki basis pengetahuan publik yang relevan secara daring serta penekanan pada soal-soal yang menguji daya ingat (*recall*), yang membuat penulisan *prompt* menjadi relatif mudah; contohnya adalah bidang hukum (lihat Deroy, Ghosh, & Ghosh, untuk contoh-contohnya). Instruksi (*prompting*) yang lebih ekstensif diperlukan untuk soal pilihan ganda (MCQ) yang tidak semata-mata mengandalkan ingatan faktual, terutama bagi soal yang menuntut peserta tes untuk menyintesis informasi dalam butir soal serta melibatkan beragam sumber pengetahuan yang mungkin menggunakan terminologi khusus bidang tertentu dan/atau tidak dapat diakses oleh masyarakat umum.

Dua tinjauan literatur mengenai pembuatan MCQ di bidang medis mendokumentasikan berbagai kemungkinan instruksi, di mana studi kedua menyajikan 19 contoh instruksi yang diambil dari berbagai publikasi. Tinjauan terhadap contoh-contoh tersebut mengindikasikan bahwa instruksi yang kompleks dapat mencakup hal-hal berikut: gaya butir soal ujian tertentu; nada atau tingkat kesulitan yang diinginkan; sumber referensi publik atau yang telah disediakan; tingkat pendidikan peserta tes; tujuan pembelajaran yang rinci; identitas penulis soal; skenario klinis dan diagnosis yang kompleks; serta format khusus untuk pokok soal (*stem*) dan pilihan jawaban. Struktur yang sudah ada untuk memandu pengembangan soal tidak perlu dikesampingkan, karena elemen-elemen tersebut dapat diintegrasikan ke dalam instruksi pembuatan soal. Program yang menggunakan pendekatan desain berbasis bukti (*evidence-centered design*) dapat mengalihfungsikan elemen-elemen terkait apa yang harus diketahui dan mampu dilakukan oleh peserta tes menjadi instruksi pembuatan soal. Instruksi juga dapat dirancang dengan mempertimbangkan struktur model kognitif melalui penyisipan komponen seperti masalah/skenario, sumber informasi, dan fitur-fitur tertentu ke dalam perintah yang diberikan. Pendekatan ini memastikan bahwa soal yang

dihasilkan selaras dengan struktur pengetahuan dan tingkat kesulitan yang ditargetkan, sehingga memperkuat validitas serta mengurangi variasi dalam konten soal.

Pengujian dan penyempurnaan instruksi secara berulang (*iteratif*) merupakan langkah krusial. Bahkan pendekatan *few-shot* yang sederhana pun dapat meningkatkan kualitas keluaran secara signifikan. Isley mendokumentasikan proses iteratif tiga tahap: instruksi awal yang menyertakan materi kursus dan contoh soal, pelabelan soal hasil buatan LLM sebagai kategori "baik" atau "buruk", serta instruksi akhir yang memasukkan label-label tersebut untuk menghasilkan soal yang lebih baik. Temuan mereka menunjukkan tingkat kesulitan soal yang sedikit lebih rendah namun daya pembeda (*discrimination*) yang lebih tinggi, dengan tingkat reliabilitas yang dapat diterima di berbagai tingkat kemampuan mahasiswa.

7.6 PENGODEAN OTOMATIS DAN PEMBUATAN PENJELASAN (*RATIONALE*)

Kegunaan LLM untuk pembuatan soal pilihan ganda (MCQ) melampaui sekadar pembuatan butir soal itu sendiri. Butir soal memerlukan metadata seperti bidang materi, format soal, dan tingkat kognitif untuk memastikan keselarasan dengan cetak biru tes. Secara tradisional, proses pengodean ini dilakukan secara manual oleh para ahli, sehingga sering kali menimbulkan hambatan yang tak terelakkan. Teknologi NLP terbukti efektif dalam mengodekan butir soal yang sudah ada; langkah logis berikutnya adalah mengintegrasikan pengodean otomatis ke dalam *prompt* (instruksi) selama atau setelah proses pembuatan soal. LLM mampu menganalisis fitur linguistik dan konseptual untuk memprediksi metadata seperti tingkat kognitif, domain materi, dan klasifikasi keterampilan.

Penggunaan materi referensi dalam *prompt* dapat meningkatkan akurasi, dan keberagaman perspektif pengodean (misalnya, perbandingan antara materi buku teks dan pandangan ahli secara kolaboratif) juga dapat disertakan. Ketika dipadukan dengan pengklasifikasi *machine learning* (ML) yang dilatih menggunakan data historis butir soal, LLM dapat menyempurnakan prediksi seiring berjalannya waktu, sehingga meningkatkan konsistensi di seluruh kumpulan soal yang besar. Butir soal yang digunakan dalam penilaian praktik biasanya menyertakan penjelasan (*rationale*) untuk pilihan jawaban yang benar maupun yang salah. Penjelasan yang efektif akan merujuk pada bukti tekstual, menjelaskan alasan pendukung jawaban yang benar, serta menguraikan mengapa pilihan pengecoh (*distractor*) dianggap salah. Penjelasan juga mendukung pengendalian mutu dengan memastikan keselarasan terhadap kompetensi yang ditargetkan serta memastikan bahwa pilihan pengecoh tampak masuk akal. Namun, penulisan penjelasan secara tradisional merupakan proses yang tidak efisien, terutama jika dilakukan oleh ahli materi pelajaran (SME) secara terpisah dari proses penulisan soal. LLM menawarkan solusi yang dapat ditingkatkan skalanya (*scalable*) untuk pembuatan penjelasan ini.

Dalam sistem bimbingan belajar cerdas (*intelligent tutoring systems*), kerangka kerja seperti yang dikembangkan oleh Hang menggabungkan GPT-4 dengan teknik *retrieval-augmented generation* (RAG) untuk memanfaatkan penjelasan yang ditulis oleh instruktur dan mahasiswa, serta menerapkan rekayasa *prompt* secara iteratif. Penilai manusia memberikan skor sempurna untuk aspek kelancaran tata bahasa dan relevansi, meskipun aspek

keberagaman dan kompleksitas dinilai lebih rendah. Bahkan pendekatan yang paling sederhana di mana pembuatan teks diminta tanpa adanya contoh konkret (dikenal sebagai pendekatan *zero-shot*) menghasilkan hasil yang menjanjikan. Ch'en menggunakan GPT-4 dengan *prompt* sederhana untuk sepuluh soal pilihan ganda berbasis skenario klinis (*clinical vignette*): "Jawablah soal pilihan ganda dengan satu jawaban terbaik berikut ini dan jelaskan mengapa setiap pilihan jawaban dianggap benar atau salah?" Direktur kursus menilai 93,8% alasan (rasional) yang diberikan sudah tepat dan 77,5% akurat, dengan tingkat akurasi tertinggi dicapai pada percobaan pertama.

Meninjau Soal Pilihan Ganda (MCQ) yang Dihasilkan

Ahli materi pelajaran (SME) dapat memasukkan bias dan kesalahan ke dalam butir soal; hal ini biasanya diatasi melalui pelatihan dan panduan penulisan soal yang menyediakan aturan eksplisit serta templat terstandarisasi. MCQ yang dihasilkan oleh AI, meskipun efisien, tidak menghilangkan risiko tersebut dan justru dapat memunculkan tantangan baru yang lebih sulit diidentifikasi dan diperbaiki. LLM publik dilatih menggunakan kumpulan data yang sangat besar dan tidak terkontrol, sehingga wajar untuk berasumsi bahwa bias dari data pelatihan dapat muncul dalam soal yang dihasilkan. Soal mungkin tampak valid namun mengandung ketidakakuratan, mulai dari kesalahan kecil hingga halusinasi yang nyata. Sebuah tinjauan sistematis dalam pendidikan kedokteran mencatat adanya masalah seperti jawaban benar lebih dari satu, soal yang terlalu sederhana, dan penggunaan bahasa yang bias. Kualitas *prompt* (instruksi) memengaruhi kompleksitas dan akurasi soal, yang menegaskan perlunya pengendalian kualitas di sepanjang proses, bukan hanya pada tahap peninjauan akhir.

Prompt sebaiknya merinci gaya, format, dan cakupan materi, serta penggunaan *prompt* secara berulang (*iterative prompting*) dapat mencakup pemeriksaan akurasi dan kesesuaian dengan tingkat pendidikan. AI menghadapi tantangan serupa dengan penulis manusia dalam menciptakan pengecoh (*distractor*) MCQ yang masuk akal, yang sangat penting bagi tingkat kesulitan dan daya pembeda soal. Secara historis, pembuatan pengecoh merupakan salah satu aspek tersulit dalam penulisan soal, dan upaya awal pembuatan pengecoh menggunakan LLM lokal memberikan hasil yang kurang memuaskan, di mana 50% pengecoh dinilai tidak masuk akal.

Namun, seiring berkembangnya model berbasis GPT, kegunaannya dalam menghasilkan pengecoh yang lebih baik pun turut meningkat. Untuk soal sejarah, ChatGPT mencapai tingkat kualitas pengecoh mendekati 60% bahkan dalam skenario *zero-shot*, meskipun kinerjanya lebih rendah untuk soal bahasa karena kesulitan dalam menghasilkan respons yang secara tata bahasa salah. Solusi potensial untuk meningkatkan kualitas pengecoh meliputi penggunaan soal yang hanya memuat jawaban benar sebagai *prompt* serta membandingkan pengecoh yang dihasilkan dengan pengecoh aktual untuk tujuan penjaminan mutu.

Menggunakan AI untuk Mengatasi Keterbatasan Pembuatan Item Secara Massal

Metode penulisan item tradisional memiliki keterbatasan terkait skalabilitas dan biaya. Solusi seperti AIG dapat menghasilkan item yang sangat mirip satu sama lain atau justru menimbulkan hambatan (*bottleneck*) yang lebih parah pada tahap pra-pengujian. Solusi AI

untuk penulisan item dapat mempercepat proses dengan biaya lebih rendah; namun, baik AIG maupun AI berpotensi menghasilkan banyak item yang sangat mirip atau tidak dapat digunakan jika tidak tersedia kapasitas tambahan untuk pra-pengujian. Untungnya, metodologi NLP yang menjadi landasan LLM dapat digunakan untuk mengatasi kedua keterbatasan tersebut. Di sini, kami menyajikan informasi bagi pengembang tes yang ingin mengeksplorasi penggunaan NLP dan LLM untuk mengukur kemiripan tekstual pada item yang dihasilkan atau meringankan beban pra-pengujian melalui prediksi tingkat kesulitan item yang didukung oleh NLP.

Mengukur Kemiripan Tekstual Item

Praktik terbaik untuk pengodean item dan pengembangan tes yang akurat mencakup tinjauan sistematis oleh pakar materi pelajaran (SME). Tujuannya adalah memastikan bahwa setiap pasangan item dengan tingkat kemiripan tekstual yang tinggi dikodekan sebagai "item musuh" (*enemy items*) dan tidak digunakan dalam formulir tes yang sama. Pasangan item musuh dapat diidentifikasi dan ditandai sebelum penyusunan formulir tes, atau ditandai setelah penyusunan melalui tinjauan formulir. Proses tinjauan ini memakan banyak waktu dan tenaga bahkan pada metode penulisan item tradisional, dan menjadi hampir mustahil dilakukan ketika item dihasilkan dalam jumlah besar. Penelitian mengenai kesesuaian metode NLP untuk identifikasi otomatis kemiripan item atau item musuh mulai muncul tepat sebelum peluncuran model berbasis *transformer* dan menunjukkan hasil yang sangat menjanjikan.

Metodologi klasterisasi NLP, yang diterapkan pada bank item hasil AIG, terbukti efektif mengelompokkan item dengan tumpang tindih tekstual tertinggi secara berdekatan, sehingga memungkinkan peninjau manusia untuk memeriksa potensi item musuh dengan lebih cepat. Studi-studi lain telah menguraikan bagaimana teknik NLP leksikal yang menggabungkan indeks kemiripan kosinus dengan model pembelajaran mesin (ML) dapat secara efisien mengukur tumpang tindih teks antar-butir soal, memprioritaskan kasus-kasus khusus (*edge cases*) untuk ditinjau oleh manusia, serta menetapkan atau mengubah penandaan (*flag*) secara efisien agar lebih akurat mencerminkan status *enemy* (butir soal yang dianggap terlalu mirip satu sama lain).

Hasil-hasil penelitian tersebut menunjukkan potensi yang menjanjikan, sehingga para pengembang tes perlu didorong untuk mengeksplorasi kegunaan model leksikal dan semantik dalam mengotomatisasi setidaknya sebagian dari proses identifikasi *enemy* antar-butir soal. Perlu dicatat bahwa studi-studi tersebut juga mengidentifikasi sejumlah tantangan. Pertama, kode *enemy* yang sebelumnya ditetapkan oleh manusia dalam bank soal hampir pasti mengandung tingkat kesalahan tertentu; artinya, upaya manual mungkin diperlukan untuk menyediakan data pelatihan yang akurat dan mendukung model prediksi. Kedua, pasangan butir soal yang berstatus *enemy* mencakup bagian yang sangat kecil dari keseluruhan bank soal, sehingga peningkatan sensitivitas meskipun kecil pada model baru dapat mengakibatkan bertambahnya kode positif palsu (*false positive*).

Terakhir, metodologi NLP ini hanya dapat mengukur dan mengurutkan tingkat kemiripan di antara semua kemungkinan pasangan butir soal. Agar proses pengodean *enemy* secara otomatis benar-benar dapat menghemat waktu dan tenaga, pakar manusia perlu

meninjau hasilnya dan menentukan ambang batas kemiripan leksikal atau semantik; di atas ambang batas tersebut, semua butir soal dapat diklasifikasikan secara otomatis sebagai *enemy*, sehingga beban kerja pakar berkurang drastis dan hanya terbatas pada pasangan butir soal yang berada di ambang batas atau kasus-kasus *outlier* yang seharusnya dikategorikan sebagai *enemy* meskipun tingkat kemiripannya relatif rendah.

Memprediksi Tingkat Kesulitan Butir Soal

Upaya menggunakan NLP untuk memprediksi tingkat kesulitan butir soal, waktu respons, dan faktor-faktor lain berdasarkan sinyal linguistik tergolong baru dan memberikan hasil yang kurang menjanjikan dibandingkan dengan identifikasi butir soal yang saling bertentangan (*item enemy*). Upaya awal untuk memprediksi tingkat kesulitan butir soal pilihan ganda (MCQ) dalam ujian lisensi medis tidak memberikan hasil yang memuaskan. Studi lain membandingkan estimasi langsung tingkat kesulitan menggunakan LLM dengan strategi dua tahap berbasis fitur pada penilaian membaca dan matematika untuk jenjang K-5. Strategi yang menggunakan LLM untuk mengekstrak fitur kognitif dan linguistik terlebih dahulu, kemudian menggunakan ML untuk memprediksi tingkat kesulitan butir soal berdasarkan fitur-fitur tersebut alih-alih langsung dari butir soal itu sendiri menghasilkan prediksi yang lebih akurat, meskipun tidak secara konsisten pada semua jenjang kelas.

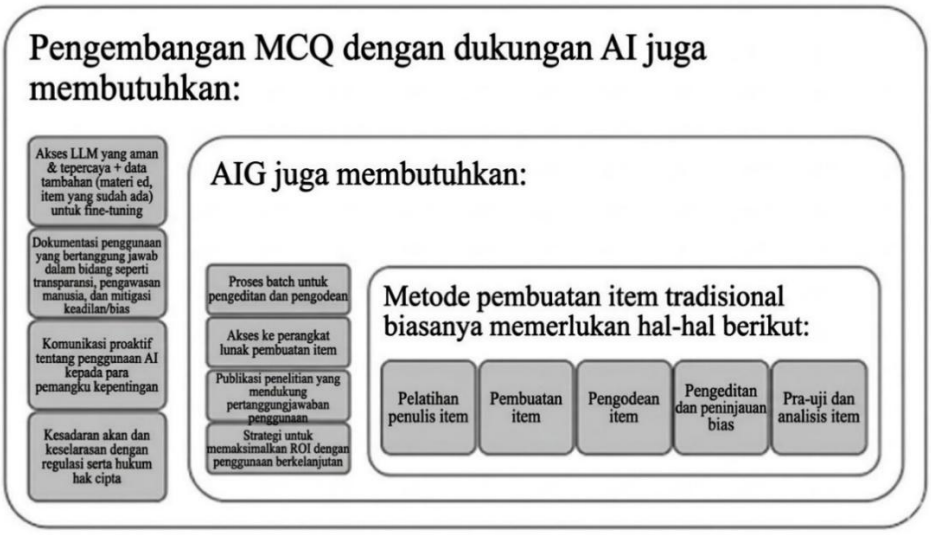
Topik ini tetap menjadi prioritas utama dalam bidang NLP untuk penilaian. Sebuah upaya besar yang dikenal sebagai *First Shared Task on Automated Prediction of Difficulty and Response Time for Multiple-Choice Items* (Tugas Bersama Pertama tentang Prediksi Otomatis Tingkat Kesulitan dan Waktu Respons untuk Butir Soal Pilihan Ganda) yang mengumpulkan solusi untuk prediksi tingkat kesulitan dan waktu respons telah merilis laporan pertamanya. Untuk inisiatif ini, sebuah dataset disusun yang mencakup teks butir soal, rata-rata waktu respons, dan nilai tingkat kesulitan model Rasch dari 667 butir soal yang baru saja ditarik dari seri USMLE®. Dataset ini diberikan kepada 17 tim dari berbagai belahan dunia; setiap tim dapat mengajukan hingga tiga solusi, dan hasilnya dinilai berdasarkan penurunan terbesar dalam *root mean square error* (RMSE) dibandingkan dengan RMSE untuk regresi *dummy* dasar (*baseline*). Berbagai model tersebut semuanya berhasil menurunkan RMSE dasar dalam jumlah kecil, meskipun terdapat variasi yang sangat minim di antara model-model tersebut. Meskipun potensi AI dalam mendukung berbagai aspek pembuatan butir soal tidak dapat disangkal, kemajuan ini juga menghadirkan tanggung jawab penting yang khas terkait penggunaan AI. Pada bagian selanjutnya, kami membahas topik-topik yang berkaitan dengan persyaratan hukum dan ekspektasi penggunaan yang bertanggung jawab. Pengembang tes harus siap menangani pertimbangan-pertimbangan ini secara proaktif demi menjaga integritas penilaian serta mempertahankan kepercayaan pemangku kepentingan terhadap proses pengujian dan skor yang dihasilkannya.

7.7 PERTIMBANGAN HUKUM DAN PENGGUNAAN YANG BERTANGGUNG JAWAB

Salah satu keuntungan besar dari pembuatan butir soal atau model secara tradisional adalah sifatnya yang transparan dan mudah dijelaskan kepada para pemangku kepentingan. Penulis soal dapat dimintai keterangan secara langsung mengenai potensi bias, dan organisasi

dapat bersikap transparan terkait metode pelatihan serta materi referensi. Meskipun pembuatan, penanganan, dan penyimpanan informasi yang dapat mengidentifikasi individu (PII) milik peserta tes mungkin terdampak oleh peraturan perundang-undangan seperti *General Data Protection Regulation* (Regulasi (UE) 2016/679) dan perlindungan privasi data lainnya, proses penulisan soal secara tradisional kemungkinan besar tidak akan pernah mengharuskan kepatuhan terhadap peraturan tambahan apa pun di tingkat yurisdiksi mana pun. Terakhir, kepemilikan atas soal yang dihasilkan tidak perlu diragukan lagi, dan perlindungan hak cipta sepenuhnya dapat diperoleh mengingat adanya kepenulisan manusia yang jelas.

Di sisi lain, penggunaan AI untuk menghasilkan soal pilihan ganda (MCQ) menghadirkan potensi tantangan bagi setiap elemen tersebut; penggunaan ini juga disertai dengan kesadaran pemangku kepentingan yang jauh lebih tinggi mengenai keberadaan, penggunaan, dan implikasi AI. Gambar 7.1 menggambarkan persyaratan yang saling melingkupi (*nested*) dalam berbagai metode penulisan soal, di mana setiap tingkatan mencakup seluruh elemen dari tingkatan yang berada di dalamnya. Dengan demikian, pelatihan bagi ahli materi pelajaran (SME) diperlukan untuk metode pengembangan soal apa pun; penggunaan AIG berarti juga mencakup proses penyuntingan dan pengodean sejumlah besar soal secara sekaligus; selain itu, penggunaan AI menuntut adanya kebijakan penggunaan yang bertanggung jawab serta kesadaran akan peraturan terkait AI.



Gambar 7.1 Diagram persyaratan yang saling melingkupi untuk tiga metode pengembangan soal.

Pemilihan dan penggunaan LLM telah dibahas sebelumnya dalam bab ini. Di sini, kita beralih membahas kekhawatiran umum pemangku kepentingan mengenai penggunaan AI yang bertanggung jawab (keadilan/bias, transparansi, dan pengawasan manusia), pedoman komunikasi proaktif, serta pembahasan mengenai dua isu hukum mendesak terkait pembuatan soal pilihan ganda dengan bantuan AI.

Penggunaan AI yang Bertanggung Jawab dan Komunikasi dengan Pemangku Kepentingan

Bias dalam butir soal penilaian baik yang bersifat budaya, linguistik, maupun sosial-ekonomi dapat mendistorsi pengukuran dengan memunculkan varians yang tidak relevan dengan konstruk (*construct-irrelevant variance*), sehingga mengganggu prinsip keadilan dan argumen validitas penggunaan skor tes. Standar untuk menjamin keadilan dan memitigasi bias dalam setiap penggunaan AI, termasuk pembuatan soal pilihan ganda (MCQ) oleh AI, muncul sebagai salah satu prinsip penggunaan AI yang bertanggung jawab yang paling umum diterapkan di berbagai organisasi. Oleh karena itu, organisasi yang menggunakan AI untuk membuat soal pilihan ganda harus siap menyediakan dokumentasi yang memadai guna mendukung argumen mengenai keadilan; dokumentasi ini mencakup rincian tentang sumber data pelatihan, arsitektur model, dan langkah-langkah pra-pemrosesan, serta dasar pemikiran di balik perancangan *prompt* dan penyempurnaan butir soal. Dokumentasi tersebut harus mencatat rincian tinjauan keadilan dan bias yang sebelumnya telah diterapkan dalam penulisan dan penerimaan soal secara konvensional, serta bagaimana prosedur tersebut diperbarui untuk penggunaan AI termasuk representasi demografis peninjau manusia terhadap soal buatan AI dan kriteria tambahan yang digunakan untuk mengevaluasi potensi bias pada soal tersebut. Catatan-catatan ini dapat menjadi kunci dalam argumen yang memperkuat validitas skor ujian.

Selain dokumentasi, organisasi perlu menetapkan proses formal untuk menyelidiki potensi bias pada soal buatan AI guna menghindari persepsi atau kenyataan adanya dampak negatif terhadap validitas tes. Praktik terbaik mencakup pelaksanaan analisis kuantitatif, seperti *differential item functioning* (DIF), untuk mendeteksi perbedaan antar-subkelompok yang mungkin terlewatkan oleh peninjau manusia. Metode deteksi bias otomatis, seperti uji asosiasi *embedding*, dapat melengkapi tinjauan manusia dengan menyediakan evaluasi bias yang dapat diskalakan dan direplikasi pada kumpulan soal yang besar. Investigasi ini harus bersifat iteratif dan terintegrasi ke dalam siklus pengembangan tes, guna memastikan bahwa bias yang teridentifikasi telah ditangani sebelum tes digunakan secara operasional. Dengan memadukan dokumentasi yang kuat dan analisis bias yang sistematis, organisasi penyelenggara tes dapat mempertahankan standar validitas serta menjaga kepercayaan publik terhadap penilaian yang didukung oleh AI.

Transparansi dalam pengembangan soal pilihan ganda yang dibantu AI menuntut komunikasi yang jelas dengan para pemangku kepentingan, termasuk peserta tes. Organisasi harus siap untuk menyampaikan secara terbuka alasan pemilihan metode AI tertentu, cara penggunaannya, serta nilai atau manfaat yang diharapkan darinya. Penelitian pendukung apa pun mungkin mengharuskan adanya publikasi awal mengenai data uji coba (*pilot data*). Tergantung pada peraturan perundang-undangan, mungkin diperlukan strategi untuk memungkinkan peserta tes memilih untuk tidak berpartisipasi dalam penelitian atau penilaian yang melibatkan AI. Salah satu tantangan utama adalah sifat "kotak hitam" (*black-box*) dari LLM, di mana kecepatan dan akurasi sering kali dicapai dengan mengorbankan kemampuan untuk memberikan penjelasan (*explainability*).

Untuk mengatasi hal ini, organisasi dapat memulai dengan penggunaan AI yang lebih sederhana dan lebih mudah dipahami seperti metode kemiripan berbasis NLP sebelum beralih ke model *encoder* yang kompleks dengan elemen *machine learning* (ML), meskipun metode yang lebih sederhana tersebut mungkin memiliki tingkat akurasi yang lebih rendah. Pengawasan manusia yang terdokumentasi (atau dikenal sebagai *humans in the loop*) ditekankan dalam pedoman etika publik serta contoh-contoh lain yang disebutkan dalam bab ini; namun, hingga saat ini belum ada standar universal mengenai sejauh mana dan kapan keterlibatan manusia tersebut harus dilakukan. Organisasi penyelenggara tes sebaiknya menyusun pedoman internal mereka sendiri dengan mempertimbangkan masukan pemangku kepentingan, analisis risiko, titik-titik kesalahan yang umum terjadi dalam penulisan butir soal, serta materi pelatihan yang telah diperbarui (untuk contoh prosesnya, lihat Burstein.

Organisasi penyelenggara tes juga harus mengomunikasikan perubahan pada program ujian sebelumnya guna menjaga transparansi dan kepercayaan. Meskipun perubahan kecil dalam metodologi penulisan butir soal biasanya tidak diumumkan kepada peserta tes dan pengguna skor lainnya, penggunaan AI bahkan dalam penilaian formatif sebaiknya diungkapkan secara proaktif. Sentimen publik terhadap AI bisa jadi negatif, dengan adanya kekhawatiran mengenai kualitas keluaran, bias otomatisasi, dan berkurangnya keterlibatan manusia dalam pengambilan keputusan yang berdampak besar (*high-stakes decisions*). Komunikasi yang efektif harus menjelaskan alasan penggunaan LLM, data pendukung, pengawasan manusia, serta langkah-langkah pengamanan untuk menjamin akurasi dan keadilan. Upaya sosialisasi perlu mencakup lebih dari sekadar dokumen teknis (*white papers*); hal ini bisa meliputi webinar, siaran pers, pembaruan situs web, dan materi peserta tes yang telah direvisi. Tantangan utamanya adalah menyeimbangkan transparansi dengan keamanan tes: detail seperti *prompt*, data *fine-tuning*, atau proses pembuatan konten mungkin perlu dirahasiakan untuk mencegah peniruan materi ujian.

Penggunaan AI Berisiko Tinggi

Regulasi yang muncul terkait penggunaan AI biasanya dikelompokkan berdasarkan tingkat risiko, dengan perhatian terbesar ditujukan pada teknologi yang berpotensi menimbulkan dampak buruk signifikan, sehingga dilarang atau memerlukan pengawasan ketat. Definisi risiko tinggi bervariasi antarwilayah hukum, namun sebagian besar regulasi termasuk EU AI Act (*Regulation* (EU) 2024/1689 (*Artificial Intelligence* Act)) dan Colorado AI Act (*Colorado General Assembly*) mengategorikan aplikasi AI yang dapat memengaruhi peluang pendidikan, pelatihan vokasi, dan pekerjaan sebagai berisiko tinggi. Hingga tahun 2025, belum ada konsensus industri yang dipublikasikan mengenai kapan butir soal buatan AI akan memicu penerapan aturan-aturan tersebut. Secara umum diasumsikan bahwa penilaian formatif akan dianggap berisiko rendah atau tanpa risiko, namun organisasi penyelenggara tes perlu mengantisipasi bahwa soal pilihan ganda (*multiple-choice questions* atau MCQ) buatan AI untuk ujian berisiko tinggi dapat dikategorikan sebagai berisiko tinggi, sehingga memerlukan dokumentasi kepatuhan dan audit terhadap sumber data.

Pertimbangan Hak Cipta

US Copyright Office menyatakan pada tahun 2023 bahwa perlindungan hak cipta hanya berlaku "untuk materi yang merupakan hasil kreativitas manusia" (*US Copyright Office*). Putusan-putusan terkini menegaskan kembali bahwa pembuatan butir soal secara massal dengan tinjauan manusia yang minim dapat menghalangi klaim hak cipta, sebagian karena kekhawatiran bahwa keluaran LLM mungkin tidak cukup transformatif untuk memenuhi ketentuan penggunaan wajar (*fair use*). Karena pengguna tidak dapat memverifikasi semua sumber pelatihan, terdapat pula risiko reproduksi materi berhak cipta secara tidak sengaja. Namun, ada juga potensi bahwa jika AI digunakan untuk membantu bukan menciptakan sepenuhnya bagian dari soal pilihan ganda (MCQ) yang didokumentasikan sebagai hasil karya manusia dapat dilindungi hak ciptanya.

Oleh karena itu, salah satu rekomendasinya adalah merancang sistem dan alur kerja AI di mana LLM hanya berfungsi sebagai alat penyusun draf awal; tinjauan dan revisi oleh manusia selanjutnya haruslah substansial, terdokumentasi, dan, jika memungkinkan, dapat diukur. Jika potensi perlindungan hak cipta masih dianggap rendah, organisasi dapat mempertimbangkan penerapan perlindungan alternatif seperti *Defend Trade Secrets Act* (DTSA). Berdasarkan DTSA, soal pilihan ganda yang dihasilkan AI dapat dikategorikan sebagai rahasia dagang jika soal tersebut memberikan nilai ekonomi karena kerahasiaannya dan jika langkah-langkah yang wajar telah diambil untuk menjaga kerahasiaan tersebut.

7.8 TANGGUNG JAWAB DALAM INTEGRASI AI PADA PEMBUATAN BUTIR SOAL

Sejarah pengembangan butir soal memberikan pelajaran penting. Penulisan butir soal secara tradisional menawarkan autentisitas namun sering kali lambat dan memakan biaya tinggi. Pemodelan butir soal meningkatkan efisiensi dan pemodelan kognitif memperkuat keselarasan teoretis, namun masalah kemiripan butir soal dan hambatan dalam tahap pra-pengujian justru semakin mengemuka. Setiap tahapan menyelesaikan beberapa masalah sekaligus menciptakan tantangan baru, yang menegaskan bahwa tidak ada metode yang bebas dari konsekuensi timbal-balik (*trade-off*). AI melanjutkan pola ini dengan menawarkan kecepatan, kemampuan adaptasi, dan kreativitas, sembari menghadirkan risiko terkait bias, kurangnya transparansi, dan tantangan hukum. Tugas bagi para psikometrisian adalah menentukan cara mengintegrasikan AI secara bertanggung jawab guna menjaga validitas dan keadilan, sekaligus memanfaatkan potensi yang dimilikinya.

Pertimbangan mengenai penggunaan yang legal, etis, dan bertanggung jawab menjadi hal yang sangat mendesak. Hukum hak cipta hanya melindungi materi yang diciptakan oleh manusia, sehingga status kepemilikan konten yang dihasilkan AI menjadi tidak pasti; di sisi lain, perlindungan rahasia dagang menuntut kerahasiaan yang ketat serta adanya justifikasi ekonomi. Berbagai pertanyaan pun muncul: mampukah organisasi secara realistis melindungi materi dalam jumlah besar yang dihasilkan AI sebagai rahasia dagang? Bagaimana keterlibatan manusia harus didokumentasikan agar tetap memenuhi syarat hak cipta? Dan apa yang terjadi ketika model secara tidak sengaja mereproduksi elemen dari data pelatihannya? Pada saat yang sama, LLM yang dilatih menggunakan kumpulan data masif dapat memuat bias budaya

atau demografis, yang berpotensi merugikan kelompok pembelajar tertentu. Metode teknis untuk mendeteksi bias memang mulai bermunculan, namun metode tersebut tidak dapat menggantikan nuansa dan penilaian dari pakar manusia. Peran UKM tetap tak tergantikan, namun upaya mereka dalam mengadaptasi teknologi ini harus dirancang dengan baik dan dikelola secara cermat.

Transparansi dan kolaborasi merupakan landasan utama dalam integrasi AI yang bertanggung jawab. Para pemangku kepentingan harus bekerja sama untuk menjamin keterbukaan mengenai bagaimana materi dibuat, konstruk apa yang diukur, serta bagaimana prinsip keadilan dijaga. Transparansi menjadi tantangan tersendiri akibat sifat AI yang menyerupai "kotak hitam" (*black-box*) dan kebutuhan untuk melindungi kekayaan intelektual; hal ini menuntut organisasi untuk menyeimbangkan antara keterbukaan informasi dan aspek keamanan, sembari mengembangkan strategi yang tangkas (*agile*) guna memungkinkan inovasi tanpa mengorbankan validitas. Integrasi AI bukanlah tujuan akhir, melainkan sebuah langkah awal sebuah ajakan untuk berdialog dan memikul tanggung jawab bersama. Keputusan yang diambil saat ini terkait efisiensi, biaya, dan keadilan akan membentuk masa depan sistem penilaian; adapun apakah penggunaan AI akan memperkuat atau justru melemahkan bukti validitas, hal itu sangat bergantung pada seberapa cermat organisasi dalam menyikapi berbagai isu yang telah diuraikan dalam bab ini.

BAB 8
MENGUKUR PEMBELAJARAN MELALUI BERMAIN

8.1 PENTINGNYA PENGUKURAN DALAM PERMAINAN EDUKASI

Pembelajaran berbasis permainan (*game-based learning* atau GBL) telah berkembang jauh melampaui peran awalnya sebagai pelengkap motivasional bagi pengajaran tradisional. Permainan edukasi kontemporer semakin berfungsi sebagai lingkungan pembelajaran kompleks yang melibatkan peserta didik dalam pemecahan masalah yang berkelanjutan, pengambilan keputusan strategis, kolaborasi, dan refleksi. Melalui interaksi ini, permainan memunculkan proses kognitif, perilaku, dan afektif yang kaya serta berkembang secara dinamis seiring waktu. Saat bermain, peserta didik menghasilkan aliran data yang berkelanjutan berupa pilihan yang dibuat, strategi yang dicoba, kesalahan yang diperbaiki, pola kegigihan, dan respons terhadap umpan balik yang memiliki potensi besar sebagai bukti pembelajaran. Namun, tanpa kerangka pengukuran yang berlandaskan prinsip yang kuat, data ini sebagian besar tetap bersifat deskriptif dan hanya memberikan wawasan terbatas mengenai apa yang diketahui peserta didik, bagaimana mereka bernalar, atau apa yang mampu mereka lakukan. Oleh karena itu, pengukuran memegang peranan sentral dalam mengubah aktivitas bermain menjadi bukti pembelajaran yang bermakna dan dapat dipertanggungjawabkan. Pengukuran pendidikan menyediakan pendekatan sistematis untuk menafsirkan perilaku peserta didik melalui definisi konstruk yang eksplisit, proses inferensi yang terstruktur, dan teori validitas. Dalam konteks permainan edukasi, pengukuran memperjelas bentuk pembelajaran apa yang dituju, bagaimana bentuk pembelajaran tersebut diharapkan muncul melalui aktivitas bermain, serta asumsi apa yang mendasari klaim mengenai pengetahuan, keterampilan, atau disposisi peserta didik berdasarkan perilaku yang teramati. Konstruk seperti kemampuan pemecahan masalah, berpikir kritis, atau kolaborasi tidak dapat disimpulkan secara langsung dari jejak aktivitas bermain.

Sebaliknya, konstruk-konstruk tersebut harus dirumuskan secara eksplisit dan didasarkan pada teori, dengan hubungan yang jelas antara konstruk laten dan tindakan yang dapat diamati di dalam permainan. Argumen validitas kemudian menghubungkan pengamatan tersebut dengan interpretasi yang dituju, guna memastikan bahwa inferensi yang diambil dari data permainan didukung oleh teori dan bukti, bukan sekadar intuisi atau korelasi dangkal. Perbedaan krusial dalam ranah ini terletak pada penggunaan permainan sebagai sarana pembelajaran versus penggunaan data permainan sebagai bukti penilaian. Banyak permainan edukatif dirancang terutama untuk mendorong pembelajaran melalui keterlibatan, eksplorasi, dan latihan. Namun, sejumlah penelitian terkini menekankan peran permainan sebagai lingkungan penilaian yang terintegrasi atau bersifat "terselubung" (*stealth assessment*), di mana penilaian menjadi fitur tak terpisahkan dari sistem permainan itu sendiri, bukan sekadar lapisan pengujian eksternal.

Dalam paradigma ini, peserta didik dievaluasi melalui interaksi alami mereka dengan dunia permainan, sering kali tanpa menyadari secara eksplisit bahwa penilaian sedang

berlangsung. Pendekatan ini berpotensi mengurangi kecemasan menghadapi tes, meningkatkan validitas ekologis, dan menangkap proses pembelajaran saat berlangsung dalam konteks yang autentik. Dalam pengertian ini, pengukuranlah yang memungkinkan data permainan berfungsi sebagai data penilaian, yang mendukung pemberian umpan balik formatif, jalur pembelajaran adaptif, serta interpretasi sumatif terhadap kinerja. Yang tak kalah penting, desain pengukuran yang cermat memungkinkan permainan edukatif berfungsi sebagai lingkungan penilaian yang valid tanpa mengorbankan keterlibatan, autentisitas, atau otonomi peserta didik. Metrik yang tidak selaras, aturan penilaian yang tidak transparan, atau mekanisme pengujian yang mengganggu dapat merusak imersi dan mendistorsi perilaku peserta didik, sehingga pada akhirnya mengurangi keterlibatan maupun validitas. Sebaliknya, sistem pengukuran yang dirancang dengan baik memanfaatkan berbagai potensi permainan seperti interaktivitas, adaptivitas, dan kekayaan konteksnya untuk menangkap proses pembelajaran secara langsung di dalam permainan (*in situ*). Oleh karena itu, penyelarasan antara mekanika permainan, tujuan pembelajaran, dan model pengukuran sangatlah penting untuk mempertahankan kekuatan motivasional permainan sekaligus menghasilkan bukti pembelajaran yang andal, dapat diinterpretasikan, dan dapat ditindaklanjuti. Singkatnya, pengukuran memegang peranan penting dalam permainan edukatif karena menjembatani kesenjangan antara keterlibatan dan bukti pembelajaran. Hal ini memungkinkan permainan edukatif untuk melampaui sekadar pengalaman belajar yang menarik, dan berkembang menjadi lingkungan yang kredibel untuk memahami, mendukung, serta mengevaluasi pembelajaran dalam konteks yang kompleks dan dinamis.

Masalah Penelitian dan Kontribusi

Meskipun lingkungan *Game-Based Learning* (GBL) berkembang dan diadopsi dengan pesat, masih terdapat tantangan besar terkait bagaimana penilaian yang terintegrasi dalam permainan dikonseptualisasikan, divalidasi, dan diinterpretasikan. Sebagian besar literatur GBL yang ada saat ini lebih menekankan pada keterlibatan, motivasi, atau hasil pembelajaran, namun memberikan perhatian yang relatif terbatas pada aspek penilaian sebagai suatu proses inferensial yang ketat.

Isu-isu mendasar terkait definisi konstruk, validitas, inferensi, dan keadilan sering kali kurang dirumuskan secara jelas atau hanya ditangani secara implisit; data permainan sering dianggap sebagai indikator pembelajaran yang sudah jelas dengan sendirinya, alih-alih dipandang sebagai bukti yang memerlukan interpretasi berbasis prinsip. Akibatnya, banyak penilaian berbasis permainan (GBA) menghasilkan temuan yang sulit untuk dipertanggungjawabkan, dibandingkan, atau digeneralisasikan lintas peserta didik, konteks, maupun populasi.

Oleh karena itu, masalah penelitian utama yang muncul adalah belum adanya kerangka kerja koheren yang mengintegrasikan teori pembelajaran, desain permainan, dan pengukuran pendidikan. Kemajuan dalam analitik pembelajaran telah memungkinkan pengumpulan dan analisis data proses yang sangat rinci dari permainan, namun pendekatan analitik ini tidak selalu didasarkan pada model konstruk yang eksplisit ataupun argumen validitas yang terumuskan dengan jelas. *Evidence-centered design* (ECD) telah banyak disebut sebagai

kerangka kerja yang menjanjikan untuk GBA, namun penerapannya sering kali bersifat parsial. Dalam banyak kasus, ECD terutama diterapkan pada tahap desain tugas, sementara dasar pembenaran (*warrants*), asumsi, dan penjelasan alternatif yang mendasari inferensi masih belum dikembangkan secara memadai.

Kekhawatiran terkait keadilan dan bias semakin memperumit penggunaan permainan sebagai lingkungan penilaian. Perbedaan dalam pengalaman bermain sebelumnya, tingkat keakraban dengan antarmuka digital, interpretasi budaya terhadap mekanisme permainan, dan proses pengambilan keputusan algoritmik dapat menimbulkan varians yang tidak relevan dengan konstruk serta kerugian sistematis bagi kelompok tertentu. Ketika faktor-faktor tersebut tidak dimodelkan atau dievaluasi secara eksplisit, inferensi yang didorong oleh kecerdasan buatan (AI) berisiko memperparah ketimpangan alih-alih menguranginya.

Bab ini membahas tantangan-tantangan tersebut dengan menyintesiskan teori pengukuran pendidikan, ECD, dan analitik pembelajaran berbasis AI untuk merumuskan pendekatan berlandaskan prinsip dalam merancang, memvalidasi, dan menginterpretasikan penilaian yang terintegrasi dalam permainan edukatif. Alih-alih memandang permainan terutama sebagai sumber data baru, bab ini menempatkan GBA sebagai masalah pengukuran yang memerlukan pemodelan konstruk secara eksplisit, rantai inferensi yang dapat dipertanggungjawabkan, dan evaluasi validitas yang sistematis. Bab ini mengkaji bagaimana perilaku yang dapat diamati di dalam permainan dapat berfungsi sebagai bukti untuk klaim penilaian, asumsi yang mendasari keabsahan inferensi tersebut, serta implikasinya terhadap keadilan, transparansi, dan kemampuan interpretasi.

Kontribusi karya ini mencakup aspek konseptual maupun metodologis. Secara konseptual, karya ini menempatkan kembali permainan edukatif sebagai sistem pengukuran yang terintegrasi dalam lingkungan pembelajaran yang dinamis. Secara metodologis, karya ini memberikan panduan bagi peneliti dan perancang yang ingin beralih dari analitik deskriptif menuju praktik pengukuran yang berlandaskan teori, transparan, dan berkeadilan. Dengan demikian, bab ini bertujuan mendukung pengembangan GBA yang tidak hanya menarik dan inovatif, tetapi juga valid, dapat diinterpretasikan, dan kredibel untuk pengambilan keputusan di bidang pendidikan.

8.2 DESAIN PENILAIAN BERBASIS PERMAINAN

Kerangka Kerja Desain untuk Penilaian Berbasis Permainan (GBA)

Salah satu kerangka kerja utama untuk desain GBA, khususnya dalam lingkungan interaktif yang didukung AI, adalah ECD. ECD menyediakan pendekatan penilaian yang berlandaskan prinsip dengan menyusun secara eksplisit rantai inferensi yang menghubungkan perilaku yang dapat diamati dengan klaim mengenai pengetahuan, keterampilan, atau disposisi peserta didik. Alih-alih memperlakukan penilaian sebagai prosedur pemberian skor yang diterapkan setelah tugas selesai, ECD mengonseptualisasikan penilaian sebagai argumen berbasis bukti yang terdiri dari tiga elemen yang saling terkait: (a) klaim yang merinci apa yang diketahui atau dapat dilakukan oleh pembelajar, (b) bukti yang diperlukan untuk mendukung klaim tersebut, dan (c) tugas yang dirancang untuk memunculkan bukti tersebut.

Dalam lingkungan berbasis permainan, elemen-elemen ini dapat diselaraskan secara sistematis dengan mekanika permainan, fitur telemetri, dan model analitik berbasis AI. Klaim berkaitan dengan kompetensi yang ditargetkan atau kondisi laten pembelajar, seperti penalaran strategis, pemikiran sistem, atau pemecahan masalah secara kolaboratif. Bukti diwujudkan dalam jejak permainan yang rinci, termasuk urutan tindakan, perubahan strategi, pola kesalahan, lintasan kegigihan, atau respons terhadap umpan balik. Tugas diwujudkan melalui tantangan permainan, skenario, dan fitur interaksi (*affordances*) yang menciptakan peluang bagi pembelajar untuk menunjukkan perilaku-perilaku tersebut. Dalam kerangka kerja ini, model AI seperti model grafis probabilistik, model sekuensial, atau pengklasifikasi pembelajaran mesin beroperasi sebagai mesin inferensi yang mengubah data permainan bervolume tinggi menjadi estimasi mengenai kemahiran atau kondisi pembelajar.

Kekuatan ECD terletak pada artikulasi eksplisit mengenai asumsi inferensi dan dasar pembenaran validitas. Dengan mengharuskan perancang untuk menetapkan konstruk, hubungan berbasis bukti, dan tuntutan tugas sejak awal (*a priori*), ECD mendukung interpretasi yang dapat dipertanggungjawabkan atas inferensi yang dihasilkan AI serta menyediakan landasan yang koheren untuk validasi. Fitur ini sangat krusial dalam lingkungan permainan adaptif yang kaya data, di mana kompleksitas interaksi dapat mengaburkan hubungan antara perilaku yang teramati dan konstruk yang dituju. ECD tidak menentukan teknik analitik spesifik; sebaliknya, ECD membatasi penggunaannya dengan menempatkan teknik-teknik tersebut dalam argumen pengukuran yang eksplisit.

Perspektif lain mengenai penilaian dan pembelajaran menawarkan wawasan pelengkap namun memberikan panduan yang kurang langsung untuk inferensi algoritmik. Penilaian sebagai pembelajaran (*assessment-as-learning*) menekankan pada refleksi pembelajar dan penggunaan informasi penilaian secara formatif, sementara perkembangan pembelajaran (*learning progressions*) berfokus pada lintasan perkembangan kompetensi dari waktu ke waktu (*National Research Council*).

Kedua perspektif tersebut berharga untuk desain dan interpretasi instruksional, khususnya dalam konteks formatif. Namun, tidak ada satu pun yang menyediakan kerangka kerja yang terartikulasi secara utuh untuk merinci bagaimana data perilaku yang kompleks harus diterjemahkan menjadi klaim yang dapat dipertanggungjawabkan melalui model komputasi. Fokus ECD pada klaim, bukti, dan tugas menjadikannya sangat tepat untuk mengintegrasikan desain permainan, pengukuran pendidikan, dan analitik berbasis AI dalam sistem penilaian yang terintegrasi (*embedded assessment systems*).

Bentuk-Bentuk Penilaian dalam Pembelajaran Berbasis Permainan

Penilaian dalam lingkungan GBL (pembelajaran berbasis permainan) dapat mengambil berbagai bentuk, yang berbeda dalam tingkat integrasinya dengan alur permainan serta kecanggihan analisisnya. Salah satu pendekatan umum melibatkan penilaian yang dilakukan di luar permainan, seperti *pre-test* (tes awal) dan *post-test* (tes akhir). Ukuran-ukuran ini memberikan informasi sumatif atau komparatif mengenai hasil belajar dan sering kali dipahami dengan baik secara psikometris. Namun, metode ini sebagian besar terpisah dari aktivitas di dalam permainan dan tidak memanfaatkan kekayaan data permainan. Akibatnya,

metode ini hanya memberikan wawasan terbatas mengenai proses pembelajaran atau penggunaan strategi selama bermain.

Bentuk kedua mengandalkan metrik di dalam permainan, seperti skor, level, lencana, waktu penyelesaian, atau indikator pencapaian. Ukuran-ukuran ini sangat terkait erat dengan alur permainan serta sering kali mudah dihitung dan dikomunikasikan. Akan tetapi, metrik ini sering kali memiliki spesifikasi yang lemah terkait konstruk dasarnya dan jarang disertai dengan argumen validitas yang eksplisit. Dalam banyak kasus, indikator semacam itu mencampuradukkan kinerja, keterlibatan, dan mekanisme perkembangan, sehingga sulit untuk menafsirkan apa yang sebenarnya direpresentasikan oleh indikator tersebut dalam konteks pembelajaran.

Bentuk ketiga yang pengaruhnya semakin besar adalah penilaian terselubung (*stealth assessment*) yang didukung oleh AI; dalam metode ini, penilaian disematkan langsung ke dalam alur permainan dan berjalan secara berkelanjutan saat peserta didik berinteraksi dengan permainan. Alih-alih menghentikan permainan untuk melakukan tes eksplisit, penilaian terselubung memanfaatkan perilaku, pilihan, strategi, kesalahan, dan kegigihan pemain yang muncul secara alami sebagai bukti pembelajaran. Tipologi ini mencerminkan perkembangan dari penilaian *di sekitar* permainan, menuju penilaian *di dalam* permainan, hingga penilaian *melalui* permainan, yang disertai dengan peningkatan validitas ekologis, resolusi diagnostik, serta keselarasan antara proses pembelajaran dan pengukuran.

Yang terpenting, pergerakan menuju penilaian yang tersemat dan terselubung ini tidak menghilangkan kebutuhan akan ketepatan pengukuran. Sebaliknya, hal ini justru meningkatkan pentingnya pemodelan konstruk yang eksplisit dan penarikan kesimpulan yang berlandaskan prinsip yang kuat. Ketika penilaian tidak dapat dipisahkan dari alur permainan, beban interpretasi beralih dari butir-butir tes yang terpisah ke aliran data interaksi yang berkelanjutan; hal ini menjadikan kerangka kerja seperti ECD (*Evidence-Centered Design*) sangat penting untuk menjaga kejelasan mengenai apa yang diukur dan mengapa hal tersebut diukur.

Penilaian Terselubung dan Penarikan Kesimpulan Berbasis AI

Penilaian terselubung (*stealth assessment*) mengacu pada penilaian yang tidak mengganggu, bersifat formatif, dan berkelanjutan, yang terintegrasi secara mulus ke dalam alur permainan. Dalam pendekatan ini, penilaian tidak dirasakan oleh peserta didik sebagai kegiatan evaluasi yang terpisah. Sebaliknya, peserta didik menunjukkan kompetensi mereka melalui interaksi alami dengan lingkungan permainan, dan interaksi tersebut direkam sebagai data telemetri. Model inferensi berbasis AI menganalisis data ini secara waktu nyata (*real-time*) untuk memperbarui estimasi mengenai pengetahuan, keterampilan, atau disposisi peserta didik.

Sistem penilaian terselubung (*stealth assessment*) umumnya didasarkan pada prinsip-prinsip ECD. Komponen klaim menetapkan kompetensi yang menjadi fokus; model bukti mengidentifikasi fitur-fitur permainan yang memberikan informasi mengenai kompetensi tersebut; dan model tugas menentukan kondisi di mana bukti semacam itu dapat diperoleh. Inferensi berbasis AI memungkinkan pembaruan dinamis terhadap model peserta didik,

sehingga penilaian dan instruksi dapat berkembang secara bersamaan selama permainan berlangsung. Integrasi ini mendukung pemberian umpan balik adaptif, *scaffolding* yang dipersonalisasi, serta urutan tugas yang responsif, semuanya tanpa mengganggu pengalaman imersif pemain.

Dari perspektif pengukuran, penilaian terselubung (*stealth assessment*) menandai pergeseran dari potret kinerja yang statis menuju representasi dinamis dari proses pembelajaran. Alih-alih hanya menanyakan apakah peserta didik dapat memberikan jawaban yang benar pada satu titik waktu tertentu, penilaian terselubung mengamati bagaimana peserta didik mendekati masalah, merevisi strategi, dan merespons tantangan sepanjang rangkaian interaksi yang berkelanjutan. Fokus yang berorientasi pada proses ini tidak hanya selaras dengan teori pembelajaran dan kompetensi terkini, tetapi juga menghadirkan tantangan baru terkait inferensi, validitas, dan keadilan. Karena inferensi didasarkan pada pola yang terakumulasi dari waktu ke waktu, pemodelan eksplisit mengenai ketergantungan temporal dan lintasan pembelajaran menjadi hal yang krusial.

Selain itu, sifat penilaian terselubung yang tidak mencolok menuntut tanggung jawab tambahan dari para perancang dan peneliti untuk menjamin transparansi dan kemampuan interpretasi. Peserta didik mungkin tidak menyadari bagaimana tindakan mereka ditafsirkan; oleh karena itu, sangatlah penting bahwa model inferensi memiliki landasan teoretis yang kuat dan argumen validitas secara jelas membenarkan interpretasi serta penggunaan informasi penilaian yang diusulkan.

Landasan Teori Pembelajaran dan Interaksi yang Dapat Diukur

Perancangan dan interpretasi penilaian berbasis permainan (*Game-Based Assessments* atau GBA) juga didasarkan pada teori pembelajaran yang menekankan interaksi, partisipasi, dan motivasi. Perspektif konstruktivis dan situatif mengonseptualisasikan pembelajaran sebagai proses aktif pembentukan makna yang terjadi dalam konteks autentik dan praktik sosial. Berdasarkan sudut pandang ini, permainan bukan sekadar sarana penyampaian materi, melainkan lingkungan partisipatif tempat peserta didik terlibat dalam pemecahan masalah, eksperimen, dan kolaborasi. Pengetahuan dibangun melalui interaksi dengan berbagai alat, representasi, dan pemain lain bukan sekadar ditransfer melalui tugas-tugas yang terisolasi.

Teori determinasi diri (*self-determination theory*) melengkapi perspektif-perspektif tersebut dengan mengidentifikasi otonomi, kompetensi, dan keterhubungan (*relatedness*) sebagai faktor pendorong utama keterlibatan dan pembelajaran. Lingkungan berbasis permainan memiliki posisi yang sangat strategis untuk mendukung kebutuhan psikologis tersebut melalui pilihan pemain, tantangan yang adaptif, dan umpan balik yang responsif. Ketika dilengkapi dengan adaptasi berbasis kecerdasan buatan (AI), permainan dapat menyesuaikan tingkat kesulitan tugas dan umpan balik secara dinamis untuk mempertahankan tingkat tantangan yang optimal, sehingga memengaruhi motivasi maupun perilaku pembelajaran yang dapat diamati.

Teori-teori pembelajaran ini juga menjadi dasar bagi aspek keterukuran. Strategi kognitif, kondisi motivasi, dan perilaku pengaturan diri diharapkan termanifestasi sebagai sinyal permainan yang dapat dideteksi, seperti pola eksplorasi, lintasan persistensi, perilaku

mencari bantuan, dinamika kolaborasi, atau respons terhadap umpan balik. Dengan memetakan konstruk teoretis secara eksplisit ke data interaksi yang dapat diamati, perancang dapat memastikan bahwa model AI mendasarkan inferensinya pada teori, alih-alih hanya mengandalkan korelasi berbasis data. Keterkaitan ini memperkuat aspek keterpahaman (*interpretability*), mendukung argumen validitas, serta meningkatkan daya penjelas sistem GBA yang didukung AI.

Pada saat yang sama, teori pembelajaran menyoroti risiko adanya varians yang tidak relevan dengan konstruk (*construct-irrelevant variance*). Perilaku yang diamati dalam permainan mungkin mencerminkan faktor-faktor seperti pengalaman bermain sebelumnya, tingkat keakraban dengan antarmuka, atau gaya bermain pribadi, alih-alih mencerminkan konstruk yang menjadi sasaran pengukuran. Memahami perbedaan ini memperkuat perlunya spesifikasi konstruk, desain tugas, dan evaluasi validitas yang cermat aspek-aspek yang menjadi semakin krusial dalam lingkungan adaptif yang digerakkan oleh AI.

8.3 PSIKOMETRI UNTUK PERMAINAN BERBASIS AI

Permainan berbasis AI yang digunakan untuk penilaian mewakili perubahan mendasar dari format tes konvensional dalam pengukuran pendidikan dan psikologi. Sistem ini mengintegrasikan AI, algoritma adaptif, dan pengambilan data secara berkelanjutan untuk mendukung pengambilan kesimpulan (*inference*) secara *real-time* mengenai kompetensi pembelajar. Meskipun kemampuan teknis sistem semacam itu sangat berbeda dari penilaian tradisional, prinsip-prinsip dasar psikometri tetap penting untuk memastikan bahwa kesimpulan yang dihasilkan oleh AI bermakna, dapat dipertanggungjawabkan, dan adil. Oleh karena itu, tantangan utamanya bukanlah apakah psikometri dapat diterapkan pada permainan berbasis AI, melainkan bagaimana prinsip-prinsip yang sudah mapan harus dikembangkan untuk mengakomodasi lingkungan penilaian yang dinamis, sekuensial, dan adaptif.

Prinsip-Prinsip Utama Pengukuran

Validitas, reliabilitas, keadilan, dan generalisasi tetap menjadi pertimbangan mendasar, bahkan ketika kesimpulan dihasilkan melalui model berbasis AI dan bukan melalui aturan penilaian yang baku (*fixed scoring rules*). Validitas berkaitan dengan apakah interpretasi yang diperoleh dari data permainan yang dimediasi AI sudah tepat untuk tujuan penggunaannya, sementara reliabilitas mengacu pada stabilitas dan konsistensi kesimpulan tersebut lintas waktu, konteks, dan kondisi algoritma. Dalam permainan berbasis AI, reliabilitas tidak dapat lagi dikonseptualisasikan hanya dalam hal konsistensi internal; sebaliknya, reliabilitas harus mencakup stabilitas temporal, ketahanan (*robustness*) di berbagai jalur adaptif, serta sensitivitas terhadap perubahan parameter model dan data pelatihan.

Sistem penilaian berbasis AI juga harus secara eksplisit memperhitungkan lintasan pembelajaran dan ketergantungan temporal. Aktivitas bermain dalam lingkungan adaptif biasanya dirancang untuk mendorong pembelajaran, eksperimen, dan penyempurnaan strategi seiring berjalannya waktu. Akibatnya, hasil pengamatan memiliki sifat saling bergantung, dan kinerja pada momen tertentu mencerminkan tingkat kemahiran saat ini

sekaligus interaksi sebelumnya dengan sistem. Karya Embretson mengenai pengujian dinamis dan pengukuran yang berorientasi pada pembelajaran memberikan landasan teoretis untuk memodelkan lintasan semacam itu; ia menekankan perlunya membedakan pertumbuhan yang bermakna dari kesalahan pengukuran serta memandang perubahan sebagai hasil substantif, bukan sekadar variabel pengganggu.

Aspek generalisasi dan keadilan menjadi sangat krusial dalam permainan berbasis AI, mengingat adanya keberagaman latar belakang pemain, gaya interaksi, dan pengalaman sebelumnya dengan teknologi digital. Model AI yang dilatih menggunakan data permainan dapat secara tidak sengaja menyerap bias sistematis yang berkaitan dengan tingkat keakraban terhadap antarmuka, konvensi permainan, atau pendekatan eksplorasi dan pengambilan risiko yang dipengaruhi oleh budaya. Teori generalisasi menawarkan kerangka kerja yang bermanfaat untuk menguraikan berbagai sumber varians seperti interaksi antara pemain dan waktu, pemain dan konteks, serta pemain dan algoritma sekaligus mengevaluasi apakah inferensi yang dihasilkan tetap stabil di berbagai populasi dan kondisi.

Aspek komparabilitas di antara berbagai kondisi permainan yang adaptif terhadap AI menghadirkan tantangan psikometris tambahan. Sistem berbasis AI sering kali menyesuaikan tingkat kesulitan, konten, atau umpan balik sebagai respons terhadap perilaku pemain, sehingga menghasilkan pengalaman yang tidak setara antarindividu. Untuk menjamin komparabilitas dalam kondisi seperti ini, diperlukan pengembangan lebih lanjut dari pendekatan tradisional terkait *linking* (penghubungan) dan *equating* (penyetaraan). Meskipun penelitian dasar mengenai *equating* telah menyediakan landasan konseptual, permainan berbasis AI menuntut metode baru yang mampu mengakomodasi adaptasi berkelanjutan, alur permainan yang bercabang, serta kumpulan tugas yang terus berkembang.

Model Inferensi untuk Data Permainan

Teori tes klasik dan teori respons butir (*item response theory*) terus memainkan peran penting dalam permainan berbasis AI, khususnya sebagai model dasar (*baseline*) atau sebagai komponen dalam sistem hibrida. Namun, asumsi-asumsi utamanya seperti observasi independen, butir soal diskret, dan parameter statis sering kali tidak terpenuhi dalam lingkungan yang dicirikan oleh interaksi berkelanjutan, pengambilan keputusan sekuensial, dan pemilihan tugas yang adaptif. Akibatnya, permainan berbasis AI semakin mengandalkan model inferensi yang lebih fleksibel dan mampu menangani data kompleks yang memiliki struktur temporal.

Jaringan Bayesian dan model diagnostik kognitif (*cognitive diagnostic models* atau CDM) terbukti sangat bermanfaat dalam konteks ini. Jaringan Bayesian mendukung integrasi berbagai sumber bukti dan memungkinkan pemodelan ketidakpastian secara eksplisit, sehingga estimasi tingkat kemahiran dapat diperbarui secara *real-time* seiring dengan diperolehnya data baru. CDM memfasilitasi inferensi diagnostik yang rinci dengan memodelkan profil penguasaan berbagai keterampilan atau atribut, yang selaras dengan tujuan instruksional dan orientasi pemberian umpan balik dalam permainan berbasis AI.

Secara paralel, algoritma pembelajaran mesin (*machine learning*) semakin banyak digunakan untuk mendeteksi pola dalam data permainan berskala besar, termasuk urutan

tindakan, informasi waktu, dan penggunaan strategi. Pendekatan-pendekatan ini dapat meningkatkan akurasi prediksi dan skalabilitas secara signifikan. Namun, dari perspektif psikometrik, pendekatan ini menimbulkan kekhawatiran penting terkait aspek keterpahaman (*interpretability*), representasi konstruk, dan generalisasi hasil. Oleh karena itu, model pembelajaran mesin harus diintegrasikan ke dalam kerangka pengukuran eksplisit yang menghubungkan keluaran algoritma dengan konstruk yang memiliki landasan teoretis, alih-alih diperlakukan sebagai prediktor mandiri yang hanya dioptimalkan untuk kinerja semata.

Ancaman terhadap Validitas dalam Permainan Berbasis AI

Permainan berbasis AI menghadirkan ancaman unik terhadap validitas, khususnya melalui varians yang tidak relevan dengan konstruk (*construct-irrelevant variance*) yang berkaitan dengan permainan dan antarmukanya itu sendiri. Faktor-faktor seperti pengalaman bermain sebelumnya, kemahiran menggunakan antarmuka, perilaku eksploratif, dan toleransi risiko dapat memengaruhi pola permainan secara sistematis, terlepas dari konstruk yang menjadi target pengukuran. Ketika model AI mempelajari pola-pola semacam itu tanpa kontrol yang memadai, terdapat risiko tercampurnya aspek keterlibatan (*engagement*), tingkat keakraban, atau preferensi gaya bermain dengan kompetensi yang sebenarnya.

Algoritma adaptif dan algoritma pembelajaran semakin memperumit masalah validitas dengan mengaburkan hubungan antara perilaku yang teramati dan konstruk yang mendasarinya. Perubahan pada tingkat kesulitan tugas, umpan balik, atau perilaku sistem dapat mengubah strategi pemain dengan cara yang sulit dipisahkan dari perubahan nyata dalam tingkat kemahiran. Sejalan dengan prinsip-prinsip yang dirumuskan dalam *Standards for Educational and Psychological Testing* (AERA, NCME, dan APA, 2014), adaptasi semacam itu menimbulkan kekhawatiran penting terkait representasi konstruk, keadilan, dan validitas interpretasi skor. Tanpa desain dan analisis yang cermat, adaptasi berbasis AI dapat memunculkan mekanisme umpan balik yang memperkuat perbedaan awal di antara peserta didik dan mengancam keadilan antar-kelompok.

Teori validitas berbasis argumen menyediakan kerangka kerja yang koheren untuk mengatasi tantangan-tantangan ini. Dari perspektif ini, validitas bergantung pada sejauh mana bukti dan teori mendukung interpretasi serta penggunaan skor dan klasifikasi yang dihasilkan oleh AI. Dalam permainan berbasis AI, hal ini menuntut adanya justifikasi eksplisit mengenai bagaimana data permainan dipilih, dikodekan, dimodelkan, dan ditransformasikan menjadi inferensi, serta transparansi terkait asumsi yang tertanam dalam algoritma AI dan mekanisme adaptifnya.

Argumen Validitas untuk Penggunaan Berbasis AI

Argumen validitas yang kokoh untuk permainan berbasis AI harus bersumber dari berbagai bukti, termasuk struktur internal, proses respons, hubungan dengan variabel eksternal, dan konsekuensi penggunaan. Bukti yang didasarkan pada struktur internal mengkaji apakah hubungan antar-indikator yang diturunkan dari AI selaras dengan ekspektasi teoretis mengenai konstruk tersebut. Bukti proses respons menyelidiki apakah proses kognitif dan perilaku yang dimunculkan oleh permainan serta disimpulkan oleh model AI sesuai dengan kompetensi yang dituju.

Hubungan dengan variabel eksternal memberikan bukti konvergen dan diskriminan dengan mengkaji asosiasi antara inferensi permainan berbasis AI dan kriteria eksternal, seperti penilaian standar, hasil pembelajaran, atau penilaian ahli. Terakhir, bukti mengenai konsekuensi membahas dampak penggunaan GBA berbasis AI, termasuk manfaat bagi pembelajaran dan pengajaran serta potensi efek yang tidak diinginkan, khususnya dalam konteks pengambilan keputusan berisiko tinggi atau otomatis.

Singkatnya, prinsip-prinsip psikometrik tetap menjadi elemen sentral dalam gim berbasis AI, namun prinsip-prinsip tersebut harus dikembangkan secara cermat untuk mengakomodasi algoritma adaptif, ketergantungan temporal, dan model inferensi yang kompleks. Melandaskan GBA (gim berbasis penilaian) yang digerakkan oleh AI pada teori pengukuran yang kokoh serta argumen validitas yang komprehensif sangatlah penting untuk memastikan bahwa inovasi tidak mengorbankan aspek keterpahaman (*interpretability*) ataupun keadilan.

8.4 TELEMETRI PERMAINAN UNTUK ANALITIK PEMBELAJARAN

Permainan edukasi digital berfungsi sebagai sistem yang dilengkapi instrumen untuk menghasilkan data telemetri yang padat dan terperinci, termasuk log dengan penanda waktu, urutan tindakan pemain, alur pengambilan keputusan, penggunaan sumber daya, dan jejak interaksi. Data semacam itu memungkinkan pengamatan perilaku belajar yang berkelanjutan dan berorientasi pada proses, serta memberikan wawasan mengenai perencanaan, penggunaan strategi, dan dinamika pemecahan masalah yang tidak dapat diperoleh melalui ukuran hasil tradisional. Metode AI sangat cocok untuk memodelkan data sekuensial berfrekuensi tinggi ini; namun, nilai analitiknya bergantung pada keselarasan yang berlandaskan prinsip dengan konstruk pembelajaran untuk memastikan bahwa inferensi yang dihasilkan bermakna, bukan sekadar deskriptif.

Data Multimodal untuk Memperkaya Bukti Pembelajaran

Selain log interaksi, analitik pembelajaran multimodal menawarkan peluang untuk memperkaya bukti mengenai proses pembelajaran dengan mengintegrasikan sinyal perilaku, fisiologis, afektif, dan sosial. Sumber data seperti arah pandangan mata, ucapan, postur tubuh, ekspresi wajah, dan pengukuran biometrik dapat dikombinasikan dengan telemetri permainan untuk melakukan triangulasi inferensi mengenai beban kognitif, keterlibatan, regulasi diri, dan kolaborasi. Jika dirancang dengan cermat, bukti multimodal dapat memperkuat validitas dengan menangkap aspek-aspek pembelajaran yang hanya dapat diamati secara tidak langsung melalui tindakan dalam permainan. Pada saat yang sama, data multimodal menghadirkan tantangan tambahan terkait sinkronisasi, derau (*noise*), privasi, dan definisi konstruk. Tanpa landasan teoretis yang eksplisit dan pra-pemrosesan yang ketat, korelasi multimodal berisiko menjadi semu atau tidak dapat ditafsirkan. Oleh karena itu, penggunaan data multimodal harus dipandu oleh tujuan pengukuran yang jelas dan diintegrasikan ke dalam model bukti yang eksplisit.

Mengintegrasikan Analitik AI dalam Desain Berbasis Bukti (ECD)

Untuk menerjemahkan analitik hasil AI menjadi bukti penilaian yang dapat dipertanggungjawabkan, metode analitik harus diintegrasikan ke dalam kerangka kerja ECD (*Evidence-Centered Design*). ECD menyediakan struktur konseptual dan inferensial yang membatasi pengembangan model dengan mengharuskan spesifikasi eksplisit mengenai klaim kompetensi, aturan bukti, dan karakteristik tugas. Integrasi analitik dalam kerangka kerja ini mendorong kemampuan interpretasi, transparansi, dan kemampuan audit, sekaligus mengurangi kekhawatiran mengenai AI yang bersifat tertutup atau "kotak hitam" (*black-box*) serta memungkinkan pelaksanaan penilaian berbasis analitik yang dapat ditingkatkan skalanya.

Tantangan Penilaian Berbasis Analitik

Terlepas dari potensinya, penilaian berbasis analitik yang didukung AI menghadapi tantangan yang terus ada. Kurangnya transparansi algoritma, keterbatasan kemampuan penjelasan (*explainability*), *overfitting* terhadap konteks permainan tertentu, serta pergeseran konstruk (*construct drift*) seiring waktu mempersulit interpretasi dan validasi. Inti dari kekhawatiran ini adalah masalah penerjemahan: prediksi yang dihasilkan AI harus memiliki landasan teoretis dan validasi empiris agar dapat berfungsi sebagai bukti penilaian, alih-alih dianggap sebagai bukti itu sendiri. Peningkatan volume data semata tidak serta-merta meningkatkan kualitas pengukuran; tanpa definisi konstruk dan pemodelan bukti yang disiplin, analitik tetap bersifat deskriptif dan bukan inferensial.

8.5 DESAIN PERMAINAN BERBASIS AI

Desain permainan berbasis AI memperluas cakupan desain permainan edukatif yang berorientasi pada pengukuran dengan cara mengintegrasikan model inferensi komputasional secara eksplisit ke dalam proses penyelarasan, pemodelan bukti, dan validasi. Dalam sistem semacam itu, AI tidak hanya mempersonalisasi pengalaman, tetapi juga berfungsi sebagai komponen inti penilaian dengan menafsirkan data permainan untuk menghasilkan inferensi mengenai kompetensi pembelajar. Oleh karena itu, desain berbasis AI yang efektif memerlukan koordinasi yang berlandaskan prinsip kuat antara tujuan pembelajaran, model bukti, strategi keterlibatan, dan alur kerja validasi.

Penyelarasan Hasil Pembelajaran

Hasil pembelajaran harus diselaraskan secara eksplisit dengan fitur dan sinyal yang digunakan oleh model AI untuk melakukan inferensi. Kompetensi perlu dioperasionalkan sebagai pola perilaku permainan yang dapat diamati, serta dapat dideteksi, direpresentasikan, dan ditafsirkan secara andal oleh model probabilistik atau model pembelajaran mesin (*machine learning*). Penyelarasan ini memastikan bahwa inferensi AI didasarkan pada tindakan yang relevan dengan konstruk, bukan sekadar artefak permainan yang bersifat insidental.

Pemodelan Bukti

Skenario permainan hanya dapat berfungsi sebagai lingkungan tugas yang dapat dipahami AI jika didasarkan pada model bukti yang eksplisit. Dengan mengacu pada pendekatan ECD (*Evidence-Centered Design*), kompetensi, perilaku yang menjadi bukti, dan

kondisi tugas harus ditetapkan terlebih dahulu agar sistem AI dapat memetakan data interaksi mentah ke konstruk laten secara sistematis dan berprinsip. Tanpa model bukti yang eksplisit, analisis AI berisiko menghasilkan inferensi yang tidak transparan atau ambigu.

Keterlibatan dan Kendali Psikometrik

Adaptasi yang didorong oleh AI harus menyeimbangkan antara motivasi pembelajar dan ketepatan psikometrik. Meskipun penyesuaian tingkat kesulitan, umpan balik, dan personalisasi dapat meningkatkan keterlibatan, elemen-elemen tersebut harus dibatasi guna menjaga komparabilitas dan kemampuan interpretasi bukti di antara para pembelajar. Adaptasi yang diatur dengan kurang baik dapat memunculkan varians yang tidak relevan dengan konstruk, sedangkan sistem adaptif yang dirancang dengan baik dapat meningkatkan kualitas bukti dengan cara mempertahankan upaya dan memunculkan pemecahan masalah yang autentik.

Siklus Validasi dan Struktur Tim

Penilaian berbasis AI memerlukan siklus iteratif yang mencakup pemodelan, pelatihan, validasi, dan revisi. Validasi tidak hanya mencakup akurasi model, tetapi juga pemeriksaan cermat terhadap pemilihan fitur, bias, dan stabilitas di berbagai konteks dan populasi. Proses-proses ini bergantung pada kolaborasi lintas disiplin antara ahli psikometri, peneliti AI, perancang gim, dan ahli ilmu pembelajaran; masing-masing memberikan keahlian khusus di bidangnya untuk memastikan sistem bersifat adaptif, menarik, dan dapat dipertanggungjawabkan secara psikometris.

8.6 CONTOH KASUS PENILAIAN BERBASIS GIM DENGAN DUKUNGAN AI

Penerapan nyata penilaian berbasis gim (GBA) dengan dukungan AI mengilustrasikan bagaimana prinsip pengukuran dan kerangka kerja desain yang dibahas pada bagian sebelumnya dapat dioperasionalkan di berbagai ranah pembelajaran. Contoh-contoh ini menunjukkan bagaimana data proses permainan yang ditafsirkan melalui model inferensi berbasis teori dapat mendukung penilaian kompetensi kompleks yang sulit ditangkap menggunakan format tradisional berbasis butir soal.

Pemecahan Masalah dan Penalaran Strategis

GBA untuk pemecahan masalah semakin mengandalkan model berbasis AI untuk menyimpulkan perencanaan strategis, fungsi eksekutif, dan penalaran adaptif peserta didik dengan menganalisis proses permainan, bukan sekadar hasil akhir. Dalam sistem semacam itu, bukti diperoleh dari urutan tindakan, pola waktu, perubahan strategi, dan respons terhadap umpan balik, yang memberikan wawasan tentang bagaimana peserta didik menghadapi masalah yang tidak terstruktur dengan jelas (*ill-structured problems*) dari waktu ke waktu. Pendekatan yang berorientasi pada proses ini selaras dengan inisiatif berskala besar seperti tugas penilaian inovatif dalam *Programme for International Student Assessment* (PISA) dari OECD, yang menekankan pemecahan masalah kompleks dalam lingkungan yang dinamis (OECD).

Dengan memodelkan perilaku pemecahan masalah sebagai sebuah lintasan (*trajectory*) dan bukan sekadar peristiwa kinerja tunggal, gim berbasis AI dapat membedakan

peserta didik yang mencapai hasil serupa melalui strategi yang secara kualitatif berbeda. Dari perspektif pengukuran, pembedaan ini sangat penting: efisiensi strategis, kemampuan adaptasi, dan kegigihan merupakan dimensi kompetensi yang bermakna secara substansial namun sering kali tidak terlihat dalam penilaian tradisional. Akan tetapi, inferensi ini bergantung pada model konstruk yang eksplisit dan pengendalian yang cermat terhadap pengaruh yang tidak relevan dengan konstruk, seperti gaya bermain eksploratif atau tingkat keakraban dengan antarmuka.

Pembelajaran Bahasa dan Literasi

Gim yang kaya akan narasi dan dialog menyediakan konteks yang menjanjikan untuk menilai keterampilan bahasa dan literasi dengan cara yang autentik dan peka terhadap konteks. Dengan dukungan teknik pemrosesan bahasa alami, permainan semacam itu dapat menganalisis bahasa lisan atau tulisan pembelajar untuk menyimpulkan praktik wacana, pemahaman bacaan, penalaran naratif, dan penggunaan bahasa secara pragmatis. Alih-alih sekadar memancing respons terisolasi terhadap instruksi yang lepas dari konteks, penilaian bahasa berbasis permainan menempatkan penggunaan bahasa dalam interaksi yang bermakna, sehingga menyelaraskan penilaian dengan praktik komunikatif.

Model inferensi berbasis AI dalam ranah ini dapat mencakup berbagai fitur seperti keragaman leksikal, kompleksitas sintaksis, koherensi, dan responsivitas terhadap isyarat percakapan. Saat diintegrasikan ke dalam kerangka kerja ECD (*Evidence-Centered Design*), fitur-fitur ini berfungsi sebagai bukti bagi konstruk linguistik yang terdefinisi dengan jelas, bukan sekadar sebagai indikator yang bersifat *ad hoc*. Pada saat yang sama, penilaian bahasa berbasis permainan memunculkan kekhawatiran khusus terkait validitas dan keadilan, termasuk sensitivitas terhadap variasi dialek, multilingualisme, dan norma komunikasi budaya. Penanganan masalah-masalah ini memerlukan pemodelan eksplisit mengenai proses respons serta validasi yang cermat di berbagai populasi pembelajar.

Pembelajaran Awal dan Penilaian yang Sensitif terhadap Perkembangan

Dalam konteks pendidikan anak usia dini, penilaian berbasis permainan (*Game-Based Assessments* atau GBA) yang didukung AI menawarkan peluang untuk mendukung pengukuran yang sesuai dengan tahapan perkembangan, sembari mengakomodasi keterbatasan terkait perhatian, komunikasi, dan variabilitas dalam lintasan perkembangan. Permainan yang dirancang untuk pembelajar usia dini dapat menyimpulkan kemajuan belajar berdasarkan pola interaksi seperti manipulasi objek, latensi respons, urutan eksplorasi, dan koreksi kesalahan alih-alih mengandalkan penjelasan verbal atau respons tertulis.

Adaptasi berbasis AI memungkinkan tugas-tugas disesuaikan secara dinamis dengan kemampuan anak yang sedang berkembang, sehingga mendukung keterlibatan sekaligus sensitivitas pengukuran. Namun, dari sudut pandang psikometrik, sistem semacam itu harus mampu membedakan secara cermat antara perubahan perkembangan dan kesalahan pengukuran, serta memperhitungkan tingginya variabilitas pada diri anak itu sendiri. Oleh karena itu, argumen validitas dalam konteks pembelajaran awal memberikan penekanan khusus pada bukti proses respons dan konsekuensi penggunaannya, mengingat sifat hasil penilaian yang formatif dan berpotensi memiliki dampak besar.

Keterampilan Sosio-Emosional dan Kolaboratif

GBA (Penilaian Berbasis Permainan) yang didukung AI semakin banyak digunakan untuk menyimpulkan kompetensi sosio-emosional seperti kegigihan, regulasi diri, kolaborasi, dan empati. Baik dalam lingkungan permainan individu maupun multipemain, model pembelajaran mesin menganalisis pola perilaku, termasuk kegigihan saat menghadapi tantangan, perilaku mencari bantuan, giliran bermain, koordinasi, dan responsivitas terhadap rekan sejawat.

Konstruk-konstruk ini pada dasarnya bersifat kontekstual dan bergantung pada situasi sosial, sehingga sulit dinilai melalui metode tradisional. Permainan menyediakan lingkungan yang terstruktur namun fleksibel, di mana perilaku sosio-emosional dapat diamati berulang kali dan dalam berbagai kondisi. Namun, terdapat tantangan interpretasi yang signifikan. Perilaku yang teramati mungkin mencerminkan strategi situasional atau norma budaya, bukan disposisi yang stabil. Akibatnya, penilaian keterampilan sosio-emosional memerlukan definisi konstruk yang sangat cermat, bukti longitudinal, dan interpretasi yang hati-hati, terutama jika digunakan untuk tujuan evaluatif atau yang memiliki konsekuensi tinggi (*high-stakes*).

8.7 ARAH MASA DEPAN PENILAIAN BERBASIS PERMAINAN

Seiring transisi GBA dari inovasi eksperimental menjadi komponen yang semakin terinstitusionalisasi dalam sistem penilaian pendidikan, fokus penelitian telah beralih dari pertanyaan mengenai kelayakan dan keterlibatan (*engagement*) menuju isu-isu kualitas, skalabilitas, integrasi, dan tata kelola. Kajian terkini mencerminkan bidang yang semakin matang, di mana ketelitian metodologis, AI, interoperabilitas sistem, dan akuntabilitas etis menjadi penentu utama kredibilitas dan adopsi.

Kualitas Penelitian dan Ketelitian Metodologis

Salah satu arah utama bagi penelitian GBA di masa depan adalah peningkatan berkelanjutan dalam ketelitian metodologis dan psikometris. Penelitian awal di bidang ini berfokus pada demonstrasi bukti konsep dan potensi motivasional (*motivational affordances*), yang penting untuk menetapkan kelayakan namun tidak memadai untuk penggunaan penilaian secara berkelanjutan. Penelitian yang lebih baru semakin menekankan pada desain berbasis teori, pemodelan konstruk yang eksplisit, dan argumen validitas yang komprehensif.

Studi kontemporer kini lebih sering menyelaraskan mekanika permainan dengan model bukti formal dan kerangka kerja penilaian yang mapan, seperti ECD. Pendekatan pengukuran tingkat lanjut termasuk pemodelan Bayesian, teknik variabel laten, dan analisis longitudinal semakin banyak diterapkan pada data permainan untuk mendukung inferensi mengenai kompetensi, ketidakpastian, dan lintasan pembelajaran. Terlepas dari kemajuan tersebut, kualitas penelitian masih belum merata; ukuran sampel yang kecil, replikasi yang terbatas, dan praktik pelaporan yang tidak konsisten membatasi kemampuan generalisasi hasil. Oleh karena itu, memperkuat bukti kumulatif melalui replikasi, berbagi dataset, dan standar pelaporan yang transparan menjadi prioritas utama penelitian.

AI Tingkat Lanjut dan Inferensi Berbasis Data

AI dan pembelajaran mesin (*machine learning*) akan terus memperluas kemampuan analitis dan skalabilitas GBA. Penambangan data pendidikan (*educational data mining*) dan analitik pembelajaran mendukung pemodelan data perilaku berfrekuensi tinggi, yang memungkinkan penilaian beralih dari hasil yang statis menuju inferensi yang dinamis dan berbasis proses. Kemajuan dalam *deep learning*, pemodelan sekuens, dan fusi data multimodal semakin memungkinkan adaptasi waktu nyata (*real-time*) serta pembuatan profil pembelajar dalam populasi berskala besar.

Pada saat yang sama, ketergantungan pada AI tingkat lanjut meningkatkan kekhawatiran mengenai interpretabilitas dan validitas. Banyak model berkinerja tinggi beroperasi sebagai sistem yang tidak transparan (*opaque systems*), sehingga menyulitkan penjabaran rantai bukti yang menghubungkan perilaku yang diamati dengan konstruk yang disimpulkan. Ketegangan antara akurasi prediksi dan kemampuan untuk memberikan penjelasan (*explainability*) ini merupakan titik balik krusial bagi bidang ini, terutama ketika GBA dipertimbangkan untuk penggunaan berisiko tinggi atau pengambilan keputusan institusional. Oleh karena itu, penelitian di masa depan kemungkinan besar akan menekankan pendekatan hibrida yang mengintegrasikan model berbasis teori dengan teknik berbasis data, serta metode AI yang dapat dijelaskan (*explainable AI*) yang disesuaikan dengan konteks penilaian.

Integrasi Sistem dan Interoperabilitas

Arah kunci lainnya berkaitan dengan integrasi GBA ke dalam ekosistem teknologi pendidikan yang lebih luas. Sistem GBA semakin banyak disematkan ke dalam sistem manajemen pembelajaran, platform analitik, dan infrastruktur data institusional, alih-alih beroperasi sebagai alat yang berdiri sendiri. Interoperabilitas mendukung akumulasi bukti longitudinal, inferensi lintas konteks, dan pengambilan keputusan instruksional yang terkoordinasi.

Penerapan standar dan protokol data bersama memungkinkan bukti yang dihasilkan dari permainan (*gameplay*) untuk digabungkan dengan penilaian tradisional, catatan instruksional, dan informasi demografis, sehingga mendukung profil pembelajar yang lebih holistik. Di sisi lain, integrasi sistem menghadirkan tantangan baru terkait kualitas data, validitas kontekstual, tata kelola, dan interpretasi lintas platform. Mengatasi tantangan ini tidak hanya memerlukan solusi teknis, tetapi juga pengembangan kapasitas organisasi dan penyelarasan kebijakan.

Etika, Keadilan, dan Tata Kelola

Pertimbangan etis muncul sebagai perhatian utama dalam fase pengembangan GBA berikutnya. Sistem-sistem ini sering kali mengumpulkan data perilaku yang rinci dan berkelanjutan terkadang tanpa sepengetahuan eksplisit pembelajar sehingga memunculkan pertanyaan mengenai persetujuan berdasarkan informasi (*informed consent*), transparansi, kepemilikan data, dan penggunaan yang tepat. Kekhawatiran semacam itu semakin menguat pada sistem yang didukung AI, di mana proses inferensi mungkin sulit untuk dijelaskan atau disanggah. Penelitian menyoroti risiko bias algoritmik, terutama ketika model dilatih menggunakan data yang mencerminkan ketimpangan historis atau ketika mekanisme

permainan memberikan keuntungan bagi pembelajar dengan latar belakang budaya atau pengalaman tertentu.

Perbedaan dalam pengalaman bermain sebelumnya, kebutuhan aksesibilitas, atau tingkat keakraban dengan konvensi permainan dapat memunculkan varians yang tidak relevan dengan konstruk (*construct-irrelevant variance*) yang pada akhirnya mengurangi aspek keadilan. Upaya di masa depan semakin menekankan pada pemodelan yang memperhatikan aspek keadilan, desain yang mempertimbangkan aksesibilitas, serta kerangka kerja tata kelola yang memperjelas akuntabilitas dan pengawasan. Memandang etika sebagai batasan desain yang mendasar bukan sekadar pertimbangan yang dipikirkan belakangan sangatlah penting untuk pengembangan skala yang bertanggung jawab.

Prioritas Penelitian dan Potensi Transformatif

Secara keseluruhan, tren-tren ini mengarah pada serangkaian prioritas penelitian yang akan membentuk arah perkembangan bidang ini. Validasi longitudinal lintas konteks, inferensi lintas permainan dan lintas platform, serta pelaporan model dan asumsi yang transparan merupakan hal krusial untuk membangun akumulasi pengetahuan. Sama pentingnya adalah pembedaan antara peningkatan berbasis AI yang bersifat inkremental terhadap praktik penilaian yang sudah ada dengan penataan ulang yang lebih transformatif mengenai bagaimana bukti pembelajaran dihasilkan, ditafsirkan, dan dimanfaatkan. Oleh karena itu, penelitian di masa depan harus mengkaji tidak hanya kemajuan teknis, tetapi juga bagaimana GBA (penilaian berbasis permainan) mengubah ekosistem penilaian secara keseluruhan. Sejauh mana GBA yang didukung AI dapat mendukung pengambilan keputusan yang valid, adil, dan dapat ditindaklanjuti dalam skala luas pada akhirnya akan menentukan perannya dalam masa depan pengukuran pendidikan.

8.8 EKOSISTEM PENILAIAN BERBASIS AI

AI memiliki potensi untuk mengubah permainan edukatif menjadi lingkungan penilaian yang adaptif dan dapat dikembangkan dalam skala luas; namun, potensi ini hanya dapat terwujud jika penilaian didasarkan secara kuat pada teori pengukuran yang sahih. GBA yang dapat dipertanggungjawabkan memerlukan konstruk yang dirumuskan dengan jelas, model inferensi yang transparan, bukti validasi yang kokoh, serta perlindungan etis yang ketat. Permainan edukatif memiliki posisi unik untuk mengintegrasikan penilaian ke dalam aktivitas yang bermakna dan menarik, sehingga memungkinkan adanya umpan balik berkelanjutan dan pembelajaran yang dipersonalisasi; namun, integrasi ini tidak boleh mengorbankan validitas, keadilan, ataupun kemampuan interpretasi hasil.

Menatap masa depan, sistem penilaian kemungkinan besar akan dibentuk oleh ekosistem berbasis AI yang memadukan pembelajaran, bermain, dan pengukuran secara harmonis. Mewujudkan visi ini menuntut kolaborasi lintas disiplin yang berkelanjutan untuk menyempurnakan teknik pemodelan, menetapkan standar validasi dan pelaporan, serta memastikan penanganan proaktif terhadap berbagai pertimbangan etis, hukum, dan sosial. Dengan memprioritaskan hal-hal tersebut, penilaian berbasis permainan (GBA) yang didukung

AI dapat berkembang dari sekadar inovasi yang menjanjikan menjadi komponen pengukuran pendidikan kontemporer yang andal, berkeadilan, dan memiliki landasan ilmiah yang kuat.

BAB 9
MODEL BAHASA BESAR DALAM PENILAIAN

9.1 APLIKASI DAN VALIDITAS LLM

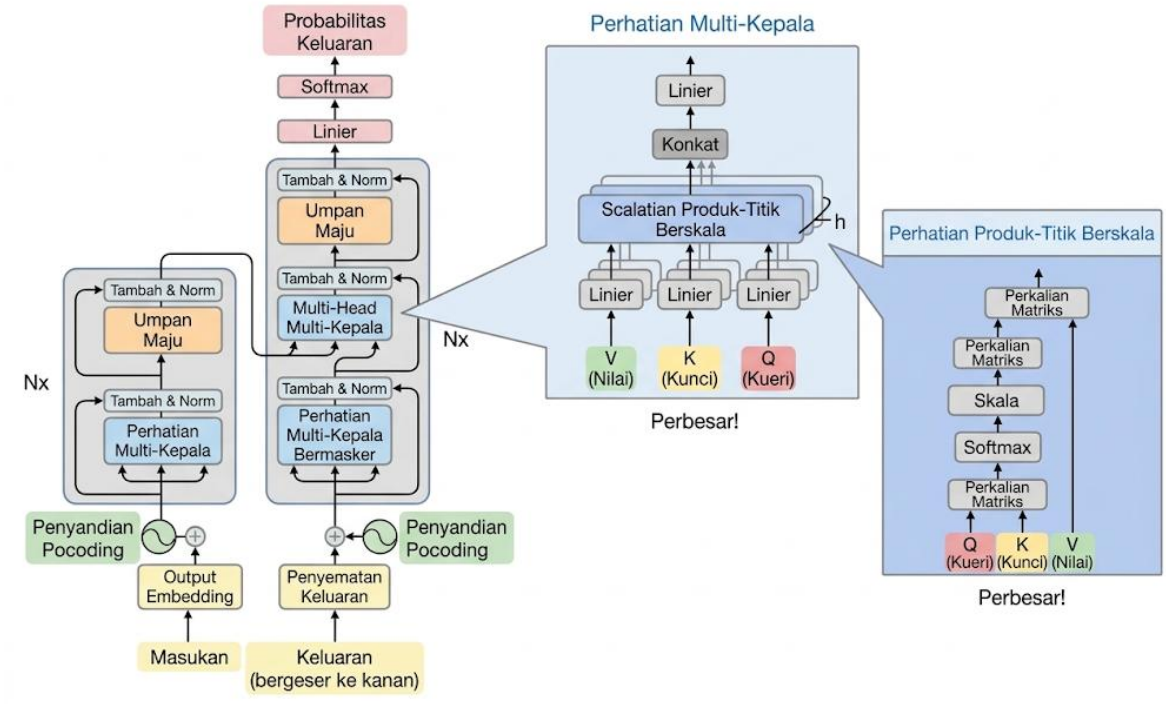
Model bahasa besar (*Large Language Models* atau LLM) telah muncul sebagai salah satu teknologi paling transformatif dalam kecerdasan buatan (AI), yang secara mendasar mengubah cara kita berinteraksi dengan mesin dan memproses informasi. Dalam bab ini, kami menyajikan gambaran umum mengenai LLM dalam penilaian pendidikan: kami membahas (1) validitas dalam konteks penerapan LLM di bidang pendidikan; (2) landasan teknis LLM; (3) bagaimana LLM diterapkan pada tiga masalah umum dalam penilaian pendidikan, yaitu penilaian otomatis (*automated scoring* atau AS), umpan balik yang dipersonalisasi, dan pembuatan butir soal otomatis (*automatic item generation* atau AIG); serta (4) bagaimana kita dapat memastikan bahwa LLM berkontribusi pada penilaian yang berkualitas tinggi.

Teori Validitas dan LLM dalam Penilaian

Integrasi LLM mengharuskan adanya landasan pada teori validitas tradisional, khususnya kerangka kerja terpadu Messick (1987) dan pendekatan berbasis argumen Kane. Penggunaan LLM mengalihkan fokus pada bagaimana model-model ini merepresentasikan konstruk pendidikan yang mendasarinya. Meskipun kita melihat bahwa LLM telah menunjukkan kemampuan luar biasa dalam logika dan pemecahan masalah, saat menerapkan LLM pada tugas akademik seperti interpretasi dan penerapan rubrik perhatian utamanya adalah representasi konstruk: memastikan LLM menangkap keterampilan yang dituju, bukan sekadar pola linguistik di tingkat permukaan. Selain itu, argumen interpretatif harus mendefinisikan secara jelas klaim apa yang dapat dibuat berdasarkan keluaran yang dihasilkan AI serta mengidentifikasi di mana varians yang tidak relevan dengan konstruk seperti bias algoritmik atau halusinasi model dapat mengganggu integritas skor tes.

9.2 LANDASAN TEKNIS UNTUK LLM

Bagian ini menyajikan tinjauan teknis mengenai LLM, mencakup landasan, kemampuan, dan pertimbangan praktisnya. Kita mulai dengan membahas arsitektur *transformer* yang mendasari model-model ini. LM (Model Bahasa) adalah fungsi yang mengestimasi probabilitas suatu urutan kata, yang memungkinkan pembuatan konten melalui prediksi kata berikutnya secara berulang. Istilah LLM biasanya merujuk pada jaringan saraf dalam (*deep neural networks*) yang berbasis arsitektur *transformer*. Bergantung pada komponen penyusunnya, model-model ini hadir dalam tiga varian utama: *encoder-only* (hanya *encoder*), *decoder-only* (hanya *decoder*), atau *encoder-decoder*. Masukan untuk setiap jenis model ditentukan oleh input embedding (yang mengodekan informasi semantik) dan *positional embedding* (yang mengodekan informasi posisi relatif), sebagaimana ditampilkan pada Gambar 9.1.



Gambar 9.1 Arsitektur *transformer*.

Mekanisme inti dari *transformer* adalah penerapan mekanisme *attention* (perhatian) secara berlapis (*layer-wise*). Berbeda dengan model jaringan saraf sebelumnya yang memproses teks secara sekuensial lalu menerapkan *attention*, penerapan *attention* secara berlapis memungkinkan model untuk secara simultan menimbang tingkat kepentingan relatif setiap kata dalam konteks masukan. Pemrosesan paralel ini jauh lebih efektif dalam menangkap dependensi jarak jauh (*long-range dependencies*) di dalam teks. Namun, hal ini juga membawa batasan mendasar: keseluruhan urutan masukan harus diproses sekaligus. Semua LLM berbasis *transformer* memiliki ukuran masukan maksimum yang tetap, yang dikenal sebagai panjang konteks (*context length*). LLM generatif modern dilatih melalui tiga tahapan berikut:

1. **Pra-pelatihan (*Pretraining*):** Bobot model dioptimalkan untuk memprediksi kata berikutnya dalam korpus teks berskala besar yang tidak berlabel.
2. ***Instruction Tuning* (Penyesuaian Instruksi)**: Bobot model kemudian disesuaikan lebih lanjut (*fine-tuned*) untuk menyelesaikan sekumpulan instruksi (*corpus*).
3. **Penyempurnaan (*Refinement*):** Bobot model kemudian disesuaikan lebih lanjut agar lebih mengutamakan keluaran yang telah diberi label.

Dua tahap pertama bergantung pada varian *stochastic gradient descent*, seperti Adam. Tahap pertama sering disebut sebagai pembelajaran semi-terawasi (*semi-supervised*), karena prediksi kata berikutnya merupakan tugas pembelajaran terawasi (*supervised*) pada data yang tidak berlabel; data tidak berlabel ini tidak harus mengikuti format tertentu dan sering kali mencakup materi akademik, konten berbasis web, prosa, dialog, serta kode pemrograman atau matematika. Tahap kedua, yaitu *instruction tuning*, menggunakan kumpulan instruksi

yang biasanya berbentuk dialog antara pengguna dan asisten; pengguna memberikan tugas beserta masukan yang relevan, dan asisten menyelesaikan tugas tersebut termasuk kemungkinan adanya tugas lanjutan dan penyelesaiannya seperti yang ditunjukkan pada Tabel 9.1.

Tabel 9.1 Contoh Templat Obrolan

Pengguna	Apa ibu kota Prancis?
Asisten	Ibu kota Prancis adalah Paris.
Pengguna	Berapa jumlah penduduknya?
Asisten	Kota Paris memiliki populasi sekitar 2,1 juta jiwa. Namun, wilayah metropolitan yang lebih luas—sering disebut sebagai Grand Paris atau wilayah Île-de-France—memiliki populasi lebih dari 12 juta jiwa.

Tahap akhir, yaitu proses penyempurnaan, menggunakan teknik pembelajaran penguatan (*reinforcement learning*) seperti *direct preference optimization* (DPO). Teknik-teknik ini dapat digunakan untuk memodifikasi bobot model agar keluarannya lebih bermanfaat dan tidak berbahaya. Masukan untuk metode ini berupa pasangan preferensi: pengguna menentukan suatu tugas, lalu ditampilkan beberapa versi penyelesaian tugas tersebut, dengan satu versi yang lebih disukai. Preferensi ini dikodekan sebagai fungsi imbalan (*reward function*), dan bobot model diperbarui untuk meningkatkan kemungkinan munculnya keluaran yang lebih disukai tersebut.

Kemampuan LLM Generatif

Pertama, LLM secara alami memiliki serangkaian kemampuan yang ditentukan oleh tahap pelatihan *instruction-tuning*; dengan demikian, model ini umumnya dapat menyelesaikan tugas-tugas dalam kategori yang tercakup dalam *corpus* yang digunakan pada tahap tersebut. *Corpus* ini biasanya terdiri dari tugas-tugas dalam kategori berikut:

1. **Pemahaman Bahasa Alami (*Natural Language Understanding*):** Contohnya mencakup tugas klasifikasi seperti analisis sentimen atau deteksi toksisitas, serta tugas tanya-jawab seperti pertanyaan pilihan ganda (MCQ) dan soal jawaban singkat berbasis pengetahuan.
2. **Pembangkitan Bahasa Alami (*Natural Language Generation*):** Ini mencakup tugas peringkasan dan pembuatan konten, seperti menulis cerita, puisi, dan esai.
3. **Pengambilan Informasi:** Mengidentifikasi informasi spesifik, seperti nama, tanggal, dan tempat, dari sebuah teks.
4. **Logika dan Pemrograman:** Penyusunan penalaran langkah demi langkah serta pemecahan masalah menggunakan berbagai bahasa pemrograman.
5. **Lainnya:** Ini mencakup transfer gaya, penerjemahan, dan integrasi data.

Salah satu temuan kunci mengenai LLM adalah bahwa peningkatan skala ukurannya umumnya memperbaiki kemampuan model dalam menyelesaikan tugas-tugas yang belum pernah ditemui sebelumnya. Karakteristik ini memunculkan gagasan tentang "kemampuan yang muncul" (*emergent ability*), yaitu kemampuan apa pun yang tidak dimiliki oleh model berukuran lebih kecil tetapi dimiliki oleh model berukuran lebih besar yang dilatih dengan cara

yang sama. Contoh paling menonjol dari kemampuan yang muncul ini dikenal sebagai pembelajaran *zero-shot*, *one-shot*, dan *few-shot*. Kemampuan-kemampuan ini mengacu pada penyelesaian tugas tanpa contoh, dengan satu contoh, atau dengan beberapa contoh. Kita dapat melihat peningkatan luar biasa dalam klasifikasi teks dengan jumlah contoh yang sangat sedikit yang kita anggap sebagai kemampuan yang muncul karena peningkatan semacam itu tidak teramati pada model yang jauh lebih kecil.

Para peneliti telah berupaya mengevaluasi kemampuan dan keterbatasan LLM menggunakan serangkaian tolok ukur (*benchmark*), yang didefinisikan oleh kumpulan data dan metrik evaluasi. Beberapa tolok ukur yang paling umum digunakan meliputi *Massive Multitask Language Understanding*, *Graduate-Level Google-Proof* Q&A *Benchmark* (GPQA), HumanEval, MATH, dan *Multilingual Grade School Math* (MGSM). Beberapa tolok ukur untuk model-model populer ini disajikan dalam Tabel 9.2, dengan kinerja yang diukur berdasarkan akurasi, yang dinyatakan sebagai persentase tugas yang diselesaikan dengan benar.

Tabel 9.2 Daftar Tolok Ukur LLM

Model	MMLU	GPQA	MATH	HumanEval	MGSM
Claude 3.5 Sonnet	88,3	59,4	71,1	92,0	91,6
GPT-4.5	90,8	69,5	87,1	88,0	86,9
Gemini 2.5 Pro	84,1	80,3	95,2	82,0	92,2
Grok 3	81,4	84,0	94,2	86,5	91,3

Tinjauan terhadap pola-pola ini menunjukkan bahwa meskipun model seperti GPT-4.5 dan Gemini 2.5 Pro memperlihatkan kemahiran tinggi dalam tolok ukur yang menuntut kemampuan penalaran kuat seperti MATH, kinerja model-model tersebut sangat bervariasi di berbagai domain, seperti pemrograman (HumanEval) dan pengetahuan tingkat pascasarjana (GPQA). Perbedaan-perbedaan ini menyoroti bahwa tidak ada satu model pun yang unggul dalam semua tugas penilaian, sehingga menegaskan perlunya pemilihan model yang spesifik untuk domain tertentu. Terakhir, perlu dicatat bahwa banyak model dilatih untuk mengintegrasikan informasi dari sumber eksternal; teknik ini dikenal sebagai *retrieval-augmented generation* (RAG). RAG merupakan metode untuk memperluas kemampuan LLM dalam memanfaatkan informasi terkini yang belum tentu tersedia pada saat pelatihan berlangsung.

Sebagai catatan penyesuaian bahasa, tanda desimal titik pada angka telah diubah menjadi koma sesuai dengan format penulisan angka dalam bahasa Indonesia. Judul kolom pertama (Model) ditambahkan untuk memperjelas, sementara nama model AI dan istilah tolok ukur pengujian (MMLU, GPQA, MATH, HumanEval, MGSM) tetap menggunakan istilah aslinya karena merupakan nama baku/teknis.

Keterbatasan dan Kendala dalam Penggunaan LLM Generatif

Keterbatasan pertama yang kami sebutkan adalah bahwa arsitekturnya membatasi panjang input. Pada LLM berbasis *transformer* awal seperti *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin d, panjang urutan maksimumnya adalah

512, yang membatasi penerapannya. LLM generatif modern mengatasi keterbatasan ini dengan menggunakan *rotational positional embeddings*, yang memungkinkan pelatihan lebih lanjut pada konteks yang panjang.

Meskipun panjang konteks yang dihasilkan tetap terbatas, dalam praktiknya, panjang tersebut sudah cukup memadai untuk sebagian besar aplikasi. *Relative Effective Context Length* (RECL), yang diperkenalkan oleh Dai, merupakan ukuran yang lebih praktis, terutama karena ukuran ini mencerminkan kemampuan model dalam mengekstrak informasi di sepanjang konteks yang panjang. Salah satu tolok ukur RECL yang populer adalah uji "jarum di tumpukan jerami" (*needle in a haystack*), yang menilai kemampuan LLM untuk mengambil informasi penting dari teks pengalih perhatian (*distractor*) yang panjang.

Penggunaan RAG juga krusial dalam mengatasi dua keterbatasan LLM lainnya: batas pengetahuan (*knowledge cutoff*) saat pra-pelatihan dan halusinasi. Batas pengetahuan mengacu pada fakta bahwa LLM bersifat statis; dengan demikian, bobotnya tidak berubah akibat pemberian *prompt*. LLM hanya mengetahui peristiwa yang dipelajari selama pra-pelatihan dan tidak akan mengetahui peristiwa terkini secara alami kecuali jika digunakan bersamaan dengan RAG.

Kendala kedua, yang sebagian diatasi oleh RAG, adalah kemungkinan LLM menyajikan konten yang tidak akurat sebagai kebenaran yang dikenal sebagai halusinasi factual serta logika yang keliru dalam tugas-tugas bertahap, yang disebut sebagai halusinasi penalaran. Beberapa contoh terkenal halusinasi faktual mencakup model yang menghasilkan kasus hukum fiktif sebagai preseden atau mengutip artikel jurnal yang tidak ada sebagai bukti konsep atau fakta tertentu. Contoh halusinasi penalaran meliputi penjelasan yang tidak akurat mengenai alasan esai Mahasiswa mendapatkan nilai tertentu atau penalaran yang tidak valid mengenai alasan jawaban tertentu dianggap benar. Beberapa peneliti berupaya mengukur tingkat halusinasi, sementara peneliti lain berupaya mendeteksinya.

Prompt Tuning dan Fine-Tuning

Karena kualitas instruksi memengaruhi kinerja LLM, *prompt engineering* merupakan praktik penyusunan *prompt* yang efektif untuk memaksimalkan kinerja tugas. Upaya peningkatan ini sering kali melibatkan pembuatan templat *prompt* yang dapat digunakan kembali, yang memuat instruksi, ketentuan, serta *placeholder* untuk konten dinamis. Untuk mengotomatisasi proses ini, kerangka kerja seperti DSPy (yang bersifat sumber terbuka/ *open-source*) menggunakan pengoptimal untuk menyempurnakan *prompt* secara sistematis berdasarkan metrik kinerja, sementara platform tertutup (*closed-source*) seperti OpenAI dan Anthropic menawarkan perangkat kepemilikan serupa. Akan tetapi, LLM tertutup sering kali diperbarui secara internal, yang menyebabkan terjadinya pergeseran model (*model shift*). Akibatnya, efektivitas *prompt* dapat berubah saat digunakan pada model tertutup yang telah diperbarui, sehingga menghasilkan keluaran yang tidak konsisten dan reproduksibilitas yang rendah.

Ketika *prompt engineering* gagal memenuhi standar validasi dan reproduksibilitas, langkah selanjutnya adalah melakukan *fine-tuning* pada model sumber terbuka (*open-source*). LLM sumber terbuka dapat disimpan secara lokal dengan bobot yang dibekukan (*frozen*),

sehingga menjadikannya lebih andal dan konsisten. Namun, *fine-tuning* penuh membutuhkan sumber daya komputasi yang besar; sebagai contoh, model dengan 1 miliar parameter memerlukan memori GPU lebih dari 16 GB saat menggunakan pengoptimal (*optimizer*) standar, yang melampaui kapasitas sebagian besar perangkat keras kelas konsumen.

Untuk mengatasi tantangan ini, para peneliti mengembangkan metode *fine-tuning* yang efisien dalam penggunaan parameter atau *parameter-efficient fine-tuning* (PEFT). Teknik yang paling populer, yaitu *low-rank adaptation* (LoRA), bekerja dengan cara membekukan bobot asli model dan hanya melatih sekumpulan kecil lapisan baru yang disebut lapisan "adapter". Praktik ini secara drastis mengurangi kebutuhan memori dan komputasi. Ketika LoRA dikombinasikan dengan kuantisasi teknik untuk menyimpan bobot model pada tingkat presisi yang lebih rendah metode tersebut dikenal sebagai QLoRA. QLoRA memungkinkan dilakukannya *fine-tuning* pada model berskala besar menggunakan perangkat keras kelas konsumen.

9.3 PENDEKATAN GENERATIF UNTUK PENILAIAN OTOMATIS

Penilaian otomatis (*Automated Scoring* atau AS) merupakan alternatif yang hemat biaya dibandingkan penilaian esai dan jawaban singkat yang dibuat sendiri oleh Mahasiswa (*short-constructed responses*) secara manual. *Prompt engineering*, strategi *fine-tuning*, dan teknik augmentasi data kemudian menarik minat besar di kalangan komunitas penilaian. Bagian ini menyajikan gambaran umum yang membandingkan metode penilaian sebelumnya dengan pendekatan yang menggunakan LLM generatif.

Perbandingan dengan Pendekatan Penilaian Sebelumnya

Pendekatan paling awal terhadap AS menggunakan fitur-fitur yang dipilih secara manual untuk menentukan kualitas jawaban mahasiswa. Fitur-fitur tersebut mencakup panjang jawaban, kompleksitas sintaksis, kosakata, keterbacaan, gaya bahasa, argumentasi, relevansi dengan instruksi (*prompt*), dan lain sebagainya. Model pembelajaran mesin tradisional menggunakan fitur-fitur yang dirancang secara manual untuk memprediksi nilai mahasiswa. Namun, pembuatan fitur-fitur ini memakan biaya dan waktu yang besar, serta memerlukan eksperimen ekstensif untuk mengoptimalkan kinerja. Fitur-fitur tersebut juga sangat bergantung pada konteks; fitur yang digunakan untuk soal matematika berbeda secara signifikan dengan fitur yang diperlukan untuk penilaian esai. *Embedding* kata dan dokumen digunakan sebagai alternatif bagi fitur-fitur yang dirancang secara manual. Ini adalah representasi vektor kata atau dokumen yang mampu menangkap makna semantik berdasarkan konteks statistik kata. Model yang dilatih menggunakan urutan *word embedding* dari korpus besar respons mahasiswa menghasilkan representasi vektor untuk menangkap hubungan semantik, yang kemudian digunakan sebagai prediktor skor. Metode untuk menghitung *embedding* kata dan dokumen meliputi Word2Vec, GloVe, dan FastText. Berbagai penelitian menunjukkan bahwa model *convolutional neural network* dan *recurrent neural network* yang dilatih menggunakan metode-metode tersebut memiliki kinerja lebih unggul dibandingkan pendekatan awal yang mengandalkan rekayasa fitur.

Model BERT menangkap dependensi jarak jauh dan representasi semantik menggunakan mekanisme *attention*. Model ini dilatih sebelumnya (*pre-trained*) menggunakan 3,3 miliar token melalui tugas *masked language modeling* dan prediksi kalimat berikutnya (*next-sentence prediction*), dengan mempelajari konteks dua arah (*bidirectional*) untuk melakukan prediksi. Untuk AS (*Automated Scoring*), lapisan klasifikasi tambahan ditambahkan pada BERT dan dilakukan fine-tuning untuk keperluan penilaian. Varian dari pendekatan ini telah mencapai kinerja terbaik saat ini (*state-of-the-art*) dengan kemampuan generalisasi lintas jenis butir soal yang lebih baik dibandingkan pendekatan berbasis rekayasa fitur. Namun, proses *fine-tuning* tetap membutuhkan dataset yang besar dan mahal biayanya untuk mencapai kinerja optimal.

Pendekatan Penilaian *Zero-Shot* dan *Few-Shot*

LLM generatif mampu menandingi pendekatan berbasis BERT tanpa perlu *fine-tuning*, dan bahkan berpotensi melampauinya setelah melalui proses fine-tuning. Peningkatan kinerja ini berasal dari dataset pelatihan awal yang jauh lebih besar Llama 3 menggunakan 15 triliun token dibandingkan 3,30 miliar token pada BERT yang menyediakan basis pengetahuan lebih luas, serta dari ukuran model yang jauh lebih besar Llama 3 memiliki 70 miliar parameter dibandingkan 330 juta parameter pada BERT. Berbeda dengan metode *masked language modeling* pada BERT, LLM generatif memanfaatkan informasi tambahan dalam *prompt*, yang memungkinkan dilakukannya prediksi *zero-shot* dan *few-shot*. Pelatihan awal ini menghasilkan model AS yang tidak memerlukan *fine-tuning* tambahan.

Mizumoto dan Eguchi memperoleh hasil yang menjanjikan saat menggunakan GPT-3.5 untuk penilaian esai bahasa kedua. Hou menunjukkan bahwa *prompt zero-shot* yang menyertakan fitur linguistik tambahan (seperti *lemma*, kata benda, dan jumlah kalimat) secara konsisten memberikan hasil lebih baik dibandingkan *prompt* tanpa fitur tersebut, bahkan menyamai kinerja dasar (*baseline*) BERT saat diuji menggunakan model GPT-4 serta Mistral 7B. Yoshida menemukan bahwa GPT-3.5 memberikan kinerja terbaik dengan *prompt few-shot*, sementara GPT-4.0 menunjukkan peningkatan minimal di berbagai jenis *prompt*. Stahl dkk. menguji penggunaan berbagai persona (asisten pengajar, peneliti pendidikan, anggota komunitas penulisan kreatif) pada model Mistral 7B. Mereka menemukan bahwa persona peneliti akademis dan asisten pengajar mencapai kinerja optimal, dengan peningkatan lebih lanjut yang diperoleh melalui contoh-contoh tambahan. Wang dan Ormerod serta Saad-Falcon dkk. menunjukkan kinerja yang sangat baik dengan menggunakan teknik *few-shot prompting* yang dikombinasikan dengan alur kerja RAG.

Penalaan Halus (*Fine-Tuning*) untuk Penilaian Otomatis

Penalaan halus (*fine-tuning*) pada LLM generatif untuk penilaian otomatis (AS) merupakan pendekatan yang belum banyak dieksplorasi dibandingkan dengan prediksi *single-shot* atau *few-shot*, yang tidak memerlukan pelatihan tambahan. Keterbatasan ini terutama disebabkan oleh waktu dan biaya yang diperlukan untuk melakukan penalaan halus pada LLM terkini, mengingat ukurannya yang besar. Namun, studi mengenai AS yang menerapkan bentuk penalaan halus tertentu memberikan tolok ukur (*baseline*) yang sangat baik, melampaui hasil dari pendekatan *zero-shot* dan *few-shot*.

Ormerod dan Kwako serta Katuka, Gain, dan Yu menyajikan beberapa kasus di mana metode PEFT, seperti QLoRA, yang diterapkan pada model sumber terbuka (*open-source*) seperti Llama menghasilkan kinerja lebih unggul dibandingkan model *single-shot* yang jauh lebih besar maupun tolok ukur BERT. Meskipun penalaan halus umumnya diterapkan pada model sumber terbuka, Latif dan Zhai melakukan penalaan halus pada GPT-3.5 untuk menilai berbagai jenis pertanyaan sains secara otomatis. Mereka melatih GPT-3.5 selama beberapa *epoch* menggunakan sekitar 7.000 respons mahasiswa, yang masing-masing telah memiliki tingkat nilai (*grade level*) yang ditetapkan sebelumnya.

Augmentasi Data untuk Penilaian

Augmentasi data adalah proses penambahan data yang dialihfungsikan atau data sintetis ke dalam dataset yang sudah ada yang digunakan untuk penyempurnaan (*fine-tuning)* model guna meningkatkan kinerja model. Data sintetis dapat berupa data yang dibuat melalui cara artifisial, seperti prediksi dari model lain atau konten yang dihasilkan oleh AI. Dalam jumlah yang memadai, data sintetis dapat meningkatkan kinerja model serta dapat dimanipulasi untuk mengurangi bias model atau karakteristik lain yang tidak diinginkan.

Koreksi kesalahan tata bahasa (*Grammatical Error Correction* atau GEC) merupakan salah satu penerapan awal augmentasi data dalam pelatihan model generatif. Dengan menerapkan penerjemahan dua arah (*round-trip translation* atau RTT), Stahlberg dan Kumar membuat korpus besar untuk melatih model T5 bagi keperluan GEC. Bayer dkk. menunjukkan bahwa, jika diterapkan secara cermat, RTT juga dapat memberikan manfaat dalam masalah klasifikasi teks terutama ketika skema klasifikasi tidak berubah akibat parafrase meskipun penerapannya dalam bidang penilaian sejauh ini belum menunjukkan keberhasilan yang serupa. *North dan Ormerod* menggunakan data sintetis untuk meningkatkan kinerja penilaian otomatis (AS) pada korpus PERSUADE. Mereka menunjukkan bahwa pelatihan LLM tradisional, seperti ELECTRA, dapat dilakukan dengan jumlah data yang jauh lebih sedikit. Melalui penyempurnaan (*fine-tuning*) model Llama-3 hanya pada 10% dari korpus PERSUADE asli dan kemudian menggunakan Llama-3 untuk memprediksi nilai mahasiswa pada 90% sisa korpus pelatihan tersebut, model ELECTRA menunjukkan kinerja yang jauh lebih baik dibandingkan saat dilatih hanya dengan 10% data asli. Hasil ini secara efektif menunjukkan bahwa penggunaan model generatif dalam alur kerja distilasi model hanya memerlukan sedikit data yang telah dinilai untuk mencapai kinerja AS yang setara dengan model penilaian tradisional.

Tantangan Implementasi dalam Penilaian Berskala Besar

LLM generatif merupakan alternatif yang menarik untuk penilaian otomatis (AS) karena kinerjanya dapat menandingi metode-metode sebelumnya tanpa memerlukan proses penyempurnaan (*fine-tuning*) yang memakan biaya besar pada dataset khusus. Selain itu, jika disempurnakan, model-model ini mampu menetapkan standar kinerja terbaik yang baru (*state-of-the-art*), khususnya untuk tugas-tugas kompleks seperti penilaian esai (Ormerod & Kwako). Terlepas dari berbagai keunggulan tersebut, penerapan LLM dalam penilaian berskala besar menghadirkan tantangan signifikan terkait bias, privasi data, dan kebutuhan sumber daya komputasi yang besar untuk pengoperasiannya.

Salah satu tantangan utama terkait keadilan muncul dari temuan *North* dan *Ormerod* bahwa skor yang dihasilkan LLM sering kali memiliki distribusi yang berbeda dibandingkan skor yang diberikan oleh pakar manusia. Penyimpangan sistemik ini dapat memunculkan bias baru yang tidak ditemukan dalam penilaian oleh manusia, sehingga menimbulkan kekhawatiran serius mengenai kesetaraan. Untuk mengatasi hal ini, para peneliti sedang menjajaki teknik *fine-tuning* tingkat lanjut seperti DPO, yang bertujuan menyelaraskan keluaran model agar lebih sesuai dengan penilaian dan preferensi manusia.

Privasi dan keamanan merupakan hambatan besar lainnya, terutama pada model *closed-source* (sumber tertutup). Karena model-model ini dilatih menggunakan data yang tidak diungkapkan, pengiriman informasi mahasiswa yang bersifat rahasia atau rubrik penilaian internal ke dalam model tersebut menimbulkan risiko kebocoran data yang tidak dapat diterima dalam konteks ujian yang krusial. Model *open-source* (sumber terbuka) menawarkan solusi langsung untuk masalah ini, karena model tersebut dapat di-hosting dan dilatih secara aman di lingkungan internal organisasi. Namun, kendali penuh ini menuntut sumber daya komputasi yang sangat besar. Meskipun teknik seperti LoRA dan QLoRA dapat mengurangi kebutuhan sumber daya GPU, penerapan praktisnya memerlukan keahlian teknis khusus.

9.4 MENGHASILKAN UMPAN BALIK YANG DIPERSONALISASI

Umpan balik terdiri dari verifikasi (misalnya, skor) dan elaborasi (misalnya, panduan yang dipersonalisasi untuk membantu pembelajar melakukan perbaikan) (Kulhavy & Stock). Meskipun panduan ini dapat berasal dari berbagai sumber, LLM dapat mengotomatisasi pembuatannya, serta memberikan masukan yang rinci dan tepat waktu yang sangat penting bagi pembelajaran yang efektif. Namun, karena sekadar menyajikan umpan balik jarang cukup untuk mendukung pembelajaran (Hattie & Timperley), umpan balik otomatis akan paling efektif jika diintegrasikan ke dalam perangkat pendidikan yang lebih luas, seperti penilaian formatif atau sistem tutor cerdas (*Intelligent Tutoring Systems* atau ITS). Bagian ini akan memberikan gambaran umum mengenai umpan balik yang efektif, meninjau bagaimana LLM menghasilkannya, serta membahas integrasinya ke dalam perangkat yang dirancang untuk memfasilitasi pembelajaran.

Apa yang Membuat Umpan Balik Menjadi Efektif?

Penelitian menunjukkan bahwa umpan balik yang efektif tidak bersifat seragam untuk semua situasi; dampaknya bergantung pada interaksi yang kompleks antara pembelajar, konteks, dan tugas. Saat merancang sistem umpan balik otomatis, sangatlah penting untuk mempertimbangkan tidak hanya umpan balik apa yang akan diberikan, tetapi juga kepada siapa dan dalam kondisi apa umpan balik tersebut diberikan. Dengan mempertimbangkan hal tersebut, kami menyoroti beberapa temuan yang sangat relevan dengan umpan balik otomatis yang dipersonalisasi:

- Umpan balik paling efektif jika diberikan tepat waktu, namun tidak terlalu cepat hingga mengganggu proses berpikir pembelajar. Idealnya, pembelajar harus bersikap terbuka terhadap umpan balik atau secara aktif mencarinya.

- Spesifisitas itu penting, namun jenisnya bergantung pada pembelajar. Meskipun umpan balik yang spesifik selalu lebih bermanfaat daripada saran umum, pendekatan terbaiknya bervariasi. Umpan balik direktif (memberikan jawaban yang benar) sering kali paling baik bagi pemula, sedangkan umpan balik fasilitatif (memberikan panduan tanpa langsung memberikan jawaban) lebih bermanfaat bagi pembelajar tingkat lanjut.
- Keterlibatan merupakan hal yang esensial. Agar umpan balik efektif, mahasiswa harus memiliki kesempatan untuk terlibat secara aktif dengannya, dan menerapkannya pada tujuan pembelajaran mereka. Umpan balik yang hanya berupa pengingat sederhana sering kali diabaikan dan kecil kemungkinannya menghasilkan peningkatan.
- Umpan balik memiliki dimensi emosional. Meskipun dimaksudkan sebagai umpan balik formatif, hal tersebut dapat mematahkan semangat serta berdampak negatif terhadap motivasi dan kinerja mahasiswa jika tidak disampaikan dengan penuh kehati-hatian.

Menghasilkan Umpan Balik dengan LLM

Umpan balik otomatis dapat diperoleh dari LLM melalui dua metode utama: memberikan instruksi (*prompting*) secara langsung kepada model generatif atau melakukan *fine-tuning* (penyesuaian halus) pada model untuk tugas spesifik terkait umpan balik. Kedua pendekatan ini dapat meningkatkan sistem umpan balik yang sudah ada serta memungkinkan bentuk-bentuk panduan pendidikan yang benar-benar baru. Sebagaimana disebutkan sebelumnya, *prompting* adalah metode di mana LLM diberikan instruksi dan hasil pekerjaan Mahasiswa untuk menghasilkan umpan balik. Strategi tingkat lanjut seperti *chain-of-thought prompting* (meminta model menjabarkan langkah penalaran), menetapkan persona ahli, atau menggunakan LLM kedua untuk mengevaluasi umpan balik tersebut (pendekatan "LLM-sebagai-penilai" atau *LLM-as-judge*) dapat meningkatkan kualitas keluaran.

Sebagai alternatif, LLM dapat disesuaikan (*fine-tuned*) untuk mengidentifikasi elemen wacana seperti "klaim" dan "bukti" dalam esai mahasiswa. Prediksi model tersebut kemudian dapat digunakan untuk menghasilkan umpan balik yang sangat terarah, seperti menyarankan mahasiswa untuk menambahkan lebih banyak bukti guna mendukung klaim atau menulis ulang kalimat agar argumen menjadi lebih kuat. Setiap metode memiliki konsekuensi pertukaran (*trade-off*) yang jelas. *Prompting* lebih mudah diterapkan pada tahap awal namun dapat menghasilkan keluaran yang tidak dapat diandalkan atau tidak terduga. Selain itu, karena sifat LLM yang terbuka (*open-ended*), kualitasnya sulit divalidasi secara otomatis dan mungkin lebih sulit untuk diintegrasikan ke dalam perangkat pendidikan. Sebaliknya, *fine-tuning* memerlukan upaya awal yang signifikan, seperti pembuatan kumpulan data dan keahlian teknis, namun menghasilkan umpan balik yang lebih konsisten, dapat diandalkan, dan terarah, serta lebih mudah dievaluasi dan diselaraskan dengan tujuan pendidikan tertentu.

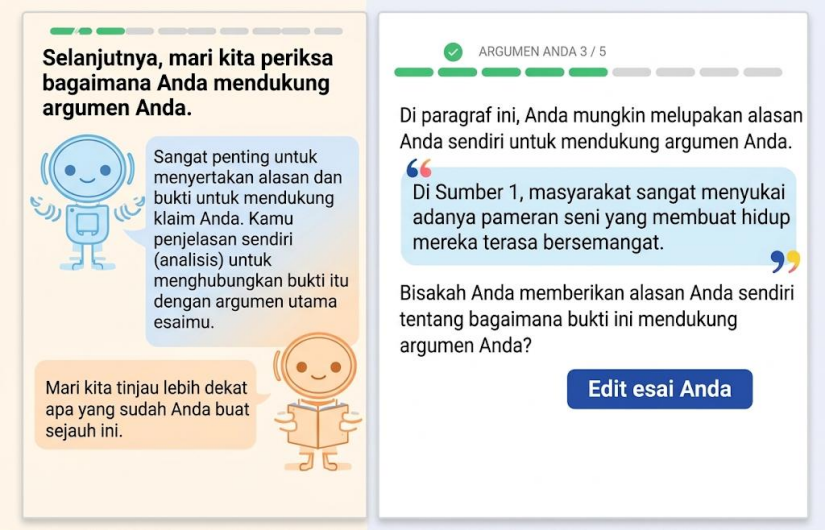
Penerapan Umpan Balik Otomatis yang Dipersonalisasi

Meskipun sistem tutor cerdas (*Intelligent Tutoring Systems* atau ITS) generasi awal telah menyediakan umpan balik otomatis yang canggih, sistem tersebut tidak memiliki fleksibilitas yang ditawarkan oleh LLM. *Write On with Cambi!* sebuah alat penilaian formatif, merupakan contoh bagaimana model bahasa besar (*Large Language Models* atau LLM) yang

telah disesuaikan secara khusus (*fine-tuned*) dapat membantu mahasiswa meningkatkan kemampuan menulis argumentatif mereka.

Dalam alat ini, mahasiswa mengumpulkan esai dan dipandu melalui proses revisi terstruktur yang terdiri dari beberapa tahap. Serangkaian model yang telah disesuaikan tersebut terlebih dahulu mengevaluasi komponen tingkat tinggi, dengan memeriksa elemen-elemen kunci seperti klaim utama, bukti pendukung, dan kesimpulan. Hanya setelah aspek-aspek struktural ini ditangani, alat tersebut memberikan saran untuk perbaikan pada tingkat permukaan, seperti ejaan dan tata bahasa. Proses ini bersifat interaktif; sistem menunggu instruksi dari mahasiswa sebelum beralih ke tahap berikutnya.

Sebagai platform pendidikan interaktif yang disediakan oleh Cambium Assessment, *Write on with Cambi* menerapkan beberapa prinsip umpan balik yang efektif. Desainnya berpusat pada mahasiswa: umpan balik diberikan tepat waktu namun tidak mengganggu alur kerja, sehingga mahasiswa dapat mengatur kecepatannya sendiri. Saran-saran yang diberikan bersifat spesifik dan dapat langsung diterapkan, sering kali dengan menyoroti bagian teks atau lokasi tertentu yang perlu direvisi. Dengan mengikuti urutan yang didasarkan pada prinsip pedagogis yang tepat, alat ini mendorong keterlibatan yang mendalam dan membantu mahasiswa memperbaiki proses penulisan mereka secara terstruktur. Contoh 9.1 menampilkan contoh umpan balik yang berpusat pada mahasiswa yang diberikan oleh *Write on with Cambi*.



9.5 MENGHASILKAN BUTIR SOAL DAN KONTEN

AIG memiliki sejarah yang panjang. Sebagian besar penelitian terkini mengenai AIG berfokus pada pembuatan soal pilihan ganda (MCQ) berbasis teks. Namun, terdapat pula penelitian mengenai konten tes yang lebih luas, seperti pembuatan diagram untuk soal matematika dan pembuatan teks informatif untuk soal pemahaman bacaan. Dalam asesmen berisiko tinggi (*high-stakes assessments*), AIG terutama digunakan untuk mengisi kumpulan butir soal (*item pools*), yang kemudian diuji coba di lapangan sebelum digunakan. LLM tidak hanya digunakan untuk menghasilkan butir soal. LLM juga banyak dimanfaatkan dalam tahap

pasca-pembuatan untuk menemukan pengecoh (*distractor*) yang tepat atau menyaring butir soal yang tidak memenuhi kriteria tertentu; selain itu, LLM terkadang digunakan untuk tahap pra-pembuatan atau penambangan data.

Perbandingan dengan Pendekatan Berbasis Templat

Secara tradisional, AIG mengandalkan pendekatan berbasis templat yang menggunakan model butir soal sangat terstruktur dengan komponen modular (misalnya, 5 + 3 = _ untuk penjumlahan satu digit). Templat menghasilkan keluaran yang dapat diprediksi dan terkendali. Sebaliknya, banyak pendekatan modern menggunakan metode berbasis probabilitas (seperti LLM) untuk menghasilkan konten. Dalam penggunaan LLM, hal ini biasanya dilakukan dengan memberikan instruksi (*prompting*) kepada model berskala besar dengan kode tertutup (*closed-source*).

Kedua metode ini menghadirkan pertukaran (*trade-off*) yang jelas dalam hal alur kerja dan pengendalian mutu. Pendekatan templat memerlukan investasi awal yang besar dalam merancang model butir soal, namun membutuhkan lebih sedikit validasi pasca-pembuatan. Sebaliknya, penggunaan LLM memudahkan pembuatan konten yang beragam dalam jumlah besar, namun memerlukan upaya verifikasi yang ekstensif. LLM dapat menghasilkan cakupan domain yang tidak seimbang akibat bias bawaannya, memunculkan ketidakakuratan faktual atau keluaran yang tidak terduga, serta mengalami kesulitan dalam menciptakan pengecoh yang efektif. Kelemahan-kelemahan ini menuntut adanya proses iteratif dalam menganalisis keluaran dan menyempurnakan instruksi (*prompt*).

Akan tetapi, tidak satu pun dari kedua pendekatan tersebut yang berhasil memecahkan tantangan dalam memprediksi tingkat kesulitan butir soal secara akurat. Bahkan butir soal berbasis templat yang sederhana pun dapat menghasilkan keluaran yang tidak terduga, sehingga menegaskan pentingnya uji coba lapangan untuk mengkalibrasi butir soal sebelum digunakan. Akibatnya, peran utama AIG dalam tes berisiko tinggi adalah mengisi bank soal untuk kalibrasi di masa mendatang, bukan untuk pembuatan soal secara dinamis dan seketika (*on-the-fly*). Meskipun model prediksi parameter IRT dapat mengurangi skala uji coba lapangan, model tersebut belum cukup akurat untuk menggantikannya.

Penerapan LLM yang Berhasil untuk AIG

Keberhasilan penerapan pemberian instruksi (*prompting*) pada LLM untuk AIG bergantung pada beberapa strategi umum yang mengatasi keterbatasan metode *prompting* yang telah dijelaskan sebelumnya. Perlu dicatat bahwa strategi 1 dan 2 merupakan proses iteratif yang mungkin melibatkan banyak eksperimen berskala kecil.

1. **Membatasi input (yaitu, *prompt*):** Sangat bermanfaat untuk memulai dengan *prompt* yang menghasilkan keluaran sedekat mungkin dengan standar ideal. Jika diperlukan batasan lebih lanjut, memecah tahap pembuatan soal menjadi beberapa tahap yang lebih kecil dapat membantu. Ahli materi pelajaran (*Subject Matter Experts* atau SME) sebaiknya dilibatkan untuk memastikan keluaran selaras dengan bidang ilmu dan standar yang relevan.
2. **Membuat banyak variasi (*over-generate*) dan menyaring (*filter*):** Membuat banyak variasi soal membantu memastikan bahwa setidaknya ada beberapa soal yang

memenuhi semua kriteria. Proses penyaringan melibatkan penerapan teknik dasar *Natural Language Processing* (NLP) (misalnya, *cosine similarity*) untuk menyingkirkan keluaran yang redundan atau kecil kemungkinannya memenuhi karakteristik yang diinginkan.

3. **Mengevaluasi atau menyunting calon soal:** Melibatkan SME untuk mengevaluasi atau sekadar menyunting soal yang telah dihasilkan. Evaluasi manual dapat memakan banyak waktu, sehingga *prompt* dan filter yang efektif harus ditetapkan terlebih dahulu.

Pemberian Perintah (*Prompting*) untuk Pembuatan Konten

Perintah standar untuk AIG biasanya mencakup instruksi, persona (misalnya, "Anda adalah ahli penyusun tes"), spesifikasi format keluaran, dan beberapa contoh untuk memandu model. Pendekatan *few-shot learning* ini sering kali memberikan hasil yang lebih baik dibandingkan *zero-shot prompting*, meskipun jumlah contoh yang optimal perlu ditentukan secara empiris untuk setiap kasus penggunaan.

Untuk mendapatkan kendali yang lebih besar, tugas pembuatan konten yang kompleks dapat dipecah menjadi serangkaian perintah yang lebih kecil dan memiliki batasan ketat. Sebagai contoh, alih-alih meminta satu paragraf utuh sekaligus, seseorang dapat memberikan perintah untuk setiap kalimat secara terpisah, dengan menentukan topik, gaya, dan tingkat kerumitan yang khas untuk setiap kalimat tersebut. Untuk menghasilkan pengecoh (*distractor*) pada soal pilihan ganda, salah satu metodenya adalah menggunakan kembali kunci jawaban dari butir soal terkait, dengan memanfaatkan kemiripan semantik untuk memilih opsi yang tampak masuk akal namun tidak terlalu mudah ditebak. Pendekatan lainnya adalah memerintahkan LLM untuk membuat pengecoh secara langsung; misalnya, dengan menginstruksikan model untuk membuat jawaban berdasarkan miskonsepsi tertentu atau konteks yang bertentangan.

Jika LLM menghasilkan variasi yang kurang memadai, elemen-elemen modular dapat ditambahkan ke dalam perintah untuk meningkatkannya. Sebagai contoh, untuk menghasilkan konten dengan cakupan topik yang lebih luas, sebuah variabel dapat ditambahkan dan diisi menggunakan daftar yang telah dikurasi pada setiap pemanggilan perintah. Templat hierarkis dapat semakin menata modularitas ini untuk tugas pembuatan konten yang lebih kompleks.

Pemrosesan Pasca-Pembuatan Konten

Strategi utama dalam AIG adalah menghasilkan konten dalam jumlah berlebih lalu menyaringnya: memproduksi konten dalam volume besar, kemudian secara sistematis menguranginya menjadi kumpulan konten yang lebih sedikit namun berkualitas lebih tinggi untuk ditinjau oleh para ahli materi pelajaran (SME). Penelitian terbaru oleh Firoozi. menyajikan pendekatan terstruktur untuk proses pengendalian kualitas ini, khususnya untuk butir soal pilihan ganda. Sebagai contoh, Runge. menghasilkan 125.000 bagian teks, yang kemudian disaring menjadi 900 bagian sebelum ditinjau oleh pakar materi pelajaran (SME).

Proses penyaringan menggunakan beberapa teknik. Teknik yang paling umum adalah penyaringan pasca-generasi, di mana butir soal atau teks dihapus jika tidak memenuhi batasan tertentu, seperti jumlah kata, keterbacaan, aturan struktural (misalnya, jumlah kata atau skor

Lexile), atau kemiripan semantik dari pengecoh/distraktor. Penyaringan juga dapat disematkan langsung dalam *prompt* (instruksi) menggunakan pendekatan *LLM-as-judge* (LLM sebagai penilai), di mana model menghasilkan beberapa opsi lalu mengevaluasi keluarannya sendiri untuk memilih yang terbaik.

Selain itu, analisis pasca-generasi sangat penting untuk memperbaiki bias yang tidak terduga dalam konten yang dihasilkan. Sebagai contoh, ketika LLM menghasilkan terlalu banyak butir soal untuk satu topik tertentu, peneliti memperbaiki *prompt* mereka untuk meminta penulisan ulang menggunakan daftar topik yang telah dikurasi. Hal ini menyoroti prinsip penting: LLM memiliki perilaku dan preferensi yang tidak intuitif, sehingga analisis berulang dan penyempurnaan *prompt* menjadi hal yang esensial untuk mencapai hasil yang diinginkan.

9.6 MEMASTIKAN KUALITAS ASESMEN

Mengintegrasikan LLM ke dalam program asesmen mengharuskan dipenuhinya prinsip-prinsip pengukuran inti, yaitu validitas, reliabilitas, dan keadilan. Validitas menjadi hal yang sangat krusial karena berkaitan dengan sejauh mana bukti mendukung interpretasi skor tes untuk tujuan penggunaannya (AERA, APA, dan NCME). Karena LLM secara mendasar mengubah proses asesmen, penggunaannya harus didasari oleh alasan kuat yang menunjukkan bagaimana perubahan ini mendukung, atau bahkan meningkatkan, interpretasi skor yang dituju. Oleh karena itu, validasi asesmen yang terintegrasi dengan LLM berfokus pada pengumpulan bukti untuk membuktikan bahwa asesmen tersebut mempertahankan atau meningkatkan integritas interpretasi skor. Prosedur ini memerlukan pendekatan multidimensi yang berpedoman pada lima sumber bukti sebagaimana diuraikan dalam *Standards for Educational and Psychological Testing*: bukti berdasarkan konten tes, proses respons, struktur internal, hubungan dengan variabel lain dan konsekuensi dari pengujian (AERA).

Bagian mendasar dari validasi ini adalah mendefinisikan konstruk secara jelas, yaitu pengetahuan, keterampilan, atau karakteristik spesifik yang hendak diukur oleh tes tersebut. Definisi ini harus memperjelas bagaimana konstruk tersebut berbeda dari konstruk lainnya dan bagaimana konstruk itu seharusnya berhubungan dengan variabel-variabel lain. Proses ini dibentuk oleh tujuan penggunaan skor serta upaya desain tes dalam memitigasi masalah-masalah seperti *construct underrepresentation* (pengukuran konstruk yang dimaksud secara tidak memadai) dan *construct-irrelevant variance* (pengukuran aspek-aspek yang tidak relevan dengan konstruk tersebut).

Metodologi Validasi

Bagian ini merinci metode validasi untuk LLM dalam konteks penilaian, sebagaimana diuraikan dalam Tabel 9.3 dan 9.4. Metode-metode validasi ini mengikuti langkah-langkah krusial dalam alur pengembangan LLM. Langkah-langkah tersebut tidak boleh dipandang sebagai tahapan yang terpisah-pisah, melainkan sebagai serangkaian keputusan yang saling memperkuat dan berkontribusi pada argumen validitas yang menyeluruh. Kerangka kerja yang

kami sajikan ini bersifat praktis dan melengkapi kerangka kerja lain yang mengkaji validitas penggunaan LLM dalam penilaian pendidikan.

Tabel 9.3 Metode Validasi untuk Penilaian dan Umpan Balik Otomatis

Metode validasi	Penilaian Otomatis	Umpan Balik Otomatis
Dasar pemikiran penggunaan	Pengurangan biaya dan waktu penilaian; konsistensi penilaian dari waktu ke waktu dan pada berbagai kesempatan tes	Informasi lebih rinci mengenai karakteristik respons Mahasiswa, biasanya tidak diberi skor
Interaksi manusia/AI, beserta dasar pemikirannya	Sepenuhnya berbasis AI, kombinasi pelengkap AI dan manusia, atau gabungan AI dan manusia (Williamson dkk)	Sepenuhnya berbasis AI, atau AI dengan penyuntingan oleh manusia
Definisi rubrik	Rubrik yang digunakan untuk menetapkan skor dirumuskan dengan jelas, selaras dengan desain dan tujuan butir soal, serta mencerminkan rentang kinerja yang diharapkan untuk butir soal tersebut	Terstruktur/Spesifik: Rubrik yang digunakan untuk mengevaluasi umpan balik dirumuskan dengan jelas, selaras dengan desain dan tujuan butir soal, serta mencerminkan rentang kinerja yang diharapkan untuk butir soal tersebut
		Tidak terstruktur/Umum: Umpan balik selaras dengan skor rubrik; evaluasi umpan balik mencakup aspek-aspek utama rubrik; rubrik umpan balik digunakan untuk mengevaluasi kualitas umpan balik dibandingkan dengan umpan balik dari manusia.
Data tolok ukur manusia	Bukti menunjukkan bahwa manusia dapat dilatih untuk menerapkan rubrik secara akurat dan adil terhadap semua peserta ujian.	Terstruktur/Spesifik: Bukti menunjukkan bahwa manusia dapat dilatih untuk memberikan umpan balik secara akurat dan adil kepada semua peserta ujian.
		Tidak Terstruktur/Umum: Terdapat umpan balik dari manusia yang memiliki bentuk dan isi serupa dengan yang diberikan oleh AI (misalnya, deskripsi pada contoh jawaban acuan mengenai alasan pemberian skor tertentu).
Desain mesin LLM	Pemilihan model, mekanisme *fine-tuning* atau *prompting*, reproduksibilitas skor, serta keamanan/privasi data sudah sesuai untuk digunakan	Pemilihan model, mekanisme *fine-tuning* atau *prompting*, reproduksibilitas skor, serta keamanan/privasi data sudah sesuai untuk digunakan.
Data pelatihan/validasi	Respons yang dihasilkan mencerminkan keragaman kinerja dan skor peserta ujian, serta memiliki	Respons yang mencakup variasi kinerja dan umpan balik peserta ujian, serta dalam jumlah yang memadai untuk mengevaluasi kinerja secara

	jumlah yang memadai untuk mengevaluasi kinerja secara andal	andal. Perlu dicatat bahwa untuk umpan balik tidak terstruktur, cakupan data ini bisa jadi cukup besar.
Metrik statistik untuk evaluasi	Metrik yang mengukur tingkat kesepakatan terhadap rubrik	Metrik yang mengkaji tingkat kesepakatan mengenai penugasan umpan balik
Kriteria kinerja	Ambang batas untuk menentukan tingkat kesepakatan yang memadai	Ambang batas untuk menentukan tingkat kesepakatan yang memadai
Hubungan dengan variabel lain	Korelasi dengan skor tes keseluruhan	Tingkat keselarasan dengan rubrik penilaian butir soal
Dampak terhadap kualitas penilaian hilir	Estimasi parameter butir IRT Reliabilitas tes secara keseluruhan Skor tes secara keseluruhan	Tidak berlaku
Potensi konsekuensi negatif	Upaya memanipulasi LLM demi mendapatkan skor lebih tinggi; kurangnya keterlibatan; ketidakpercayaan terhadap AI; potensi bias dalam penilaian.	Umpan balik diterima tanpa dipertanyakan; umpan balik keliru dan mengalihkan perhatian peserta tes; potensi bias dalam umpan balik; umpan balik memicu perilaku mengakali sistem (*gaming behavior*)
Rencana pemantauan beserta langkah mitigasi	Pemantauan dengan penilaian manusia merupakan bagian dari penerapan berdasarkan kriteria kinerja.	Pemantauan dengan penilaian oleh manusia merupakan bagian dari penerapan berdasarkan kriteria kinerja

Strategi validasi ditentukan secara langsung oleh alasan utama penggunaan LLM. Tujuan umum yang ingin dicapai antara lain pengurangan biaya, peningkatan kecepatan produksi, atau peningkatan kualitas penilaian (skor), umpan balik, maupun konten tes. Prioritas yang dipilih akan menentukan fokus validasi; misalnya, tujuan pengurangan biaya mungkin hanya mengharuskan pembuktian bahwa kinerja LLM setara dengan kinerja manusia, sedangkan tujuan peningkatan program akan menuntut bukti adanya kualitas yang lebih unggul.

Bagian inti dari validasi adalah menetapkan model interaksi antara manusia dan LLM. Definisi ini dapat berkisar dari penggunaan LLM yang sepenuhnya otomatis hingga berbagai pendekatan hibrida, seperti LLM menyusun draf konten untuk ditinjau manusia, LLM menyempurnakan draf buatan manusia, atau manusia dan LLM merancang konten secara interaktif dan bersama-sama. Pendekatan yang dipilih harus memiliki dasar pembenaran yang jelas sesuai dengan tujuan penilaian serta didukung oleh bukti dari proses validasi.

Tujuan utama penggunaan LLM dalam penilaian adalah mereplikasi pekerjaan manusia yang terlatih dengan baik, baik untuk pemberian skor, penyediaan umpan balik, maupun pengembangan konten tes. Oleh karena itu, validasi yang memadai sebelum penerapan sistem memerlukan beberapa komponen kunci: penilaian manusia yang terdefinisi dengan jelas sebagai acuan kebenaran (*ground truth*), rubrik evaluasi yang jelas, serta data yang representatif untuk pelatihan dan evaluasi. Analisis ini harus mencakup subkelompok peserta tes guna memastikan keadilan, dan juga dapat mencakup pemeriksaan dampak lanjutan, seperti pengaruh otomatisasi terhadap skor tes secara keseluruhan.

Setelah diterapkan, validasi harus terus berlanjut melalui pemantauan berkala untuk memastikan kualitas komponen LLM maupun komponen manusia dalam sistem tetap terjaga. Pengawasan ini diperlukan karena penilaian manusia dapat mengalami pergeseran standar (*drift*), respons mahasiswa dapat berkembang seiring perubahan bahasa dan instruksi, norma serta nilai sosial dapat berubah dengan cara yang memengaruhi pendidikan, dan spesifikasi tes dapat berubah dari waktu ke waktu. Akibatnya, sebagian besar program penilaian menerapkan LLM dengan pengawasan manusia secara berkelanjutan untuk memantau kinerja dan menyempurnakan model jika diperlukan.

Tabel 9.4 Metode Validasi untuk Pembuatan Konten Tes Otomatis

Metode validasi	Pembuatan Konten Tes secara Otomatis
Dasar pemikiran penggunaan	Waktu pengembangan lebih singkat; potensi kualitas butir soal yang lebih baik
Interaksi manusia/AI	Proses sepenuhnya berbasis AI dengan tinjauan AI; pengembangan interaktif antara AI dan manusia; pengembangan oleh manusia dengan tinjauan AI
Definisi rubrik	Rubrik yang digunakan untuk mengevaluasi konten tes yang dihasilkan telah dirumuskan dengan jelas dan mencakup berbagai aspek penting dalam pengembangan tes
Data tolok ukur manusia	Bukti menunjukkan bahwa manusia dapat dilatih untuk menerapkan rubrik secara akurat dan adil terhadap butir soal yang dihasilkan
Desain mesin LLM	Pemilihan model, mekanisme *fine-tuning* atau *prompting*, reproduksibilitas skor, serta keamanan/privasi data sudah sesuai untuk digunakan.
Data pelatihan/validasi	Ranah konten tes dan spesifikasi pengembangan konten telah didefinisikan dengan jelas; tersedia butir soal yang memenuhi persyaratan; serta tersedia tinjauan terhadap butir soal menggunakan rubrik yang terdefinisi dengan baik.
Metrik statistik untuk evaluasi	Tingkat kesepakatan antarpenilai manusia memadai untuk menafsirkan hasil rubrik; hasil umumnya ditafsirkan berdasarkan skala rubrik untuk tinjauan butir soal (misalnya, konten dapat digunakan dengan modifikasi minimal).
Kriteria kinerja	Kriteria menentukan kapan butir soal yang dihasilkan dapat diterima untuk digunakan atau untuk ditinjau dan disunting oleh manusia.
Dampak kualitas penilaian hilir	Kualitas psikometrik butir soal yang dihasilkan AI dihitung, biasanya melalui uji coba lapangan.
Konsekuensi potensial	Aspek negatif: Peserta tes menunjukkan keterlibatan yang rendah terhadap konten buatan LLM; konten LLM tidak mencakup keluasan dan kedalaman topik; staf pengembang butir soal terlalu bergantung pada konten buatan AI sehingga mengorbankan kualitas; butir soal yang dihasilkan tidak dapat dilindungi hak cipta atau justru melanggar hak cipta/plagiarisme.
Rencana pemantauan beserta langkah mitigasi	Peninjauan berkelanjutan oleh manusia.

Pengawasan Manusia dan Pendekatan Hibrida

Pengawasan manusia merupakan hal mendasar bagi penggunaan LLM yang bertanggung jawab dalam penilaian. Dalam semua kasus, materi yang dihasilkan manusia baik berupa skor, evaluasi umpan balik, maupun butir soal tes berfungsi sebagai tolok ukur penting untuk melatih dan mengevaluasi model-model tersebut. Untuk AS (penilaian otomatis), pengawasan manusia sangat penting dalam dua tahap. Pertama, penilaian oleh manusia digunakan untuk mengevaluasi dan menyetujui kinerja LLM sebelum sistem tersebut dapat digunakan secara operasional. Kedua, setelah diterapkan, digunakan pendekatan hibrida, terutama untuk tes yang memiliki konsekuensi tinggi (*high-stakes*).

Proses ini sering kali melibatkan penilaian oleh manusia terhadap sebagian respons (biasanya 20%–35%) untuk memantau kinerja LLM, mirip dengan audit acak. Jika sistem menunjukkan kinerja buruk pada sampel ini, sistem dapat dilatih ulang dan semua respons dinilai kembali. Sebagai alternatif, proses ini dapat digunakan untuk mengarahkan respons yang tidak biasa atau merupakan pencilan (*outlier*) kepada manusia agar mendapatkan penilaian yang akurat. Dalam konteks berkonsekuensi rendah (*low-stakes*), manusia sering kali dapat mengesampingkan skor otomatis yang tidak mereka setujui.

Sama halnya dengan penilaian, sistem umpan balik otomatis memerlukan pengawasan manusia selama tahap pelatihan dan validasi. Setelah diterapkan, pemantauan bervariasi tergantung pada konteksnya. Karena sebagian besar umpan balik otomatis ditujukan untuk tujuan berkonsekuensi rendah (misalnya, penilaian formatif), pengguna biasanya bebas untuk menggunakan, mengabaikan, atau memodifikasi saran tersebut. Namun, sangat penting untuk mengumpulkan bukti mengenai bagaimana mahasiswa dan guru menggunakan serta mengubah umpan balik ini guna memastikan efektivitasnya tetap terjaga dan selaras dengan rubrik penilaian.

Untuk konten tes, hampir semua penggunaannya akan melibatkan peninjauan signifikan oleh manusia, karena LLM belum dapat beroperasi dengan cukup baik tanpanya, dan butir soal harus memenuhi standar kualitas tertentu. Proses yang melibatkan manusia dalam alur kerja (*human-in-the-loop*) ini umumnya mengambil salah satu dari tiga bentuk berikut: (1) LLM menghasilkan konten yang kemudian ditinjau dan disetujui oleh manusia, sering kali dengan penyuntingan kecil. (2) Manusia menulis draf awal, dan LLM membantu dalam proses revisi. (3) Manusia dan LLM menghasilkan konten secara kolaboratif. Pendekatan hibrida spesifik yang dipilih akan bergantung pada kebutuhan organisasi, kemampuan LLM, dan penelitian berkelanjutan untuk menyempurnakan model-model tersebut.

Mekanisme Pengendalian Kualitas

Mekanisme pengendalian kualitas memerlukan pengawasan manusia namun juga bergantung pada implementasi dan kasus penggunaan tertentu. Tiga area utama yang perlu dipertimbangkan adalah metrik kinerja, ketangguhan model, dan kebersihan data (*data hygiene*). Pertama, metrik kinerja yang presisi merupakan langkah pengendalian kualitas yang mendasar, yang memungkinkan interpretasi dan komunikasi hasil. Meskipun kriteria tersebut telah didefinisikan dengan baik untuk AS (misalnya, kesesuaian dengan penilaian manusia,

karakteristik distribusi), kriteria ini belum banyak dikembangkan untuk umpan balik otomatis dan pembuatan konten.

Kedua, ketangguhan model sangat penting untuk menangani pergeseran data dan input yang tidak lazim. Dalam hal penilaian, model yang memberikan nilai tinggi pada teks yang tidak bermakna (*gibberish*) akan mengikis kepercayaan; masalah ini sering kali diatasi dengan menggunakan model sekunder untuk menandai data pencilan (*outlier*) semacam itu. Untuk pembuatan konten, ketangguhan berarti kemampuan menangani perintah (*prompt*) pengguna yang tidak tersusun dengan baik, sembari memastikan hasil keluarannya akurat, adil, dan tidak menjiplak teks yang sudah ada.

Ketiga, kebersihan data yang baik mendukung pengendalian kualitas (QC). Penggunaan pembagian data latih/validasi/uji yang standar praktik umum dalam ilmu data memastikan bahwa kinerja model dapat digeneralisasi dan direproduksi saat beralih dari tahap pelatihan ke tahap penerapan (*deployment*). Pengendalian kualitas sangatlah krusial, karena pelaporan teknis harus mencerminkan kinerja operasional model secara akurat.

Standar dan Praktik Terbaik

Standar dan praktik terbaik telah dirumuskan dengan jelas untuk AS (penilaian otomatis), mengingat sejarahnya yang relatif panjang dalam bidang penilaian; namun, standar dan praktik serupa untuk umpan balik otomatis dan pembuatan konten tes baru mulai berkembang. Menurut pandangan kami, bagian mengenai metode validasi menyajikan pendekatan yang sangat baik untuk mengembangkan standar bagi penggunaan-penggunaan tersebut; akan tetapi, perincian pelaksanaannya belum ditetapkan. Kami memperkirakan hal ini akan muncul dalam beberapa tahun mendatang.

Terlepas dari hal itu, AS memang menyediakan contoh standar yang sangat baik yang kemungkinan besar dapat diadaptasi untuk umpan balik otomatis dan pembuatan tes. Contoh-contoh unggul tersebut mencakup bagian penilaian dalam *ITC and ATP Guidelines for Technology-Based Assessment* dan *ETS's Best Practices for Constructed-Response Scoring*. Kedua contoh tersebut sangat selaras dengan metode validasi yang diuraikan sebelumnya serta memberikan perincian yang substansial untuk setiap elemennya.

Implikasi terhadap Kebijakan Pendidikan

Kebijakan pendidikan mengacu pada serangkaian pedoman untuk pengambilan keputusan; kebijakan ini dapat bersifat mengikat secara hukum dan tidak dapat dinegosiasikan. Integrasi LLM ke dalam penilaian baik untuk pemberian skor, umpan balik, maupun pengembangan tes dapat memengaruhi dan berdampak pada kebijakan. Kami menguraikan beberapa di antaranya dalam bagian ini.

Pertama, integrasi LLM harus mematuhi kebijakan AI yang sudah ada, yang biasanya mencakup aspek privasi dan keamanan data, keselamatan mahasiswa melalui mekanisme pengaman (*guardrails*), serta masalah kekayaan intelektual seperti hak cipta dan plagiarisme, khususnya dalam pembuatan tes. Kedua, kebijakan prosedural yang mengatur aspek-aspek seperti jadwal pemberian skor atau persyaratan pengembangan butir soal dapat menjadi dasar pertimbangan penggunaan LLM (misalnya, demi kecepatan) sekaligus menjadi landasan bagi analisis dampak. Kebijakan ini mungkin mewajibkan partisipasi manusia, seperti melibatkan

pendidik setempat dalam proses penilaian, sehingga membentuk pendekatan hibrida yang melibatkan LLM. Ketiga, komunikasi merupakan komponen kebijakan yang krusial.

Pengembang tes harus menyusun materi yang menjelaskan penggunaan AI, alasan penggunaannya, manfaat, serta keterbatasannya menggunakan bahasa yang jelas dan tidak teknis agar dapat dipahami oleh khalayak luas. Waktu pelaksanaan komunikasi ini sangat menentukan dukungan pemangku kepentingan, mengingat kebijakan dapat berubah bergantung pada persepsi publik mengenai keberhasilan atau kegagalan LLM. Terakhir, Komite Penasihat Teknis, yang terdiri dari para ahli pengukuran, meninjau validitas dan kualitas aplikasi LLM serta memberikan rekomendasi terkait kualitas tersebut. Pelaporan teknis yang komprehensif dan andal tetap menjadi hal yang krusial untuk menyediakan bukti yang diperlukan guna membenarkan penggunaan LLM dalam program penilaian.

9.7 INTEGRASI LLM KE DALAM APLIKASI

Integrasi LLM ke dalam aplikasi dunia nyata, khususnya di bidang-bidang krusial seperti pendidikan, mencakup aspek yang jauh melampaui sekadar kinerja teknis. Penerapan model-model ini secara bertanggung jawab menuntut kajian mendalam mengenai implikasi sosial dan etisnya. Bagian ini membahas pertimbangan-pertimbangan penting tersebut, dimulai dengan analisis mengenai bias dan keadilan model. Selanjutnya, bagian ini mengupas pertukaran praktis (*trade-off*) antara model sumber terbuka (*open-source*) dan sumber tertutup (*closed-source*), diikuti dengan pembahasan mengenai risiko privasi dan keamanan yang signifikan. Terakhir, bagian ini menelaah dampak teknologi tersebut terhadap kesetaraan dan kesenjangan digital, dengan menyoroti ketimpangan dalam hal adopsi dan akses.

Pertimbangan Bias dan Keadilan

Saat mengevaluasi AS (*Automated Scoring* atau Penilaian Otomatis) dan pemberian umpan balik, sangatlah penting untuk membedakan antara dua jenis bias. Bias intrinsik berasal dari data pelatihan itu sendiri, sedangkan bias ekstrinsik muncul dari proses pembuatan (*generation*) oleh model dan diukur berdasarkan skor yang diberikan manusia, yang berfungsi sebagai *ground truth* atau acuan kebenaran. Berbeda dengan metode AS lainnya, LLM dapat mewarisi bias dari data pra-pelatihan yang sangat besar, yang mungkin tetap ada bahkan setelah proses *fine-tuning* (penyesuaian halus) dilakukan. Dua metode standar untuk mendeteksi bias semacam ini dalam penilaian adalah:

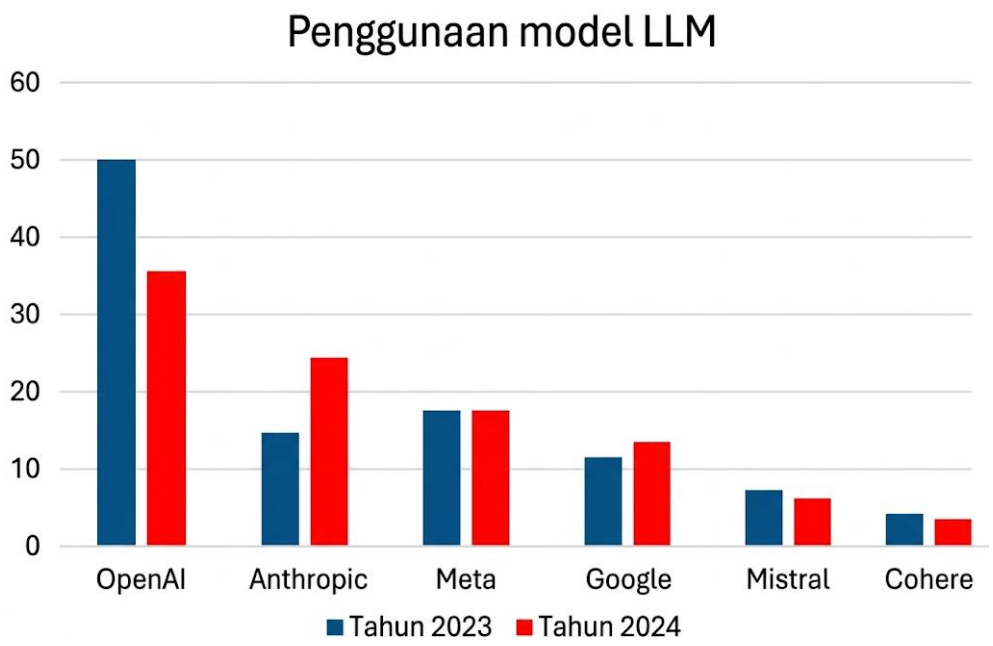
1. menghitung perbedaan rata-rata terstandarisasi dalam skor antar-subkelompok demografis serta antara skor AI dan skor manusia di berbagai subkelompok, dan
2. menggunakan pengambilan sampel yang disesuaikan (*matched sampling*), yaitu membandingkan suatu subkelompok dengan sampel dari populasi umum yang telah disesuaikan berdasarkan tingkat kemampuan untuk mengontrol perbedaan yang sudah ada sebelumnya.

Studi yang menggunakan metode-metode ini menemukan bahwa LLM yang telah melalui proses *fine-tuning* sering kali menunjukkan sedikit bias positif terhadap perempuan dibandingkan laki-laki, serta bias negatif terhadap pembelajar bahasa Inggris. Penjelasan yang mungkin mendasarinya mencakup ketergantungan model pada panjang respons yang dapat

berkorelasi dengan gender serta kompleksitas linguistik, yang dapat berbeda antara penutur asli dan bukan penutur asli. Meskipun sering kali tergolong kecil, bias-bias ini memerlukan pemantauan dan mitigasi yang berkelanjutan.

Model Sumber Terbuka (*Open-Source*) vs. Sumber Tertutup (Closed-Source)

Antarmuka yang ramah pengguna seperti ChatGPT telah mendorong adopsi luas AI generatif, dengan model sumber tertutup mendominasi pasar (pangsa 81%) dan aplikasi pendidikan. Namun, kurangnya transparansi dan potensi risiko privasi pada model-model tersebut menciptakan kebutuhan yang kuat akan alternatif sumber terbuka. Meskipun secara historis terdapat kesenjangan kinerja yang menguntungkan model sumber tertutup, peluncuran model sumber terbuka berkualitas tinggi baru-baru ini dari pengembang seperti DeepSeek, Qwen, dan Meta dengan cepat memperkecil kesenjangan tersebut. Gambar 9.2, yang menampilkan pangsa pasar penggunaan LLM antara tahun 2023 dan 2024, menunjukkan dominasi nyata model sumber tertutup dibandingkan model sumber terbuka.



Gambar 9.2 Pangsa pasar LLM, 2023–2024.

Kekuatan sesungguhnya dari model sumber terbuka tidak terletak pada kemampuannya menyamai tolok ukur umum, melainkan pada fleksibilitas penyesuaiannya (*customizability*). Perdebatan ini bukan sekadar soal "terbuka vs. tertutup", melainkan lebih kepada "yang disesuaikan (*customized*) vs. skala besar". Model sumber terbuka yang lebih kecil dan telah disesuaikan secara khusus (*fine-tuned*) untuk tugas tertentu seperti AS dapat mengungguli model tujuan umum yang jauh lebih besar. Hasil ini menunjukkan bahwa model yang sangat terspesialisasi dapat memberikan kinerja unggul pada tugas-tugas yang spesifik.

Selain itu, kemampuan replikasi yang merupakan faktor krusial dalam penilaian menjadi tantangan signifikan bagi model sumber tertutup. Temuan penelitian bisa menjadi mustahil untuk direplikasi jika penyedia mengubah model atau *prompt* sistemnya, karena kinerja dapat menurun atau berubah secara tak terduga seiring waktu. Model sumber terbuka,

yang dapat di-hosting secara mandiri dan dikontrol versinya, menawarkan landasan yang lebih stabil dan andal untuk penelitian ilmiah.

Pertimbangan Privasi dan Keamanan

Dampak AI terhadap privasi dan keamanan bersifat multidimensi. Penggunaan LLM *closed-source* (sumber tertutup) untuk penilaian berskala besar menjadi rumit karena adanya undang-undang privasi seperti *Family Educational Rights and Privacy Act* (FERPA). Mengingat respons mahasiswa dapat memuat informasi yang dapat mengidentifikasi individu secara pribadi (*Personally Identifiable Information* atau PII), pengiriman data tersebut ke API pihak ketiga untuk penilaian dianggap sebagai tindakan pengungkapan informasi. FERPA menetapkan batasan kontraktual yang ketat; aturan ini melarang penyedia AI menggunakan data Mahasiswa untuk pelatihan model di masa mendatang serta mewajibkan penerapan langkah-langkah keamanan yang dapat diaudit, seperti enkripsi dan kontrol akses. Kendala-kendala ini menyulitkan penerapan sistem penilaian yang menggunakan API eksternal.

Di sisi lain, AI juga menawarkan sarana untuk meningkatkan keamanan dan privasi. PII, sebagaimana didefinisikan dalam berbagai kerangka regulasi, mencakup tidak hanya pengenal langsung seperti nama, tetapi juga informasi tidak langsung yang memungkinkan identitas seseorang disimpulkan secara wajar. Proses penghapusan informasi semacam ini disebut anonimisasi. Metode anonimisasi tradisional yang menggunakan teknik pengenalan entitas bernama (*named entity recognition*) sering kali hanya terbatas pada penyembunyian PII secara langsung. Namun, LLM memiliki kemampuan unggul dalam menyimpulkan karakteristik pribadi dari informasi tidak langsung, sehingga memunculkan teknik-teknik yang lebih canggih. Salah satu metode tersebut adalah anonimisasi adversarial, di mana satu LLM berupaya menghapus PII sementara LLM kedua berusaha mengidentifikasi kembali individu dalam data tersebut. Pendekatan ini lebih selaras dengan kerangka regulasi yang lebih ketat.

Terakhir, AI dapat mengatasi masalah keamanan butir soal, yang merupakan aspek krusial bagi validitas tes. Ketika butir soal tes bocor atau diketahui sebelumnya, peserta tes yang memiliki pengetahuan awal tersebut akan memperoleh keuntungan yang tidak adil. Pembuatan butir soal berbasis AI dapat memitigasi risiko ini dengan cara menciptakan bank soal yang sangat besar atau menghasilkan konten dinamis sesuai kebutuhan, sehingga menghilangkan ketergantungan pada pengetahuan awal.

Kesetaraan dan Kesenjangan Digital

Konsep kesenjangan digital, yang secara tradisional berfokus pada akses terhadap teknologi dan infrastruktur, telah mengalami perkembangan. Seiring dengan maraknya penggunaan AI, kesenjangan ini kini mencakup aspek literasi dan pemanfaatan AI, sehingga menciptakan garis pemisah baru yang sering kali mencerminkan dan berisiko memperparah ketimpangan sosial yang sudah ada. Kesenjangan ini tampak sangat nyata jika ditinjau dari segi status sosial-ekonomi, ras dan etnis, serta kelompok generasi.

Status sosial-ekonomi merupakan faktor penentu utama dalam kesenjangan terkait AI. Individu dari rumah tangga berpenghasilan lebih tinggi dan mereka yang memiliki tingkat pendidikan lebih tinggi cenderung memiliki akses ke perangkat AI serta pelatihan yang diperlukan untuk memanfaatkannya secara efektif. Sebagai contoh, sebuah studi dari Pew

Research Center menemukan bahwa 42% orang dewasa di rumah tangga berpenghasilan tinggi pernah menggunakan perangkat seperti ChatGPT, dibandingkan dengan hanya 14% di rumah tangga berpenghasilan rendah. Kesenjangan ini sangat krusial, mengingat laporan dari Revelio Labs menyatakan bahwa pekerja yang mengadopsi AI memperoleh penghasilan 10% lebih tinggi dibandingkan mereka yang tidak; hal ini mengindikasikan bahwa kesenjangan AI dapat memperlebar ketimpangan ekonomi secara signifikan.

Perbedaan ini juga terlihat jelas berdasarkan generasi. Menurut laporan Randstad, generasi milenial memanfaatkan AI dalam pekerjaan mereka dengan tingkat yang lebih tinggi dibandingkan generasi lainnya, diikuti oleh Gen Z, Gen X, dan generasi *Boomer*. Tren ini sebagian didorong oleh sikap terhadap AI, di mana baik milenial maupun Gen Z tampak lebih optimistis mengenai dampak positif AI di tempat kerja. Pola serupa juga terlihat berdasarkan gender, dengan tingkat adopsi AI di kalangan pria mencapai dua kali lipat dibandingkan wanita; kondisi ini berisiko memperparah kesenjangan upah gender seiring dengan meningkatnya nilai keterampilan AI (Tabel 9.5).

Tabel 9.5 Sikap terhadap AI (*Revelio Labs*)

	Milenial (%)	**Generasi Z (%)**	**Generasi X (%)**	**Generasi Baby Boomer (%)**
Penggunaan AI generatif di tempat kerja saat ini	50	45	34	19
AI akan membuat pekerjaan saya lebih menarik	53	52	41	35
AI akan berdampak negatif terhadap kesetaraan	45	46	40	28
Siapa yang mendapatkan lebih banyak peluang untuk unggul dalam bidang AI	44	34	28	17
Perusahaan tertinggal dalam adopsi AI	47	43	35	72

Kesenjangan yang signifikan dalam adopsi AI juga terlihat di berbagai kelompok ras dan etnis. Sebuah laporan dari Revelio Labs menyoroti tren yang mencolok: bahkan setelah memperhitungkan peran pekerjaan dan tingkat paparan, para profesional keturunan Asia

mengadopsi AI dengan tingkat sekitar dua kali lipat dibandingkan rekan-rekan mereka yang berkulit putih, berkulit hitam, dan Hispanik. Data populasi yang lebih luas menunjukkan pola serupa; survei Pew Research Center menemukan bahwa 39% orang dewasa keturunan Asia telah menggunakan ChatGPT, dibandingkan dengan 23% orang dewasa berkulit putih, 21% orang dewasa Hispanik, dan 16% orang dewasa berkulit hitam. Perbedaan tingkat adopsi ini menimbulkan risiko tertinggalnya kelompok minoritas tertentu dalam ekonomi yang semakin didorong oleh AI. Selain itu, hal ini memunculkan kekhawatiran mengenai pengembangan dan penerapan AI, karena kurangnya keberagaman di kalangan pembuat dan pengguna awal (*early adopters*) dapat melanggengkan serta memperparah bias algoritmik terhadap kelompok yang kurang terwakili.

9.8 EVOLUSI LLM DALAM PENILAIAN

Pada bagian ini, kami menyoroti beberapa kemungkinan yang muncul akibat diperkenalkannya LLM. Secara khusus, kami membahas teknologi-teknologi baru yang sedang berkembang serta potensinya dalam memengaruhi penilaian pendidikan. Mulai pertengahan tahun 2025 dan seterusnya, bidang LLM dalam penilaian pendidikan berkembang pesat melampaui penerapan awal berupa pemberian perintah (*prompting*) dan penyesuaian halus (*fine-tuning*) dasar. Beberapa teknologi baru mulai membentuk generasi berikutnya dari perangkat penilaian, dengan fokus pada aspek autentisitas, reliabilitas, dan efisiensi yang lebih tinggi.

Evolusi berikutnya dari LLM adalah pengembangan agen otonom yang mampu menjalankan tugas-tugas kompleks dan bertahap tanpa intervensi manusia secara terus-menerus. Agen-agen ini dapat menggunakan berbagai perangkat, mengakses basis data, dan menjalankan seluruh alur kerja. Sebuah agen "Perancang Penilaian AI" dapat diberi tugas dengan tujuan tingkat tinggi, misalnya: "Buatlah penilaian formatif berdurasi 45 menit untuk materi genetika biologi kelas 10 yang selaras dengan standar NGSS."

Agen tersebut kemudian dapat menelusuri kurikulum secara otonom, menghasilkan beragam jenis soal (pilihan ganda, jawaban singkat, analisis diagram), memeriksa adanya bias menggunakan perangkat eksternal, serta menyusunnya menjadi kuis siap pakai di dalam sistem manajemen pembelajaran sekolah, sembari hanya menandai soal-soal yang memerlukan pertimbangan mendalam untuk ditinjau ulang oleh manusia pada tahap akhir. Teknologi ini mengotomatisasi seluruh alur pembuatan penilaian, sehingga secara drastis mengurangi beban kerja guru sekaligus memastikan keselarasan dengan standar yang berlaku.

Masa Depan Multimodal

LLM berkembang pesat melampaui antarmuka berbasis teks semata, dengan semakin mengadopsi kemampuan multimodal yang memungkinkan model-model tersebut memproses dan menghasilkan informasi dalam berbagai format, termasuk gambar, audio, dan visualisasi data. Model seperti GPT-4o dari OpenAI dan Gemini dari Google kini dapat "melihat" hasil kerja mahasiswa, "mendengar" respons lisan, dan menafsirkan diagram yang kompleks. Pergeseran mendasar dari pemahaman informasi yang murni linguistik menuju pemahaman multisensori ini siap membuka ranah baru bagi penelitian pendidikan. Transisi ini

memungkinkan peneliti untuk melangkah lebih jauh dari sekadar menganalisis teks tertulis dan mulai mengeksplorasi cakupan kompetensi mahasiswa yang lebih luas dan autentik. Dampak potensial bagi penelitian pendidikan sangatlah signifikan:

- **Menilai Keterampilan Kompleks yang Bersifat Non-Teks**: Peneliti kini dapat merancang dan mengotomatisasi penilaian tugas yang sebelumnya mustahil dievaluasi dalam skala besar. Kemampuan ini mencakup kemampuan mahasiswa dalam menafsirkan diagram ilmiah, menjelaskan proses matematika secara lisan sembari menggambar di papan tulis digital, atau mendemonstrasikan pengetahuan prosedural dalam lingkungan simulasi. Sebagai contoh, LLM dapat menganalisis jawaban akhir maupun langkah-langkah visual yang diambil mahasiswa untuk menyelesaikan soal geometri.
- **Menganalisis Penilaian Berbasis Lisan dan Presentasi**: Model multimodal dapat mentranskripsi dan menganalisis konten, kelancaran, serta struktur respons lisan atau presentasi mahasiswa. Transkripsi tersebut memungkinkan peneliti untuk mempelajari karakteristik argumentasi yang efektif dalam konteks verbal dan mengembangkan metode baru untuk memberikan umpan balik terkait keterampilan berbicara di depan umum dan komunikasi.
- **Meningkatkan Aksesibilitas dan Validitas Konstruk**: Multimodalitas menawarkan peluang baru bagi penilaian yang aksesibel, asalkan terdapat pembedaan yang jelas antara akomodasi dan preferensi belajar. Peneliti harus mengkaji apakah penggunaan moda respons alternatif (misalnya, lisan versus tertulis) dapat menghilangkan hambatan yang tidak relevan bagi mahasiswa penyandang disabilitas sehingga tetap menjaga relevansi konstruk atau justru mengubah secara mendasar aspek yang sedang diukur. Meskipun preferensi mahasiswa dapat mendorong keterlibatan, rubrik penilaian lintas-moda yang valid sangatlah penting untuk memastikan skor tetap dapat dibandingkan antar-moda.

Potensi Dampak terhadap Pendidikan

Berdasarkan pembahasan kita, integrasi LLM multimodal yang canggih berpotensi memengaruhi fondasi pendidikan itu sendiri. Perubahan yang mungkin terjadi bukan sekadar perubahan bertahap; Hal-hal tersebut merepresentasikan perubahan paradigma dalam cara kita menilai pengetahuan, menyampaikan pengajaran, serta mengonseptualisasikan peran mahasiswa dan dosen. Dampak paling signifikan akan terasa pada aspek apa yang kita ukur dan bagaimana cara kita mengukurnya.

Sistem pendidikan dapat mengalihkan fokus dari sekadar hafalan menuju evaluasi keterampilan kompleks yang dibutuhkan di dunia modern: berpikir kritis, komunikasi lisan, kolaborasi, dan pemecahan masalah secara kreatif. Bentuk penilaian bisa berupa rekaman presentasi mahasiswa, video saat mereka melakukan eksperimen sains, atau dialog interaktif dengan AI yang dirancang untuk menggali kemampuan penalaran mereka. Keragaman metode ini memberikan gambaran yang jauh lebih kaya dan valid mengenai kemampuan sesungguhnya seorang mahasiswa, sehingga mengubah penilaian dari sekadar peristiwa

tunggal yang sangat menentukan nasib menjadi proses berkelanjutan yang menyatu dengan kegiatan pembelajaran itu sendiri.

Alih-alih menggantikan peran pendidik, teknologi ini justru akan mengubah secara mendasar dan meningkatkan peran mereka di dalam kelas. AI akan menjadi "rekan pendamping" (*co-pilot*) yang andal dalam menangani tugas-tugas rutin yang menyita waktu, sehingga guru dapat lebih leluasa berfokus pada aspek-aspek pendidikan yang sangat manusiawi. Dampak akhir dari teknologi ini tidaklah ditentukan sejak awal. Hal tersebut sepenuhnya bergantung pada pilihan yang diambil oleh para pendidik, pembuat kebijakan, peneliti, dan pengembang. Jika dipandu oleh penelitian yang cermat serta komitmen teguh terhadap prinsip-prinsip pedagogis dan kesetaraan, kemajuan ini berpotensi menjadikan pendidikan lebih personal, autentik, dan memberdayakan dibandingkan sebelumnya. Sebaliknya, jika tidak dikelola dengan tepat, terdapat risiko melanggengkan ketimpangan yang sudah ada serta memunculkan ketimpangan baru. Tugas utama di bidang ini adalah mengarahkan transformasi tersebut menuju masa depan pendidikan yang lebih efektif dan berkeadilan.

BAB 10
PEMBELAJARAN MESIN UNTUK DETEKSI KECURANGAN ASESMEN

10.1 EVOLUSI KECURANGAN ASESMEN DI ERA AI

Keadilan tes merupakan pertimbangan penting dalam validitas inferensi yang dibuat berdasarkan skor tes pada asesmen skala besar yang memiliki konsekuensi tinggi (*high-stakes*). Hal ini memastikan bahwa inferensi yang ditarik dari skor tersebut baik untuk tujuan lisensi, penerimaan perguruan tinggi, maupun akuntabilitas pendidikan dapat dipertanggungjawabkan. Dalam beberapa tahun terakhir, integrasi kecerdasan buatan (AI) yang pesat telah mengubah lanskap ini secara mendasar, menciptakan persaingan dinamis antara inovasi keamanan dan metode kecurangan yang semakin canggih.

Dalam aspek pencegahan dan pelaksanaan tes, AI telah menjadi mitra yang krusial. Teknologi ini kini banyak digunakan dalam pembuatan butir soal otomatis untuk menghasilkan sejumlah besar butir soal unik, sehingga memitigasi risiko pengambilan butir soal (*item harvesting*) serta mengidentifikasi butir soal yang bermasalah atau pertanyaan yang saling eksklusif. Selain itu, meningkatnya penggunaan tes jarak jauh selama pandemi telah mendorong munculnya solusi pengawasan jarak jauh berbasis AI dengan berbagai fungsi, seperti pemantauan data biometrik, pelacakan pola tatapan mata, dan penandaan perilaku mencurigakan secara *real-time*.

Bersamaan dengan terobosan yang dibawa oleh AI, kemajuan teknologi ini juga menghadirkan ancaman yang belum pernah terjadi sebelumnya terhadap keamanan tes. Komunitas asesmen kini menghadapi berbagai tantangan, mulai dari deteksi teks buatan AI dan *deepfake* hingga penggunaan teknologi *wearable* pintar (misalnya, kacamata pintar) dan AI yang tertanam dalam perangkat keras yang mampu menyamarkan perilaku kecurangan. Meskipun lanskap AI yang lebih luas mencakup pencegahan, pemantauan, dan keamanan fisik, landasan analisis forensik tetaplah deteksi anomali di dalam data asesmen itu sendiri.

Oleh karena itu, sembari mengakui luasnya pengaruh AI terhadap keamanan tes, bab ini secara khusus berfokus pada deteksi perilaku kecurangan yang didasarkan pada data asesmen. Bab ini menunjukkan bagaimana pembelajaran mesin (*machine learning*) dan teknik augmentasi data dapat dimanfaatkan untuk mengidentifikasi kasus-kasus paling menantang yang melibatkan kemungkinan pelaku kecurangan dalam asesmen skala besar.

Pentingnya Deteksi Kecurangan

Kecurangan merupakan ancaman terhadap keadilan tes dalam asesmen skala besar. Pada umumnya, tujuan dari kecurangan adalah untuk meningkatkan skor tes secara tidak sah, terlepas dari kemampuan laten yang sebenarnya dimiliki oleh peserta tes. Ketika peserta tes tertentu menggunakan bantuan eksternal yang tidak diizinkan untuk mendongkrak skor mereka secara artifisial, hal ini menciptakan ketidakadilan bagi sebagian besar peserta tes yang mengerjakan tes tanpa bantuan semacam itu.

Perilaku demikian menimbulkan kekhawatiran mengenai keamanan tes dan dapat mengganggu validitas interpretasi skor. Oleh karena itu, deteksi kecurangan sangat penting

untuk menjamin keadilan tes, menjaga integritas hasil penilaian, serta mempertahankan kualitas teknis dalam proses pengujian. Dengan mengidentifikasi dan menangani perilaku curang, peserta tes terlindungi dari ketidakadilan, validitas interpretasi skor terjaga, dan kredibilitas hasil tes pun tetap terpelihara.

Jenis-jenis Perilaku Kecurangan

Kecurangan dalam asesmen berskala besar dapat terjadi dalam berbagai bentuk, baik sebelum, selama, maupun setelah pelaksanaan tes. Sebelum tes dilaksanakan, sebagian peserta tes mungkin memperoleh akses tidak sah terhadap butir soal, sehingga mereka dapat melihat soal-soal tersebut terlebih dahulu. Pengetahuan awal mengenai butir soal ini bisa diperoleh melalui pencurian soal yang terencana demi keuntungan finansial. Meskipun kasus-kasus masa lalu sering kali melibatkan lembaga bimbingan belajar yang mengirimkan instruktur untuk mengikuti tes secara langsung guna menghafal soal pada naskah tes yang sesungguhnya (operasional) lalu membagikannya kepada peserta kursus persiapan tes praktik pengambilan soal (*item harvesting*) modern telah berkembang menjadi aktivitas kriminal terorganisasi yang canggih.

Praktik ini mencakup penggunaan kamera tersembunyi dan perangkat perekam canggih untuk mengambil konten soal, yang kemudian disebarluaskan dengan cepat melalui media sosial, ruang obrolan pribadi, dan forum komunitas terenkripsi. Aspek komersial dari pengambilan pengetahuan awal soal ini menimbulkan ancaman yang sangat serius. Jenis kecurangan lainnya berkaitan dengan situasi di mana instruktur yang cerdik membuat soal latihan yang serupa atau hampir sama dengan soal operasional. Dengan kemajuan teknologi kecerdasan buatan generatif (genAI), jenis pengetahuan awal soal seperti ini dapat menjadi masalah yang lebih serius, mengingat soal dapat dibuat dengan memberikan perintah (*prompt*) pada alat genAI seperti ChatGPT, Gemini, Claude, dan Grok menggunakan soal latihan yang telah dirilis atau soal operasional yang bocor melalui saluran tidak resmi.

Selama pelaksanaan tes, potensi perilaku kecurangan mencakup menyontek jawaban peserta tes lain, baik dengan melihat langsung jawaban peserta di dekatnya maupun melalui pertukaran isyarat, gerakan tubuh, atau pesan berkode. Hal ini terutama mungkin terjadi pada tes linear, di mana butir-butir soal disajikan dalam urutan yang sama. Selain itu, peserta tes bisa mendapatkan jawaban dari pengawas tes akibat adanya kecurangan yang melibatkan kelompok. Bentuk kecurangan lainnya adalah penggunaan materi terlarang, seperti catatan atau perangkat elektronik yang dapat menyimpan jawaban atau menyediakan akses internet. Upaya mencari bantuan secara *real-time* dari pihak lain juga dapat dilakukan melalui pesan teks atau penggunaan *earpiece* maupun kacamata tersembunyi untuk menerima panduan dari seseorang di luar ruang tes.

Pencarian informasi secara daring semakin mungkin dilakukan dalam penilaian berbasis komputer atau penilaian jarak jauh jika peserta tes tidak diawasi dengan ketat, sehingga memungkinkan mereka mencari jawaban secara daring secara *real-time*. Dalam beberapa kasus, praktik perwakilan peserta tes (*proxy test-taking*) terjadi, di mana orang lain yang sering kali memiliki pengetahuan lebih luas mengerjakan tes atas nama peserta yang terdaftar. Selain itu, pada pelaksanaan tes jarak jauh atau tes di rumah, muncul berbagai

modus baru, seperti penggunaan beberapa perangkat secara bersamaan, berbagi layar dengan rekan kerja sama, atau penggunaan alat berbasis AI untuk membantu pemecahan masalah maupun penggunaan *generative AI* (genAI) untuk menyusun jawaban bagi soal dengan format jawaban terbuka (*constructed-response*) atau esai.

Setelah tes selesai dilaksanakan, kecurangan dapat berupa perubahan jawaban, baik yang dilakukan oleh peserta tes itu sendiri maupun oleh pihak lain. Dalam beberapa kasus, pengawas tes seperti dosen mungkin mengubah jawaban mahasiswa demi meningkatkan nilai mereka. Kolusi kelompok semacam ini bisa dipicu oleh tekanan untuk mencapai target kinerja, memperbaiki peringkat sekolah, atau mendapatkan insentif yang dikaitkan dengan prestasi mahasiswa; namun, hal ini pada dasarnya merusak keadilan dan integritas proses serta hasil penilaian. Selain itu, perlu disoroti pula kecurangan yang melibatkan genAI, yang bisa jadi sangat berbeda secara mendasar dari plagiarisme atau kolusi tradisional yang tidak melibatkan AI. Kecurangan tradisional sering kali dapat dideteksi dengan membandingkan hasil kerja seorang peserta tes dengan hasil kerja peserta lain atau dengan materi yang sudah dipublikasikan, mengingat kolusi biasanya meninggalkan jejak kemiripan pada jawaban peserta terutama pada jawaban terbuka seperti esai.

Sebaliknya, teknologi genAI pada dasarnya mampu menghasilkan konten yang sepenuhnya orisinal namun tampak masuk akal, tanpa menyalin secara langsung dari sumber yang ada, berkat kemampuan meringkas dan kemampuan penalaran yang kuat. Model genAI atau *large language models* (LLM) seperti *generative pre-trained transformer* (GPT), Gemini, Claude, Grok, atau DeepSeek dapat membuat esai, jawaban singkat, atau bahkan menjawab soal pilihan ganda dengan cara yang tampak autentik, koheren, dan sesuai dengan instruksi (*prompt*), sehingga deteksinya menjadi jauh lebih sulit.

Spektrum kecurangan berbasis AI ini mencakup pembuatan esai dan jawaban terbuka, penulisan kode pemrograman, penyusunan langkah-langkah penalaran atau pemecahan masalah untuk soal matematika atau sains, serta potensi pemilihan jawaban atau penalaran otomatis untuk soal pilihan ganda. Setiap jenis kecurangan tersebut menghadirkan tantangan deteksi yang unik; Meskipun esai mungkin terdeteksi oleh alat pendeteksi AI atau analisis gaya penulisan, jawaban pilihan ganda yang dibantu AI dapat sepenuhnya lolos dari pemeriksaan plagiarisme karena menyatu secara mulus dengan performa alami peserta tes.

Singkatnya, berbagai bentuk kecurangan baik yang dilakukan pada berbagai tahapan pelaksanaan tes maupun yang melibatkan penggunaan teknologi AI generative tidak hanya memberikan keuntungan yang tidak adil bagi pelakunya, tetapi juga mengganggu reliabilitas, validitas, dan keadilan dalam penggunaan hasil penilaian. Untuk mengatasi dampak negatif dari potensi kecurangan tersebut, berbagai metode dan teknologi telah dikembangkan guna menganalisis perilaku respons peserta tes serta mendeteksi adanya indikasi kecurangan.

10.2 METODE DETEKSI KECURANGAN

Deteksi kecurangan telah lama menjadi perhatian krusial dalam asesmen berskala besar yang memiliki konsekuensi tinggi (*high-stakes*). Berbagai metode dan teknik telah dikaji untuk mengatasi masalah ini. Teknik dan metode tersebut dapat diklasifikasikan ke dalam

pendekatan berbasis perangkat, pengawasan manusia, atau pendekatan berbasis data. Beberapa perangkat atau alat teknologi sering digunakan untuk memantau dan mencegah kecurangan. Ini mencakup kamera pengawas di ruang ujian, perangkat lunak pengawasan jarak jauh berbasis AI, peramban pengunci layar (*screen-locking browsers*), serta detektor untuk perangkat elektronik, *earpiece*, kacamata dengan sinyal nirkabel, atau perangkat komunikasi.

Selain itu, pengawas manusia sering kali dilatih untuk mengamati secara langsung dan melaporkan perilaku peserta ujian yang mencurigakan. Pengawasan langsung oleh petugas sering diterapkan dalam pelaksanaan asesmen berskala besar. Observasi oleh guru kelas dan sistem pelaporan antar-peserta didik sering kali diterapkan untuk keperluan investigasi pasca-ujian. Lapisan perlindungan tambahan mengandalkan pendekatan berbasis data untuk mendeteksi pola respons atau perilaku peserta ujian yang menyimpang. Hal ini mencakup pemodelan statistik perilaku perubahan jawaban, analisis psikometrik berdasarkan ukuran kesesuaian individu (*person-fit measures*), analisis waktu respons, pelacakan alamat IP dan lokasi untuk ujian daring, deteksi plagiarisme pada jawaban tertulis berdasarkan ukuran kemiripan, dan yang terbaru, deteksi jawaban tugas asesmen yang dibuat oleh AI generatif (genAI).

Meskipun terdapat beragam metode psikometrik forensik yang komprehensif, termasuk analisis kemiripan dan analisis selisih skor, bab ini berfokus pada kemajuan terkini dalam deteksi kecurangan menggunakan metode pembelajaran mesin (*machine learning*). Sebagai demonstrasi, kami secara khusus memusatkan perhatian pada metode berbasis AI untuk mendeteksi kemungkinan pelaku kecurangan yang telah teridentifikasi dalam program pengujian operasional; metode ini tidak mengharuskan adanya asumsi *a priori* mengenai jenis perilaku kecurangan tertentu.

Walaupun beberapa studi menggunakan ukuran kesesuaian individu untuk menandai potensi pelaku kecurangan, bab ini menunjukkan bagaimana indeks psikometrik spesifik tersebut memiliki posisi unik untuk dijadikan fitur rekayasa (*engineered features*) bagi augmentasi data dalam kerangka kerja pembelajaran mesin yang dibahas pada bagian contoh empiris selanjutnya. Oleh karena itu, bagian berikut menyoroti metode psikometrik dan pembelajaran mesin spesifik yang paling layak diterapkan dalam pendekatan hibrida yang menggunakan augmentasi data ini.

Metode Psikometrik untuk Deteksi Kecurangan

Salah satu metode tradisional untuk mendeteksi kecurangan pada era tes berbasis kertas dan pensil (*paper-and-pencil tests* atau PPT) adalah analisis penghapusan (*erasure analysis*) yang didasarkan pada pola perubahan jawaban yang menyimpang. Analisis penghapusan merupakan metode forensik untuk mendeteksi perubahan jawaban tidak wajar yang berkaitan dengan potensi kecurangan dalam PPT, khususnya ketika penilaian dilakukan menggunakan hasil pemindaian lembar jawaban berbentuk lingkaran (*bubble answer sheets*).

Analisis ini berfokus pada tingkat perubahan jawaban dari salah menjadi benar yang sangat tinggi. Pola-pola tersebut biasanya diperiksa untuk menentukan apakah jawaban peserta tes diubah setelah tes berlangsung oleh pihak ketiga (misalnya, guru, administrator, atau pengawas yang memanipulasi lembar jawaban) atau apakah peserta tes melakukan

perubahan jawaban secara strategis yang menyimpang dari perilaku populasi peserta tes pada umumnya. Berbeda dengan analisis kemiripan (*similarity analysis*), analisis penghapusan bukanlah alat utama untuk mendeteksi penyalinan jawaban antarpeserta tes, melainkan untuk mengidentifikasi perubahan anomali yang mengindikasikan adanya manipulasi catatan jawaban atau ketidakkonsistenan data. Dalam tes berbasis komputer, data penghapusan atau perubahan jawaban dapat direkam dengan tingkat presisi yang lebih tinggi, sehingga urutan pilihan atau jawaban yang dipilih dapat dilacak secara tepat.

Ukuran Kesesuaian Individu (*Person-Fit Measures*)

Kategori lain dari metode psikometrik untuk mendeteksi kecurangan yang banyak digunakan melibatkan analisis data respons butir soal guna menilai kesesuaian individu (*person fit*). Ranger meneliti 18 indikator untuk mendeteksi kecurangan dalam ujian elektronik (*e-exam*) di perguruan tinggi. Studi mereka mencakup indeks kesesuaian individu yang didasarkan pada respons butir soal, seperti U1 (van der Flier), U3 (van der Flier), CS (Sato), dan statistik HT. Studi tersebut juga mengkaji indikator berbasis waktu respons (Sinharay,) termasuk statistik KL, statistik Z2, dan statistik KT serta indikator revisi respons, seperti jumlah revisi (N1, N2), revisi dari jawaban salah menjadi benar (NC1, NC2), dan waktu revisi (T1, T2, T3).

Selain itu, studi tersebut mempertimbangkan indikator yang menilai hubungan antara respons dan waktu respons, beserta dua indikator lain yang didasarkan pada kemiripan pola respons. Hasil studi menunjukkan bahwa jumlah revisi jawaban merupakan prediktor paling penting dalam deteksi kecurangan. Sementara itu, indikator berbasis waktu respons mampu mengidentifikasi potensi pelaku kecurangan. Namun, perlu ada kehati-hatian terkait efektivitas indeks kesesuaian individu. Penelitian menunjukkan bahwa statistik-statistik ini dapat memiliki daya uji (*statistical power*) yang rendah dibandingkan dengan model forensik tertentu dan sering kali menghasilkan tingkat positif palsu yang tinggi.

Oleh karena itu, ketika kecurangan terdeteksi, diperlukan upaya investigasi yang mendalam untuk memastikan bahwa ketidaksesuaian (*misfit*) tersebut memang disebabkan oleh kecurangan dan bukan oleh sumber varians lain yang tidak relevan dengan konstruk yang diukur. Sebagai contoh, respons yang menyimpang tidak serta-merta mengindikasikan kecurangan, melainkan bisa jadi bersumber dari varians yang tidak relevan dengan konstruk (*construct-irrelevant variance*), seperti kecemasan, kelelahan, atau gaya pemrosesan kognitif yang khas. Oleh karena itu, meskipun indeks kesesuaian individu (*person-fit indices*) merupakan strategi yang masuk akal untuk memulai penyelidikan, indeks tersebut biasanya memerlukan bukti pendukung yang kuat untuk menyimpulkan secara pasti adanya kecurangan atau melakukan pembatalan skor.

Pendekatan Berbasis Model Variabel Laten

Sementara metode psikometrik tradisional menggunakan respons terhadap butir soal, perkembangan terkini telah mengintegrasikan data proses dalam deteksi kecurangan. Sebagaimana diulas oleh Ranger, data waktu respons butir soal dapat dikumpulkan dengan mudah dan telah diteliti untuk mendeteksi butir soal yang bocor dan/atau peserta tes yang melakukan kecurangan. Berbagai model variabel laten yang memanfaatkan respons butir soal

dan/atau waktu respons telah diteliti untuk deteksi kecurangan, termasuk pemodelan campuran (*mixture modeling*), analisis titik perubahan Bayesian (*Bayesian change point analysis*), pemodelan respons butir soal dengan mekanisme gerbang (gated item response modeling), dan pemodelan respons nominal.

Baru-baru ini, Man dan Harring mengevaluasi pengetahuan awal mengenai butir soal dengan menerapkan model gabungan multi-kelompok yang mengintegrasikan respons butir soal, waktu respons, dan jumlah fiksasi visual. Pemodelan gabungan dari berbagai sumber data yang saling terkait semacam itu diharapkan dapat meningkatkan akurasi deteksi kecurangan. Namun, ketika struktur data proses menjadi semakin kompleks, model statistik parametrik khususnya pendekatan pemodelan berbasis variabel laten mungkin tidak cukup layak untuk mengakomodasi kompleksitas data dari berbagai sumber jika asumsi model tidak terpenuhi.

Secara umum, efektivitas metode psikometrik yang ada bergantung pada terpenuhinya asumsi-asumsi yang ketat. Selain bergantung pada seberapa baik model pengukuran sesuai dengan data, metode-metode ini sering kali mengasumsikan bahwa jenis kecurangan yang terjadi sudah diketahui. Sebagai contoh, deteksi terhadap pengetahuan awal (*pre-knowledge*) menggunakan teknik psikometrik mungkin mengharuskan kita mengetahui secara pasti butir soal mana yang telah bocor, apakah kunci jawaban turut bocor bersama butir-butir soal tersebut, atau seberapa jauh lebih cepat peserta tes yang memiliki pengetahuan awal memberikan respons dibandingkan peserta tes yang jujur.

Jika asumsi-asumsi tersebut tidak dapat dipertahankan, peneliti berisiko memilih indeks yang tidak tepat atau menghasilkan alat deteksi yang tidak memiliki daya cukup untuk mengidentifikasi pelanggaran tersebut. Sebaliknya, pendekatan pembelajaran mesin (*machine learning*) sering kali mampu menandai pola respons yang tidak lazim tanpa harus bergantung pada asumsi yang kuat maupun spesifikasi awal yang kaku. Lebih jauh lagi, ketika perangkat AI generatif (genAI) seperti ChatGPT digunakan untuk menghasilkan respons berbasis teks terhadap butir soal penilaian, metode psikometrik tradisional mungkin tidak lagi memadai untuk menganalisis data teks tersebut dalam rangka mendeteksi kecurangan.

Metode Pembelajaran Mesin untuk Deteksi Kecurangan

Seiring dengan semakin maraknya penggunaan asesmen digital, data produk maupun proses kini dapat dikumpulkan dengan mudah. Dengan memanfaatkan kemajuan dalam pembelajaran mesin (*machine learning*) dan pembelajaran mendalam (*deep learning*), para peneliti telah mengkaji berbagai pendekatan untuk mendeteksi kecurangan, termasuk algoritma pembelajaran terawasi (*supervised algorithms*) dan algoritma pembelajaran tak terawasi (*unsupervised algorithms*). Baru-baru ini, metode pembelajaran mendalam juga telah diteliti untuk mendeteksi perilaku respons yang menyimpang serta kecurangan.

Keunggulan utama penerapan algoritma pembelajaran mesin dan pembelajaran mendalam untuk deteksi kecurangan terletak pada kemampuannya dalam menangani data yang kompleks, non-linear, multidimensi, dan multimodal hal yang tidak dapat dilakukan oleh model dan metode psikometrik tradisional. Pertama, pendekatan ini mampu menangkap hubungan yang rumit dan non-linear di antara berbagai sumber data dalam asesmen berskala besar termasuk respons butir soal, waktu respons, serta jumlah dan pola perubahan jawaban

tanpa terikat oleh asumsi-asumsi kuat yang mendasari model dan metode psikometrik variabel laten. Selain itu, metode pembelajaran mesin sangat fleksibel dan dapat ditingkatkan skalanya (*scalable*), sehingga memungkinkan integrasi data multimodal seperti data biometrik (misalnya, fiksasi mata) dan bentuk data proses sekuensial lainnya (berdasarkan dinamika penekanan tombol dan gerakan tetikus) seiring ketersediaan data tersebut.

Metode ini juga mendukung pemilihan fitur secara otomatis, yang memungkinkan identifikasi pola halus atau kombinasi indikator yang mungkin terlewatkan oleh model tunggal atau indikator yang didasarkan pada metode psikometrik tradisional. Lebih lanjut, teknik *ensemble* dan pembelajaran mendalam dapat meningkatkan akurasi prediksi dengan menggabungkan beberapa model prediktor sederhana (*weak learners*) atau dengan mempelajari representasi hierarkis dari data yang kompleks. Terakhir, model pembelajaran mesin dapat terus diperbarui dan dilatih ulang seiring bertambahnya data asesmen baru; hal ini memungkinkan sistem deteksi kecurangan untuk beradaptasi dengan cepat terhadap strategi kecurangan yang terus berkembang, berbeda dengan model dan metode psikometrik tradisional yang mungkin hanya mampu mendeteksi jenis strategi kecurangan yang bersifat tetap.

Algoritma Pembelajaran Mesin Terawasi (*Supervised Machine Learning*)

Metode pembelajaran mesin terawasi dapat digunakan untuk melatih pendeteksi kecurangan ketika tersedia data berlabel, misalnya dalam kasus di mana pelaku kecurangan potensial telah ditandai. Cavalcanti dkk. meninjau metode penambangan teks untuk deteksi kecurangan dan memperkenalkan dua model klasifikasi berdasarkan algoritma Pohon Keputusan (*Decision Tree*). Untuk menangani plagiarisme yang melibatkan parafrase dan peringkasan, Chitra dan Rajkumar mengembangkan sistem pengenalan parafrase menggunakan pendekatan *support vector machine* (SVM). Sahu menerapkan algoritma *K-nearest neighbor* (K-NN) untuk pengenalan pola guna mengidentifikasi teks yang disalin dengan membandingkannya terhadap berkas yang sudah ada. Zopluoglu menggunakan *eXtreme Gradient Boosting* (XGBoost) untuk mendeteksi pengetahuan awal mengenai butir soal dalam program tes berskala besar.

Demikian pula, Man meneliti algoritma terawasi (K-NN, *random forest*, dan SVM) maupun algoritma tak-terawasi (*K-means clustering* dan *self-organizing maps*) untuk deteksi kecurangan. Melanjutkan penelitian tersebut, Zhou dan Jiao menerapkan pendekatan pembelajaran mesin *stacking ensemble* untuk mengintegrasikan respons butir soal peserta tes, waktu respons, dan data tambahan guna mendeteksi kecurangan. Mereka kemudian membandingkan kinerja model stacking tersebut dengan dua teknik *ensemble* lainnya (*random forest* dan *gradient boosting*) serta enam model pembelajaran mesin individual non-*ensemble*, termasuk Pohon Keputusan, Naïve Bayes, jaringan saraf (*neural network*), analisis diskriminan, regresi logistik, dan SVM. Studi-studi ini mengeksplorasi berbagai model pembelajaran mesin untuk berbagai kumpulan data. Tidak ada satu model pun yang secara bulat diidentifikasi sebagai model dengan kinerja terbaik. Mengingat deteksi kecurangan lebih merupakan persoalan empiris, disarankan agar berbagai metode dibandingkan secara empiris untuk mengidentifikasi model dengan akurasi deteksi tertinggi.

Algoritma Pembelajaran Mesin Tanpa Pengawasan (*Unsupervised Machine Learning*)

Sementara algoritma pembelajaran mesin dengan pengawasan (*supervised*) mengadopsi pendekatan konfirmatif, algoritma tanpa pengawasan (*unsupervised*) menawarkan perspektif yang lebih eksploratif untuk mendeteksi potensi pelaku kecurangan. Perlu dicatat bahwa ketika metode pembelajaran mesin tanpa pengawasan diterapkan untuk deteksi kecurangan, sebagian besar penelitian berfokus pada identifikasi perilaku respons yang menyimpang. Hal ini dikarenakan klasterisasi atau anomali pencilan (*outlier*) tidak serta-merta mengindikasikan adanya kecurangan. Meskipun demikian, deteksi anomali semacam itu memberikan titik awal yang berharga untuk penyelidikan lebih lanjut terhadap potensi kasus kecurangan. Klasterisasi dan deteksi anomali adalah dua pendekatan tanpa pengawasan yang paling umum digunakan dalam deteksi kecurangan. Liao dkk. meneliti pelaku *item harvesting* (pengumpul butir soal) yang mencoba menghafal butir soal dan membagikannya kepada peserta tes berikutnya.

Studi mereka memanfaatkan data proses dan algoritma klasterisasi K-means untuk mempelajari pola dasar (*archetype*) dari proses pengerjaan tes; pola yang menyimpang dari pola dasar respons tersebut kemudian ditandai untuk penyelidikan lebih lanjut. Man dkk. mengeksplorasi algoritma tanpa pengawasan (klasterisasi K-means dan *self-organizing maps*) untuk deteksi kecurangan. Pan dan Wollack menggunakan pendekatan berbasis pembelajaran tanpa pengawasan tipe ensemble untuk mendeteksi pengetahuan awal (*pre-knowledge*) terkait peserta maupun butir soal secara bersamaan.

Gorgun dan Bulut mengeksplorasi enam metode deteksi anomali tanpa pengawasan (*Gaussian mixture model*, *Bayesian Gaussian mixture model*, *Isolation Forest*, jarak Mahalanobis, *local outlier factor*, dan *Elliptic Envelope*) untuk mengidentifikasi respons menyimpang dalam sistem tutor cerdas (*intelligent tutoring systems*) dengan data respons yang sangat jarang (*sparse*). Temuan mereka menunjukkan bahwa sebagian besar metode menandai respons menyimpang yang serupa, meskipun metode *local outlier factor* menghasilkan hasil yang sangat berbeda. Jarak Mahalanobis terbukti menjadi detektor yang konservatif, sedangkan *Isolation Forest* dan *Gaussian mixture models* cenderung lebih liberal. Secara keseluruhan, studi tersebut menunjukkan bahwa metode deteksi anomali tanpa pengawasan menjanjikan untuk mengidentifikasi respons menyimpang dan merekomendasikan penggunaan berbagai metode untuk meningkatkan akurasi.

Algoritma Pembelajaran Mendalam (*Deep Learning*)

Selain itu, penelitian terkini juga mengeksplorasi berbagai metode pembelajaran mendalam (*deep learning*) untuk deteksi kecurangan. Tang mengeksplorasi data *click-stream* untuk mendeteksi perilaku pengerjaan tes yang tidak lazim menggunakan jaringan *long short-term memory* (LSTM). Data *clickstream* mencatat rincian tindakan pengguna. Data tak terstruktur semacam itu sulit dimodelkan secara efektif menggunakan model psikometrik tradisional karena kompleksitas dan sifat dinamisnya. Studi ini memperkenalkan *model agreement index* (MAI), sebuah ukuran untuk menguantifikasi seberapa tipikal atau atipikal *clickstream* seorang peserta tes dibandingkan dengan model perilaku yang diharapkan (yang telah dilatih sebelumnya).

Temuan empiris menunjukkan bahwa peserta tes dengan skor MAI rendah sering kali mengalami kesulitan akibat tindakan berulang membuka dan menutup *widget* alat, yang berujung pada pemborosan waktu. Studi ini menguraikan proses pelatihan model LSTM, membandingkannya dengan tolok ukur dasar berupa "tindakan berikutnya yang paling umum", serta menganalisis korelasi antara MAI dan indikator perilaku lainnya seperti perubahan jawaban dengan menggunakan data dari program tes tingkat negara bagian.

Sebagaimana disoroti oleh Ranger, algoritma pembelajaran terawasi (*supervised learning*) dapat mencapai akurasi tinggi dalam mendeteksi kecurangan; namun, algoritma ini memerlukan kumpulan data berlabel yang memuat contoh pelaku kecurangan potensial sesuatu yang biasanya belum tersedia pada tahap awal siklus penilaian. Studi mereka mengeksplorasi penggunaan pembelajaran transfer (*transfer learning*) dengan mengadaptasi detektor yang telah dilatih pada satu kumpulan data berlabel agar dapat digunakan pada kumpulan data baru yang tidak berlabel dari tes yang berbeda.

Hasil empiris menunjukkan bahwa mentransfer detektor yang telah dilatih sebelumnya dapat meningkatkan akurasi dibandingkan dengan pendekatan tanpa pengawasan (*unsupervised*): transfer sederhana (*naive transfer*), yang hanya menggunakan kembali detektor tersebut, secara substansial meningkatkan akurasi, sementara transfer melalui algoritma pelabelan mandiri (*self-labeling*) memberikan peningkatan tambahan yang moderat. Temuan ini mengindikasikan bahwa pembelajaran transfer menawarkan opsi praktis untuk meningkatkan deteksi kecurangan di tahap awal proses penilaian dengan memanfaatkan detektor terlatih yang sudah ada.

Model Bahasa

Model bahasa merupakan kemajuan terkini dalam bidang pemrosesan bahasa alami (*Natural Language Processing* atau NLP). *Bidirectional Encoder Representations from Transformers* menonjol karena kemampuannya memanfaatkan komponen *encoder* dari arsitektur Transformer untuk menangkap konteks dua arah (*bidirectional*) yang mendalam. Model satu arah (*unidirectional*) membatasi pembelajaran hanya pada konteks yang muncul tepat sebelumnya dengan memproses teks secara sekuensial (misalnya, dari kiri ke kanan), sedangkan model dua arah mempertimbangkan konteks secara utuh dengan mengakses informasi dari kata-kata sebelum dan sesudahnya secara bersamaan. Berbeda dengan model bahasa satu arah tradisional, BERT menggunakan tujuan pelatihan berupa pemodelan bahasa bertopeng (*masked language modeling*).

Dalam proses ini, token (yaitu unit dasar teks, seperti kata atau sub-kata) disembunyikan (*masked*) secara acak, dan model berupaya memprediksinya berdasarkan konteks yang disediakan oleh token-token di sekitarnya. Hal ini memungkinkan model untuk menangkap hubungan dua arah yang mendalam. Tujuan pra-pelatihan ini memungkinkan BERT mempelajari representasi semantik dan sintaksis yang kaya serta peka terhadap konteks, sehingga menghasilkan kinerja yang unggul dalam berbagai tugas hilir (*downstream tasks*). Arsitekturnya yang hanya menggunakan *encoder* menghasilkan representasi bahasa yang tangguh dan dapat disesuaikan lebih lanjut (*fine-tuned*) untuk berbagai aplikasi, menjadikan BERT sebagai model fundamental dalam penelitian NLP modern. Kesuksesan BERT telah

menginspirasi berbagai adaptasi yang bertujuan untuk meningkatkan kinerja dan efisiensi model. RoBERTa menyempurnakan tahap pra-pelatihan BERT dengan menghilangkan tugas prediksi kalimat berikutnya (*next sentence prediction*), menerapkan *masking* dinamis, serta menggunakan dataset yang lebih besar dengan durasi pelatihan yang lebih lama, sehingga menghasilkan generalisasi dan kemampuan menangkap pola linguistik yang lebih baik. DeBERTa memperkenalkan mekanisme *disentangled attention* yang memisahkan informasi konten dan posisi, sehingga meningkatkan pemodelan ketergantungan antar-kata.

ELECTRA mengganti metode *masked language modeling* dengan tujuan deteksi token yang digantikan (*replaced token detection*), yang memungkinkan pembelajaran yang efisien dalam penggunaan sampel serta kinerja yang setara atau bahkan lebih unggul dibandingkan BERT dengan kebutuhan sumber daya yang lebih sedikit. ConvBERT mengintegrasikan mekanisme konvolusional ke dalam lapisan *self-attention* untuk menangkap ketergantungan lokal maupun jarak jauh secara efisien, sementara DistilBERT) menggunakan teknik *knowledge distillation* untuk menghasilkan model yang lebih kecil dan lebih cepat namun tetap mempertahankan sebagian besar tingkat akurasi BERT. Secara keseluruhan, berbagai inovasi ini menunjukkan bagaimana modifikasi pada arsitektur, pra-pelatihan, dan mekanisme *attention* dapat menghasilkan model bahasa yang lebih efektif dan efisien dalam penggunaan sumber daya.

Model bahasa untuk mendeteksi respons teks buatan AI dapat dilatih atau disesuaikan (*fine-tuned*) untuk menganalisis teks guna mengidentifikasi pola yang mengindikasikan hasil buatan mesin. Model-model ini memanfaatkan petunjuk linguistik, konsistensi gaya, dan ketidakwajaran statistik yang sering kali berbeda dari tulisan manusia, seperti pengulangan frasa, sintaksis yang tidak lazim, atau kombinasi kata yang kemungkinannya kecil terjadi secara alami. Model deteksi dapat berupa model berbasis klasifikasi yang dilatih menggunakan contoh esai buatan AI dibandingkan dengan esai tulisan manusia atau model probabilistik yang mengestimasi kemungkinan bahwa suatu teks dihasilkan oleh model bahasa. Yan dkk. melakukan penyesuaian (*fine-tuning*) pada model RoBERTa untuk mengklasifikasikan esai buatan AI versus esai tulisan manusia. Mereka kemudian membandingkan kinerja model bahasa ini dengan model SVM yang menggunakan fitur hasil ekstraksi *e-rater* sebuah mesin penilaian otomatis internal milik perusahaan dan mendapati bahwa model bahasa tersebut menghasilkan akurasi deteksi yang lebih tinggi. Jiao meneliti deteksi esai yang dihasilkan oleh ChatGPT 3.5 serta membandingkan pendekatan berbasis fitur dengan pendekatan berbasis model bahasa.

Model pembelajaran mesin terbimbing (*supervised machine learning*) digunakan untuk pendekatan berbasis fitur, mencakup regresi logistik, SVM, Naïve Bayes, *gradient boosting*, *random forest*, dan *Decision Tree*; sementara model bahasa yang diteliti meliputi BERT, ELECTRA, dan ConvBERT. Secara umum, model bahasa menunjukkan kinerja yang lebih baik dibandingkan model pembelajaran mesin terbimbing berbasis fitur, dengan ELECTRA menghasilkan akurasi deteksi tertinggi. Namun, studi oleh Hua dan Yao menguji baik model pembelajaran mesin terbimbing (SVM, regresi logistik, dan XGBoost) maupun model bahasa (BERT, ELECTRA, dan RoBERTa) dalam mendeteksi esai yang dihasilkan oleh berbagai LLM,

termasuk GPT 3.5, GPT 4, GPT-4o, Gemini, Gemini Advanced, dan Claude 3.5. Studi mereka menemukan bahwa SVM berbasis fitur memiliki kinerja terbaik di antara semua model yang diteliti, termasuk model bahasa. Dalam ranah pengujian berisiko tinggi (*high-stakes testing*), organisasi seperti Duolingo telah memelopori penerapan teknik NLP canggih ini serta deteksi teks buatan AI untuk mengamankan pelaksanaan asesmen jarak jauh.

Temuan penelitian mengenai model dengan kinerja terbaik untuk mendeteksi teks buatan AI menunjukkan variasi yang cukup besar; hal ini mencerminkan perbedaan dalam metode pembuatan data menggunakan AI generatif (genAI), kriteria evaluasi, serta proses pelatihan model. Meskipun metode-metode deteksi ini menunjukkan keberhasilan parsial dalam mengidentifikasi teks sintetis, efektivitasnya semakin ditantang oleh pesatnya kemajuan teknologi AI generatif (genAI). Hal ini dikarenakan genAI semakin mampu menghasilkan teks yang tidak hanya lebih alami tetapi juga memiliki kecanggihan kontekstual, serta memuat pola-pola linguistik bernuansa yang sangat menyerupai konten buatan manusia.

Evolusi ini memicu persaingan dinamis dan berkelanjutan antara model bahasa generative yang bertujuan menghasilkan keluaran yang tidak dapat dideteksi dan model deteksi yang bertugas membedakan teks buatan mesin dari teks buatan manusia. Penyempurnaan model generatif yang terus-menerus dengan memanfaatkan kumpulan data yang lebih besar, arsitektur yang lebih baik, serta teknik pelatihan mutakhir seperti *reinforcement learning* atau *transfer learning* semakin memperberat tantangan dalam mendeteksi respons teks buatan AI pada penilaian berskala besar di tengah lanskap teknologi yang terus berkembang.

10.3 AUGMENTASI DATA DALAM DETEKSI KECURANGAN

Ketidakseimbangan Kelas dalam Deteksi Kecurangan

Dalam deteksi kecurangan, ketidakseimbangan kelas menghadirkan tantangan kritis bagi model pembelajaran mesin (*machine learning*) yang diawasi (*supervised*), karena kelangkaan kasus positif dapat menyebabkan bias pada pelatihan, membatasi kinerja prediksi, dan mendistorsi evaluasi kinerja. Sebagai contoh, dalam dataset yang tersedia untuk umum yang dilaporkan oleh Cizek dan Wollack, hanya 46 dari 1.636 peserta tes yang ditandai sebagai kemungkinan pelaku kecurangan, yang mencakup hanya 2,81% dari sampel. Ketidakseimbangan ekstrem ini tidak hanya menyoroti jarangnya kecurangan yang teramati dalam penilaian berskala besar, tetapi juga menggarisbawahi kesulitan metodologis dalam mengembangkan model yang mampu mengidentifikasi perilaku yang jarang terjadi tersebut secara andal.

Ketidakseimbangan kelas antara kemungkinan pelaku kecurangan (kelas minoritas) dan sejumlah besar peserta tes yang tidak melakukan kecurangan (kelas mayoritas) dapat menyebabkan kurangnya representasi kelas minoritas dalam pelatihan model. Tanpa contoh positif yang memadai, model pembelajaran mesin yang diawasi dapat mengalami overfitting pada kelas mayoritas (bukan pelaku kecurangan) dan gagal mendeteksi perilaku kecurangan yang samar atau baru muncul. Dalam hal ini, augmentasi data menjadi sangat penting untuk pengembangan model pembelajaran mesin bagi deteksi kecurangan karena kasus berlabel

mengenai potensi pelaku kecurangan sering kali langka, terbatas, dan mencakup porsi yang sangat kecil dari total peserta tes.

Augmentasi Data

Augmentasi data adalah teknik yang digunakan dalam pembelajaran mesin untuk meningkatkan ukuran, representasi, dan keragaman dataset dengan menghasilkan versi yang dimodifikasi atau sintetis dari titik data yang sudah ada. Tujuan utamanya adalah untuk meningkatkan kinerja, ketangguhan (*robustness*), dan kemampuan generalisasi model pembelajaran mesin dengan memaparkan model yang dilatih pada rentang variasi dan skenario yang lebih luas.

Pendekatan ini sangat berharga ketika data pelatihan terbatas, tidak seimbang, atau kurang beragam, karena membantu memastikan model menjadi lebih tangguh, mampu melakukan generalisasi, dan dapat beradaptasi dengan berbagai kondisi. Secara umum, augmentasi data dapat mengatasi masalah terkait ketidakseimbangan kelas, keterbatasan data, kelangkaan data (*data sparsity*), *overfitting*, serta kekhawatiran privasi data, sekaligus meningkatkan *transfer learning*.

Teknik augmentasi seperti menghasilkan pola respons sintetis, menyimulasikan perilaku *clickstream* yang tidak lazim, atau membuat versi yang dimodifikasi (*perturbed*) dari kasus kecurangan yang diketahui dapat membantu memperluas dataset secara artifisial serta menghadirkan keragaman dan representasi perilaku kecurangan yang lebih baik. Hal ini meningkatkan ketangguhan model, memperkuat kemampuannya untuk menggeneralisasi model yang telah dilatih terhadap perilaku kecurangan baru atau yang belum pernah ditemui sebelumnya, serta mengurangi risiko bias terhadap kelas mayoritas yang tidak melakukan kecurangan.

Pada akhirnya, augmentasi data menjadikan sistem deteksi kecurangan lebih andal, adaptif, dan efektif dalam lingkungan pengujian nyata, di mana data kecurangan yang autentik jumlahnya terbatas dan terus berkembang. Meskipun sebagian besar penelitian mengenai augmentasi data berasal dari bidang visi komputer (untuk data citra) dan pemrosesan ucapan, bab ini berfokus pada augmentasi data tabular dan tekstual untuk deteksi kecurangan dalam asesmen berskala besar, mengingat data produk dan proses butir soal sering kali disajikan dalam format tabel atau teks, seperti pada penulisan esai dan jawaban uraian. Menurut Bayer dkk, metode augmentasi data secara umum dapat diklasifikasikan ke dalam dua kategori: ruang fitur dan ruang data. Augmentasi ruang fitur bertujuan untuk menambah jumlah variabel yang terkait dengan setiap sampel dalam sebuah dataset. Dasar pemikirannya serupa dengan penyertaan kovariat dalam analisis regresi, di mana informasi tambahan ditambahkan untuk memperkaya representasi setiap kasus dan menyediakan landasan analisis yang lebih komprehensif.

Di sisi lain, augmentasi ruang data bertujuan untuk meningkatkan ukuran sampel sekaligus keterwakilan dataset, yang sering kali dilakukan menggunakan teknik *resampling*. Jika fitur disusun dalam kolom dan kasus dalam baris, maka augmentasi ruang fitur mengacu pada penambahan jumlah kolom (yaitu variabel), sedangkan augmentasi ruang data mengacu pada penambahan jumlah baris (yaitu ukuran sampel atau jumlah kasus). Dalam deteksi

kecurangan, fitur biasanya berupa indikator perilaku seperti pola waktu respons, perilaku navigasi, dan urutan perubahan jawaban yang dapat menangkap potensi pelanggaran. Namun, fitur yang digunakan untuk augmentasi data dalam deteksi kecurangan perlu dikaji secara cermat dengan mempertimbangkan aspek validitasnya.

Metode untuk Data Tabular

Augmentasi data tabular sangat relevan dalam deteksi kecurangan karena data respons butir soal, data proses yang dikumpulkan dari *clickstream*, atau berkas log sering kali terstruktur dalam bentuk baris (peserta tes) dan kolom (fitur). Augmentasi ruang fitur berkaitan dengan pembuatan variabel (kolom) baru yang memperkaya representasi setiap sampel. Bentuk augmentasi ruang fitur yang paling sederhana adalah menyertakan kinerja peserta tes sebelumnya atau riwayat keikutsertaan tes (misalnya, apakah peserta merupakan pengulang tes atau bukan), serta data kontekstual seperti kondisi tes (waktu pelaksanaan, perangkat yang digunakan, atau durasi tes).

Selain itu, fitur baru dapat dibuat berdasarkan fitur yang sudah ada, dengan mengacu pada teori atau pengetahuan spesifik bidang terkait. Sebagai contoh, rasio antara waktu respons dan tingkat kesulitan butir soal, atau jumlah klik per menit, dapat diturunkan menggunakan informasi tingkat butir soal yang ada serta data waktu respons setiap peserta tes. Variabel lain yang mampu menangkap hubungan interaksi antarvariabel juga dapat dibuat. Misalnya, fitur baru dapat diturunkan berdasarkan interaksi antara kemampuan laten peserta tes dalam hal akurasi respons dan kecepatan laten.

Di samping itu, statistik ringkasan yang diagregasi dari berbagai butir soal seperti nilai minimum, maksimum, rata-rata, dan varians waktu respons butir soal serta durasi jeda maksimum dapat disertakan untuk augmentasi data. Lebih lanjut, beberapa transformasi matematis dapat diterapkan, termasuk transformasi polinomial untuk menangkap hubungan non-linear antarfitur mentah, transformasi logaritmik atau eksponensial untuk mengubah distribusi yang menceng (*skewed*) menjadi distribusi normal, atau pengkategorian variabel/fitur kontinu ke dalam tingkat kategori untuk memperjelas pola yang ada. Untuk data *clickstream* atau urutan tindakan, *embedding* urutan menggunakan *hidden state* LSTM atau Transformer dapat ditambahkan untuk memperluas ruang fitur. Secara umum, pemilihan fitur hasil augmentasi didasarkan pada teori atau studi empiris yang menunjukkan bahwa fitur-fitur baru tersebut valid dalam meningkatkan deteksi potensi perilaku kecurangan.

Representasi vektor fitur memegang peranan krusial terhadap kinerja model pembelajaran mesin (*machine learning*). Faktanya, representasi fitur yang tidak memadai terkadang dapat menurunkan akurasi prediksi secara lebih drastis dibandingkan dengan jumlah sampel pelatihan yang terbatas. Heaton menekankan bahwa pembuatan fitur baru yang diturunkan dari fitur yang sudah ada dapat memperkaya representasi input dan meningkatkan kinerja model. Augmentasi ruang fitur ini meningkatkan keragaman dan kandungan informasi dalam dataset. Beberapa penelitian mengilustrasikan pendekatan ini. Micen-ková menerapkan fungsi penilaian pencilan (*outlier*) untuk menghasilkan fitur yang telah ditransformasi, sehingga memperkaya representasi fitur untuk proses pembelajaran. Devries dan Taylor mentransformasi vektor data yang ada di dalam ruang fitur yang telah

dipelajari dengan cara mengodekan urutan, mengekstrak vektor konteks, dan menerapkan transformasi seperti injeksi derau (*noise injection*).

Zopluoglu mengubah respons butir soal menjadi pola respons nominal untuk memperluas ruang fitur, sementara Zhou dan Jiao melakukan augmentasi data pada tingkat model dasar dalam kerangka kerja *stacking ensemble* dengan menyertakan fitur-fitur seperti jumlah percobaan pengerjaan tes dan statistik ringkasan. Zhou dan Jiao juga memasukkan hasil deteksi anomali untuk mengaugmentasi ruang fitur pada tingkat meta-model dalam pemodelan *stacking* guna meningkatkan akurasi deteksi kecurangan. Jiao dkk. menyertakan ukuran kesesuaian individu (*person-fit measures*) untuk mengaugmentasi ruang fitur bagi deteksi kecurangan. Secara umum, strategi-strategi ini membantu mengaugmentasi ruang fitur guna meningkatkan ketangguhan dan kemampuan generalisasi model pembelajaran mesin untuk deteksi kecurangan.

Di sisi lain, augmentasi ruang data bertujuan untuk meningkatkan ukuran sampel sekaligus keterwakilan dataset sehingga memperbaiki pelatihan dan generalisasi model dengan cara menciptakan titik data sintetis. Hal ini sering kali dicapai melalui teknik pengambilan sampel ulang (*resampling*), seperti *oversampling* pada kelas minoritas atau pembuatan sampel sintetis. Dengan memperkaya dataset melalui contoh tambahan atau penyeimbangan ulang, augmentasi ruang data membantu mengatasi masalah seperti ketidakseimbangan kelas dan kurangnya keragaman dalam data pelatihan.

Salah satu teknik yang umum digunakan adalah *Synthetic Minority Oversampling Technique* (SMOTE), yang menghasilkan sampel baru untuk kelas minoritas melalui interpolasi di antara sampel minoritas yang sudah ada. Pendekatan *Adaptive Synthetic Sampling* untuk pembelajaran dengan data tidak seimbang (ADASYN) sebagai pengembangan adaptif dari SMOTE—berfokus pada pembuatan sampel sintetis untuk kasus minoritas yang lebih sulit dipelajari (misalnya, kasus yang posisinya dekat dengan kelas mayoritas atau berada di wilayah yang tumpang tindih) guna meningkatkan representasi kasus-kasus yang sulit diklasifikasikan tersebut.

Berbagai penelitian telah mengeksplorasi metode pengambilan sampel ulang ini untuk keperluan deteksi kecurangan. Sebagai contoh, Zhou dan Jiao menerapkan *oversampling*, *undersampling*, dan SMOTE berbasis algoritma KNN untuk menghasilkan data sintetis pelaku kecurangan serta mengatasi ketidakseimbangan kelas dalam deteksi kecurangan. Demikian pula, Tiong dan Lee melakukan augmentasi pada dataset mereka dengan menambahkan 60 sampel sintetis yang merepresentasikan perilaku menyimpang dalam konteks ujian daring. Secara keseluruhan, augmentasi dalam penelitian-penelitian tersebut memperkaya dataset dengan penambahan sampel baru, sehingga meningkatkan ketangguhan (*robustness*) model.

Metode untuk Data Tekstual

Augmentasi data untuk data tekstual juga diklasifikasikan ke dalam dua kategori: ruang fitur dan ruang data. Untuk kategori ruang fitur, Bayer menyoroti teknik interpolasi dan injeksi derau (*noise injection*). Wang dan Yang menggunakan *embedding* yaitu representasi vektor numerik padat yang menangkap makna semantik teks untuk melakukan augmentasi pada ruang fitur. Dalam ranah data, augmentasi data dapat diterapkan pada tingkat karakter

(misalnya, menukar huruf untuk menyimulasikan kesalahan ketik), tingkat kata (misalnya, penggantian sinonim), tingkat frasa (misalnya, parafrasa klausa atau idiom tertentu), dan tingkat dokumen (misalnya, menerjemahkan dan menerjemahkan balik keseluruhan esai). Bagian berikut menyoroti metode yang lebih umum digunakan dalam augmentasi data untuk data teks. Untuk tinjauan lengkap mengenai metode augmentasi data, pembaca dapat merujuk pada Feng, Li, dan Bayer.

Memasukkan *embedding* kata atau kalimat ke dalam ruang fitur merupakan metode augmentasi data yang umum dilakukan. Ketika metode pembelajaran mesin tradisional atau pembelajaran mendalam (*deep learning*) mengandalkan fitur hasil rekayasa, baik *embedding* statis maupun kontekstual dapat disertakan untuk memperluas ruang fitur. *Embedding* statis adalah representasi vektor kata yang tetap sama terlepas dari konteksnya. Model seperti Word2Vec, GloVe, dan FastText menghasilkan *embedding* ini dengan melatih model pada korpus teks berskala besar, serta menetapkan satu vektor tunggal untuk setiap kata dalam kosakata yang menangkap informasi semantik dan sintaksis secara umum. Meskipun *embedding* statis efisien secara komputasi dan berguna untuk berbagai tugas NLP, metode ini tidak dapat menangkap polisemi atau variasi makna halus yang bergantung pada konteks. Keterbatasan inilah yang mendorong pengembangan *embedding* kontekstual. *Embedding* kontekstual adalah representasi vektor kata atau token yang menangkap maknanya secara dinamis berdasarkan konteks di sekitarnya dalam sebuah kalimat atau dokumen; hal ini berbeda dengan *embedding* statis tradisional yang menetapkan satu vektor tetap untuk setiap kata.

Dengan mempertimbangkan kata-kata di sekitarnya dan struktur linguistik secara keseluruhan, *embedding* kontekstual dapat membedakan berbagai makna dari kata yang sama serta mengodekan informasi semantik dan sintaksis yang lebih kaya. *Embedding* ini biasanya dihasilkan oleh model jaringan saraf tiruan (*neural network*) yang mendalam, seperti *transformer*, yang memproses keseluruhan urutan dan memodelkan ketergantungan antar-kata, sehingga menghasilkan representasi yang beradaptasi dengan berbagai konteks dan secara signifikan meningkatkan kinerja dalam tugas pemahaman bahasa alami (*natural language understanding*). Model yang digunakan untuk menghasilkan embedding kontekstual antara lain mencakup *embedding* dari model bahasa, BERT, RoBERTa, DeBERTa, seri GPT, dan T5 (*Text-to-Text Transfer Transformer*). Daftar lengkap model embedding dengan performa terbaik dapat ditemukan di *Embedding Leaderboard* pada *https://huggingface.co/spaces/mteb/leaderboard*.

Dalam ranah data, augmentasi data tingkat kata mencakup penggantian sinonim, pertukaran acak (*random swap*) yang menukar posisi pertanyaan secara acak, penghapusan acak, dan penyisipan acak sebagai bentuk induksi derau/noise. Untuk augmentasi data tingkat frasa dan dokumen, *back-translation* yaitu penerjemahan dua arah (*round-trip translation*) sering digunakan untuk menambah jumlah sampel. *Back-translation* sangat bermanfaat dalam deteksi kecurangan karena metode ini menghasilkan parafrasa berkualitas tinggi yang menyerupai hasil alat *spinning* atau strategi penulisan ulang manual yang sering digunakan peserta tes untuk menyamarkan plagiarisme.

Dengan melatih model menggunakan variasi sintetis ini, model pembelajaran mesin menjadi sangat efektif dalam mendeteksi konten hasil parafrasa yang menyimpang dari materi sumber aslinya. Terkait perkembangan terkini, AI generatif/genAI dapat digunakan untuk membuat lebih banyak sampel jawaban guna augmentasi data melalui berbagai strategi pemberian *prompt* dan *fine-tuning*. Meskipun augmentasi data merupakan praktik standar dalam tugas-tugas NLP umum, penerapannya sangat krusial dalam ranah keamanan tes untuk mendeteksi esai buatan AI dan plagiarisme yang disamarkan. Dalam situasi operasional, memperoleh korpus besar yang telah diberi label berisi respons kecurangan terverifikasi sering kali sulit dilakukan, sehingga menyebabkan ketidakseimbangan kelas yang parah.

Selain itu, peserta tes yang cerdik mungkin menggunakan strategi adversarial seperti melakukan parafrasa atau menerjemahkan konten secara bolak-balik untuk menghindari pemeriksaan kemiripan yang sederhana. Akibatnya, teknik augmentasi data untuk teks bukan sekadar upaya meningkatkan ukuran sampel. Teknik ini justru berfungsi untuk menyimulasikan perilaku penghindaran yang bersifat adversarial tersebut, sehingga memungkinkan pelatihan model deteksi tangguh yang mampu mengidentifikasi variasi halus dari pelanggaran integritas akademik.

10.4 APLIKASI EMPIRIS AUGMENTASI DATA

Untuk mengilustrasikan penerapan metode augmentasi data yang dibahas dalam bab ini, bagian ini menyajikan contoh empiris berdasarkan penelitian Zhou dan Jiao serta Jiao et al. untuk menunjukkan peningkatan akurasi deteksi kecurangan ketika data ditambah (diaugmentasi) dari segi kuantitas dan kategori.

Data

Kumpulan data yang digunakan untuk demonstrasi ini tersedia untuk umum dan berasal dari ujian lisensi berskala besar yang telah menandai adanya potensi kasus kecurangan. Namun, tidak ada informasi yang diberikan mengenai jenis perilaku kecurangan yang dilakukan oleh peserta ujian yang ditandai tersebut. Format ujian mencakup 170 butir soal dikotomis yang diberikan kepada 1.636 peserta ujian; dari jumlah tersebut, 46 orang diidentifikasi sebagai pihak yang kemungkinan melakukan kecurangan. Kumpulan data tersebut memuat respons butir soal dan waktu pengerjaan untuk setiap dari 170 butir soal, total waktu pengerjaan per peserta ujian, serta jumlah upaya pengerjaan ujian. Secara keseluruhan, kumpulan data asli tersebut terdiri dari 342 variabel.

Karena hanya sebagian kecil peserta ujian yaitu 2,81% dari 1.636 peserta yang ditandai sebagai pihak yang kemungkinan melakukan kecurangan, kumpulan data tersebut menunjukkan ketidakseimbangan kelas yang signifikan. Untuk mengatasi masalah ini, teknik *resampling* diterapkan pada set pelatihan, di mana masing-masing dari ketiga studi tersebut bereksperimen dengan metode augmentasi data yang berbeda.

Kriteria Evaluasi

Mengingat adanya masalah ketidakseimbangan kelas dalam deteksi kecurangan, akurasi keseluruhan yang didefinisikan sebagai proporsi kasus yang diklasifikasikan dengan benar bukanlah metrik yang tepat untuk mengevaluasi kinerja deteksi. Untuk masalah

klasifikasi dengan data yang tidak seimbang, kinerja model terutama dievaluasi menggunakan skor F1, yaitu rata-rata harmonik dari *recall* dan presisi.

Selain itu, *recall*, presisi, dan tingkat positif palsu (*false-positive rate* atau FPR) juga dilaporkan. *Recall* (atau sensitivitas) mengukur proporsi positif sejati (pelaku kecurangan yang teridentifikasi dengan benar) dari keseluruhan pelaku kecurangan yang sebenarnya. Presisi mengukur proporsi positif sejati di antara semua peserta tes yang diprediksi sebagai pelaku kecurangan. FPR mencerminkan proporsi bukan pelaku kecurangan yang secara keliru diklasifikasikan sebagai pelaku kecurangan. Setiap kriteria evaluasi didefinisikan sebagai berikut:

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$SkorF1 = \frac{2 \times Presisi \times Recall}{Presisi + Recall} = \frac{2TP}{2TP + FP + FN}$$

$$TingkatPositifPalsu(FPR) = \frac{FP}{FP + TN}$$

di mana TP = *true positive* (positif benar), FP = *false positive* (positif semu), FN = *false negative* (negatif semu), dan TN = *true negative* (negatif benar).

Augmentasi Data

Studi 1 yang didasarkan pada Zhou dan Jiao meneliti penggunaan algoritma *machine learning* ansambel *stacking* untuk mendeteksi perilaku kecurangan dengan menganalisis respons butir soal dan waktu respons dari 170 butir soal, serta data peserta tes yang telah diaugmentasi. Studi ini mengeksplorasi augmentasi data pada ruang fitur dan ruang data. Pada ruang fitur, fitur-fitur yang diaugmentasi mencakup jumlah percobaan pengerjaan tes dan beberapa statistik ringkasan (total skor tes dan total waktu respons untuk seluruh 170 butir soal), serta nilai rata-rata, median, maksimum, dan minimum waktu respons butir soal, dengan total 7 fitur augmentasi. Penyertaan fitur-fitur augmentasi ini merupakan bentuk augmentasi data yang bertujuan memberikan informasi tingkat lebih tinggi di luar apa yang ditangkap oleh respons butir soal secara individual. Secara keseluruhan, terdapat 347 variabel yang disertakan dalam dataset.

Untuk melakukan augmentasi pada ruang data, studi ini menguji beberapa strategi *resampling* guna mengatasi ketidakseimbangan kelas, termasuk *oversampling*, *undersampling*, dan kombinasi keduanya (yang disebut sebagai *dual resampling*). Secara khusus, fungsi SMOTE digunakan untuk *oversampling* dan fungsi RandomUnderSampler digunakan untuk *undersampling* dalam Python. Untuk membandingkan pengaruh augmentasi data terhadap kinerja model, diuji berbagai rasio kelas antara peserta yang tidak curang dan

yang curang, termasuk 1:1, 2:1, 5:1, dan 10:1, melalui strategi *resampling* yang berbeda. Semua strategi augmentasi data diterapkan pada tingkat model dasar (*base-model*) dalam model ansambel *stacking*. Meta-model (model tingkat lebih tinggi yang menggabungkan hasil dari model-model dasar) dikembangkan hanya dengan menggunakan keluaran dari model-model dasar yang memanfaatkan fitur-fitur augmentasi tersebut (lihat Zhou dan Jiao untuk rinciannya).

Studi 2 yang didasarkan pada Zhou dan Jiao meneliti augmentasi data pada ruang fitur dalam kerangka kerja ansambel *blending* di tingkat meta-model. Secara khusus, empat teknik deteksi anomali digunakan untuk menghasilkan skor *outlier*; skor-skor ini kemudian digabungkan dengan fitur-fitur paling informatif yang dipilih dari dataset asli melalui empat metode seleksi fitur. Empat algoritma deteksi anomali yang umum digunakan yaitu *Isolation Forest, Elliptic Envelope, One-Class SVM*, dan DBSCAN digunakan untuk menangkap berbagai jenis pencilan. Selain itu, fitur-fitur paling efektif (total 33 fitur) yang diidentifikasi untuk pelatihan model dasar dipilih guna menambah data dalam ruang fitur bagi pengembangan meta-model. Saat melatih model dasar, teknik *resampling* baik *oversampling* (SMOTE) dengan rasio 0,4 maupun *undersampling* (RandomUnderSampling) dengan rasio 0,5 digunakan untuk menambah data pada kelompok pelaku kecurangan serta menyeimbangkan ukuran sampel antara kedua kelas tersebut.

Studi 3 memperluas ruang fitur dengan menyertakan ukuran anomali maupun ukuran kesesuaian respons individu (*person-fit measures*). Keempat metode deteksi *anomali Isolation Forest, Elliptic Envelope, Local Outlier Factor*, dan *One-Class SVM* menghasilkan skor pencilan bagi setiap peserta tes. Selain itu, 29 ukuran kesesuaian respons individu yang memperluas cakupan penelitian Ranger digunakan untuk menambah ruang fitur. Ukuran-ukuran tersebut meliputi indeks berbasis Guttman: statistik G, *Norm Conformity Index* (NCI; Tatsuoka); indeks berbasis kesesuaian (*agreement*): *Agreement Index* (A), *Disagreement Index* (D), *Dependability Index* (EI); serta indeks berbasis korelasi/kovarians: korelasi *point-biserial*, *Caution Index* atau C.Sato, C, Gnormed, U3 dan U3 terstandarisasi/ZU3, H^T, serta enam indeks yang diusulkan oleh Tatsuoka, yaitu ECI1, ECI2, ECI3, ECI4, ECI5, dan ECI6.

Selain itu, indeks berbasis *likelihood* juga digunakan untuk augmentasi data, mencakup L_0, L_z, dan L_z. Nilai *mean square outfit* dan *infit* serta statistik Z untuk responden disertakan sebagai fitur tambahan setelah data respons butir disesuaikan dengan model Rasch. Tiga indeks berbasis waktu respons turut memperkaya ruang fitur, meliputi statistik KT dari Ranger dkk. dan Man dkk, serta statistik Z-kuadrat. Mengingat setiap indeks kesesuaian responden (*person-fit index*) menangani jenis perilaku respons yang menyimpang secara berbeda, penyertaan ke-29 indeks ini diharapkan dapat memperkaya ruang fitur dan menghasilkan deteksi kecurangan yang lebih akurat. Dengan mengacu pada rasio *resampling* yang digunakan oleh Zhou dan Jiao, penelitian ini menerapkan rasio *resampling* yang sama, yaitu 0,4 untuk *over-sampling* SMOTE dan 0,5 untuk *random undersampling*.

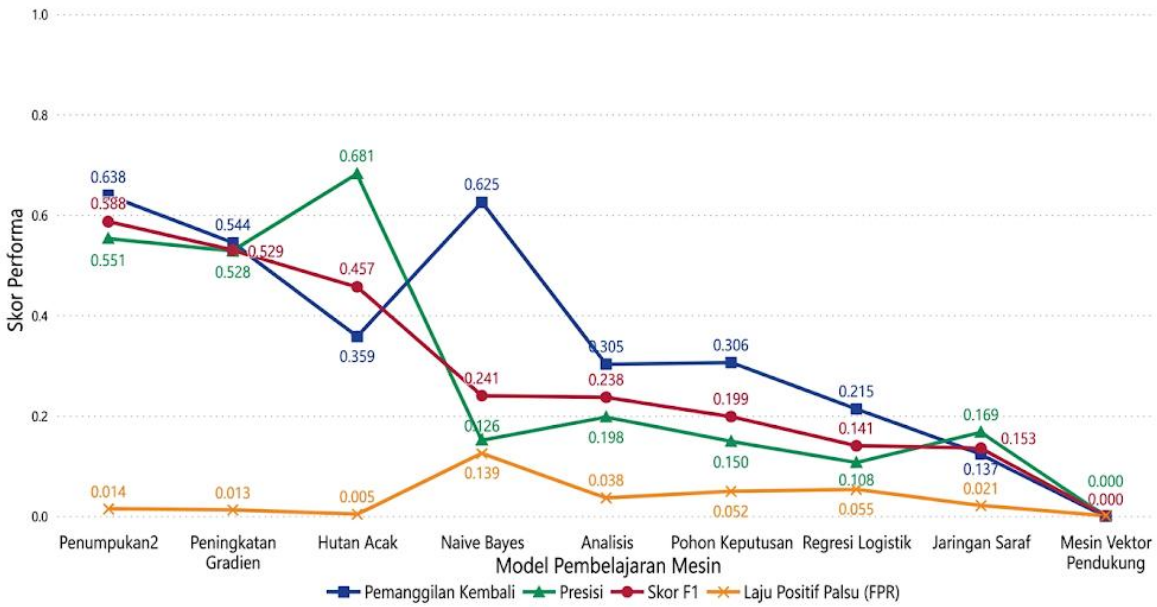
Hasil

Studi 1 menguji enam model pembelajaran mesin termasuk *Decision Tree*, SVM, regresi logistik, *Naïve Bayes*, analisis diskriminan, dan jaringan saraf tiruan serta dua model ensemble

(*random forest* dan *gradient boosting*) sebagai model dasar. Prediksi dari kedelapan model dasar ini kemudian digunakan untuk menyusun sebuah meta-model. Untuk pengembangan meta-model, data masukan diuji secara bertahap, mulai dari keluaran seluruh model dasar hingga hanya dua model dasar dengan kinerja terbaik.

Gambar 10.1, yang diadaptasi dari Gambar 8 dalam Zhou dan Jiao, merangkum model dengan kinerja terbaik yaitu meta-model yang disusun (*stacked*) berdasarkan dua model dasar terbaik serta membandingkannya dengan model-model dasar tersebut. Meta-model terbaik ini menggunakan kumpulan fitur yang menggabungkan respons butir soal dan statistik ringkasan, dengan rasio *undersampling* 10:1 dan penggunaan analisis diskriminan untuk proses *stacking* dalam meta-model.

Meta-model yang disusun berdasarkan dua model dasar terbaik tersebut mengungguli seluruh pengklasifikasi dasar individual; model ini mencapai skor F1 sebesar 0,586, *recall* 0,638, presisi seimbang 0,551, dan FPR rendah sebesar 0,014. Secara keseluruhan, di antara semua model yang dievaluasi, meta-model stacking berbasis analisis diskriminan yang menggunakan respons butir soal dan statistik ringkasan yang diperkaya (*augmented*) dengan rasio undersampling 10:1 mencapai skor F1 tertinggi dan kinerja seimbang. Semua metrik utama (F1, *recall*, dan presisi) melampaui angka 0,5, yang menunjukkan efektivitas model dalam mendeteksi pelaku kecurangan.

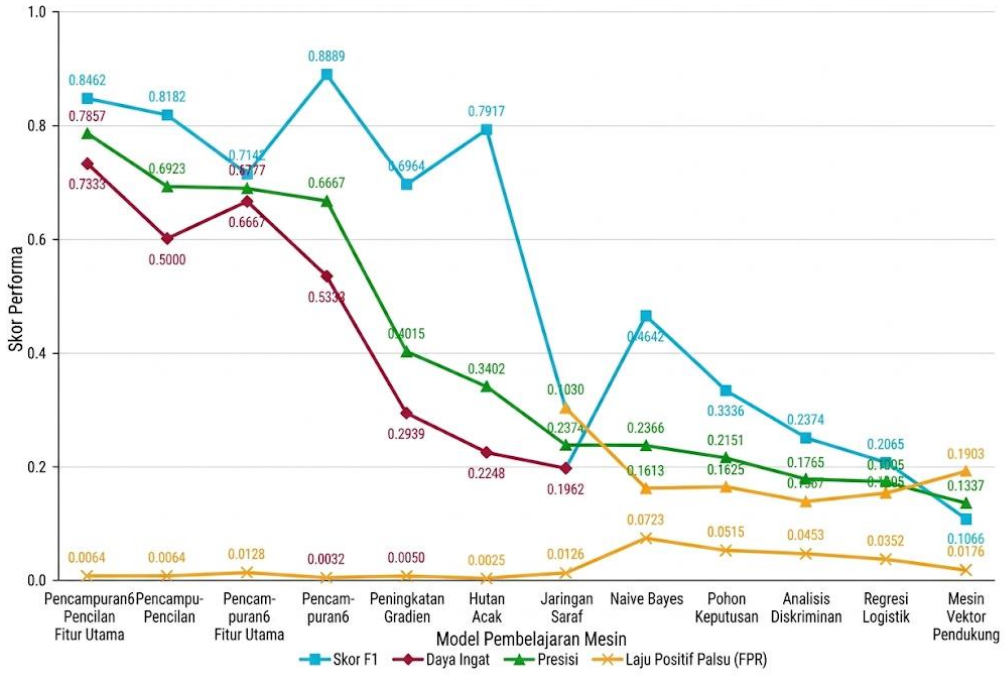


Gambar 10.1 Perbandingan antara model-model dasar dan meta-model yang menggunakan analisis diskriminan berdasarkan respons butir soal dan statistik ringkasan. Diadaptasi dari Gambar 8 dalam Zhou dan Jiao.

Studi 2 melatih delapan model dasar yang sama dengan yang dilatih pada Studi 1, dengan tingkat SMOTE sebesar 0,4 dan tingkat *Random Undersampling* sebesar 0,5 untuk menangani ketidakseimbangan kelas. Selanjutnya, studi ini melatih sebuah meta-model menggunakan metode *blending*, yang merupakan kasus khusus dari *stacking*. Sepuluh fitur paling efektif

dipilih dan digunakan sebagai fitur untuk melatih meta-model tersebut guna keperluan augmentasi data. Selain itu, empat skor outlier yang diperoleh dari *Isolation Forest*, *Elliptic Envelope*, *One-Class SVM*, dan DBSCAN digunakan sebagai data tambahan (augmentasi) untuk pelatihan meta-model. Di samping data tambahan tersebut, meta-model dilatih menggunakan berbagai variasi data masukan, mulai dari keluaran seluruh model dasar hingga hanya keluaran dari dua model dasar dengan kinerja terbaik.

Gambar 10.2, yang diadaptasi dari Gambar 5 dalam Zhou dan Jiao, mengilustrasikan kinerja meta-model *blending* terbaik yang menggabungkan dua jenis data tambahan, serta kinerja kedelapan model dasar tersebut. Gambar ini juga menampilkan kinerja model dasar dan meta-model dengan berbagai data masukan yang berbeda. Secara keseluruhan, meta-model *blending* yang didasarkan pada enam model dasar berkinerja terbaik dengan kedua jenis data tambahan tersebut mencapai skor F1 tertinggi sebesar 0,7857, keseimbangan antara *recall* dan *precision* (keduanya di atas 0,5), serta nilai FPR yang relatif rendah sebesar 0,0064. Meta-model yang hanya menggunakan satu jenis data tambahan baik itu ukuran *outlier* maupun fitur efektif terpilih tetap menunjukkan kinerja yang lebih baik dibandingkan pendekatan *blending* tanpa augmentasi data. Jika dibandingkan dengan model dasar termasuk model *ensemble* seperti *gradient boosting* dan *random forest* seluruh meta-model berbasis *blending* menunjukkan kinerja yang jauh lebih unggul.



Gambar 10.2 Perbandingan model antara model dasar dan meta-model dengan serta tanpa augmentasi data. Diadaptasi dari Gambar 5 karya Zhou dan Jiao.

Studi 3 melatih delapan model dasar yang sama untuk pengembangan meta-model. Dalam melatih model dasar tersebut, selain menggunakan 347 fitur yang dipakai pada Studi 1 dan 2, studi ini melakukan augmentasi pada ruang fitur untuk pelatihan model dasar dengan menambahkan 29 ukuran kecocokan (*fit measures*) dan 4 skor anomali. Delapan metode

seleksi fitur diterapkan untuk mengidentifikasi 10 fitur paling efektif dari setiap metode. Setelah menghapus duplikat, tersisa 55 fitur unik. Dengan menggabungkan keluaran model dasar dan fitur-fitur tambahan tersebut, total 63 fitur digunakan untuk melatih meta-model. Meta-model tersebut mencapai skor F1 sebesar 0,85, recall 0,78, presisi 0,99, dan FPR di bawah 0,001 pada ketiga jenis meta-model: *random forest, gradient boosting*, dan analisis diskriminan. Dibandingkan dengan penelitian sebelumnya yang menggunakan dataset serupa, meta-model yang dikembangkan dalam penelitian ini menunjukkan kinerja yang lebih unggul.

10.5 DETEKSI KECURANGAN BERBASIS MACHINE LEARNING DAN AI

Kemajuan terkini dalam *machine learning* dan AI generatif (genAI) telah menghadirkan perspektif baru sekaligus tantangan baru dalam deteksi kecurangan. Bab ini memperkenalkan metode deteksi kecurangan yang menggunakan teknik *machine learning*, model *deep learning*, dan model bahasa. Selain itu, melalui serangkaian studi empiris, ditunjukkan bahwa augmentasi data yang mengintegrasikan metode psikometrik dan algoritma *machine learning* dapat meningkatkan akurasi deteksi. Terakhir, bab ini secara singkat memperkenalkan metode untuk mendeteksi esai yang dihasilkan oleh genAI.

Meskipun metode *machine learning* yang bersifat *supervised* (terawasi) maupun *unsupervised* (tidak terawasi) dapat digunakan untuk mendeteksi kecurangan, isu penting mengenai ketidakseimbangan kelas (*class imbalance*) perlu ditangani ketika kelas pelaku kecurangan yang sering kali merupakan kelas minoritas hanya mencakup sebagian kecil dari total sampel. Lebih lanjut, analisis psikometrik tradisional mengenai ukuran kesesuaian individu (*person-fit measures*) dapat digunakan untuk mendeteksi kecurangan.

Ukuran-ukuran tersebut dapat disertakan sebagai data tambahan (*augmented data*) dalam deteksi kecurangan berbasis *machine learning*, melengkapi ukuran deteksi anomali yang dihasilkan oleh *machine learning* itu sendiri. Bab ini menunjukkan efektivitas augmentasi data untuk data tabular yang didasarkan pada statistik ringkasan, metode deteksi anomali, dan ukuran kesesuaian individu dari analisis psikometrik tradisional. Namun, metode augmentasi data lainnya dapat dieksplorasi lebih jauh, seperti transformasi data mentah, penambahan suku interaksi untuk memperhitungkan hubungan non-linear antar-data mentah, serta penambahan *noise* (derau) untuk meningkatkan keragaman atau variabilitas data yang ada.

Selain itu, genAI memunculkan jenis kecurangan baru, karena alat-alat ini mampu menghasilkan respons butir soal termasuk esai dan jawaban uraian Panjang yang menyerupai respons autentik dari peserta tes. Berbeda dengan bentuk kecurangan tradisional, respons yang dihasilkan AI bisa jadi sepenuhnya orisinal namun sangat meyakinkan, sehingga sulit dideteksi menggunakan metode konvensional. Model bahasa dapat memainkan peran penting dalam mendeteksi respons buatan AI pada asesmen berskala besar. Berbagai studi melaporkan bahwa, secara umum, model bahasa seperti RoBERTa dan ELECTRA memiliki kinerja lebih baik dibandingkan model *machine learning supervised* berbasis fitur dalam mendeteksi esai buatan AI. Bab ini terutama berfokus pada deteksi kecurangan terhadap esai dan respons teks buatan AI untuk butir soal uraian (*constructed-response items*). Namun, perlu diakui bahwa AI generatif (genAI) juga dapat digunakan untuk memanipulasi format soal lain seperti soal

pilihan ganda atau jawaban singkat dengan menyarankan jawaban atau pola yang kemungkinan besar benar, sehingga semakin mempersulit upaya pendeteksian.

Perlu dicatat bahwa kemajuan terkini dalam pendeteksian kecurangan semakin mengedepankan model hibrida yang menggabungkan teknik *unsupervised* (tanpa pengawasan) dan *supervised* (dengan pengawasan), serta pendekatan *ensemble learning*. Bukti empiris menunjukkan kinerja model *stacking ensemble* yang lebih unggul dibandingkan model dasar (*base model*) secara individual. Model-model ini memadukan kelebihan dari berbagai metode: pembelajaran *unsupervised* mampu mengungkap pola perilaku tidak wajar yang luas dan sebelumnya tidak diketahui tanpa memerlukan data berlabel, sementara pembelajaran *supervised* dapat mendeteksi anomali yang lebih spesifik dan mendetail berdasarkan pengetahuan awal serta contoh kecurangan yang telah diberi label.

Metode *ensemble*, yang menggabungkan beberapa pengklasifikasi atau detektor, semakin meningkatkan ketangguhan sistem dengan mengurangi risiko *overfitting* terhadap perilaku tertentu serta memperbaiki kemampuan generalisasi di berbagai konteks. Pendekatan hibrida dan *ensemble* ini semakin umum digunakan dalam pendeteksian kecurangan karena menawarkan kerangka kerja komprehensif untuk mengidentifikasi ketidakwajaran, baik pada tingkat individu maupun kelompok, termasuk yang bersifat samar.

Perlu dicatat bahwa deteksi kecurangan merupakan tugas yang menantang dalam analisis psikometrik. Perilaku curang dapat mengambil berbagai bentuk. Dalam praktiknya, sulit untuk membedakan perilaku curang dari perilaku respons menyimpang lainnya seperti tebakan cerdas atau faktor lain yang dapat menyebabkan masalah kesesuaian model. Analisis respons item mungkin tidak mengungkapkan informasi yang cukup untuk mendeteksi perilaku curang. Kami sangat menganjurkan penggunaan data proses dan integrasi data dari berbagai sumber dan berbagai modalitas untuk melakukan deteksi kecurangan guna meningkatkan kekuatan dan akurasinya. Ketika data menjadi lebih besar dan lebih kompleks, kemajuan terbaru dari pembelajaran mesin dan teknologi AI pasti akan menjadi paradigma analisis yang lebih layak untuk deteksi kecurangan ketika metode dan model psikometrik parametrik mungkin tidak memadai.

Bab ini berfokus pada mendeteksi potensi pelaku kecurangan, daripada mengidentifikasi item yang dikompromikan, meskipun mendeteksi item yang dikompromikan secara efektif sangat penting untuk menjaga kualitas bank item. Pembaca yang tertarik untuk mendeteksi item yang telah dimanipulasi atau deteksi simultan pengetahuan sebelumnya baik pada item maupun peserta tes dapat merujuk pada studi yang dilakukan oleh peneliti lain.

BAB 11
PENGGUNAAN AI DALAM PENILAIAN GLOBAL

11.1 TANTANGAN PENILAIAN DI ERA TRANSFORMASI PENDIDIKAN

Globalisasi telah mengubah sistem pendidikan di seluruh dunia secara mendasar, membentuk ulang kebijakan, kurikulum, dan akses terhadap peluang pembelajaran. Pergeseran ini telah mendorong terciptanya ekonomi berbasis pengetahuan yang saling terhubung, di mana peserta didik dan institusi saling bertukar sumber daya lintas batas negara, sehingga memajukan kolaborasi dan daya saing global. Namun, meskipun globalisasi telah mendukung pemerataan pendidikan yang lebih baik, hal ini juga dibarengi dengan transformasi digital yang pesat di dunia pendidikan. Transformasi ini menghasilkan data pembelajaran dalam volume yang sangat besar, yang tidak hanya membuka peluang baru tetapi juga memperberat tantangan bagi sistem penilaian.

Salah satu tantangan utama terletak pada skala dan heterogenitas data pendidikan masa kini. Pertumbuhan *Massive Open Online Courses* (MOOC), kelas virtual, dan program lintas batas menghasilkan aliran data yang berkelanjutan berupa teks, log interaksi, dan konten multimedia yang tidak dapat diproses secara efisien oleh sistem penilaian tradisional. Akibatnya, pendekatan manual dan berbasis kertas mengalami kesulitan dalam memberikan umpan balik yang tepat waktu, konsisten, dan personal.

Pada saat yang sama, kurikulum internasional dan penilaian berskala besar seperti *International Baccalaureate* (IB) dan *Programme for International Student Assessment* (PISA) meningkatkan tuntutan akan evaluasi yang terstandarisasi namun tetap peka terhadap konteks, sekaligus menyoroti keterbatasan kemampuan adaptasi sistem yang ada saat ini dalam menghadapi lingkungan yang beragam secara linguistik dan budaya.

AI sebagai Katalisator bagi Inovasi Penilaian dan Reformasi Pembelajaran Global

Untuk mengatasi tantangan-tantangan ini dan membuka potensi penuh pendidikan yang terglobalisasi, kecerdasan buatan (AI) telah muncul sebagai kekuatan transformatif yang mampu mengubah tidak hanya praktik penilaian, tetapi juga arsitektur sistem pembelajaran internasional itu sendiri. Dengan memanfaatkan teknik komputasi tingkat lanjut, AI menghadirkan kemungkinan baru untuk mencapai skalabilitas, presisi, dan kemampuan adaptasi yang tidak dapat dicapai oleh metode penilaian tradisional. Sistem berbasis AI memanfaatkan teknologi seperti pemrosesan bahasa alami (NLP), pembelajaran mesin (*machine learning* atau ML), pembelajaran mendalam (*deep learning* atau DL), dan jaringan saraf tiruan (*neural networks*), yang memungkinkannya memproses kumpulan data yang sangat besar dan beragam secara efisien.

Yang terpenting, peran AI melampaui sekadar penerapan penilaian; AI secara fundamental berperan dalam mentransformasi pembelajaran. Sistem-sistem ini mampu menganalisis data tidak terstruktur (misalnya, respons terbuka, catatan interaksi perilaku), memberikan umpan balik yang dipersonalisasi, dan menyesuaikan diri secara dinamis dengan kebutuhan individu pembelajar, sehingga menutup siklus umpan balik dan menjadikan

penilaian sebagai bagian integral dan formatif dari proses pembelajaran itu sendiri. Kemampuan-kemampuan ini menempatkan AI sebagai landasan utama praktik penilaian masa depan, sekaligus meningkatkan efektivitas dan inklusivitas dalam skala global.

Dampak praktis AI dalam penilaian pendidikan semakin nyata melalui berbagai penerapan, termasuk penilaian otomatis, analitik pembelajaran dan alat diagnostik, pemodelan prediktif untuk intervensi dini, serta sistem penilaian adaptif. Secara bersama-sama, pendekatan-pendekatan ini memungkinkan evaluasi yang dapat ditingkatkan skalanya, umpan balik yang dipersonalisasi, dan dukungan instruksional yang tepat waktu di berbagai konteks pembelajaran. Penerapan baru yang sedang berkembang, seperti penilaian berbasis realitas virtual (VR) dan realitas tertambah (AR) yang menciptakan lingkungan simulasi atau lingkungan yang diperkaya secara digital semakin memperluas cakupan evaluasi hingga ke ranah keterampilan yang kompleks dan autentik.

Mendorong Transformasi Global: Dari Implikasi Sistem hingga Kontribusi Bab Ini

Di luar inovasi teknis, penggunaan AI dalam penilaian memiliki implikasi sosial yang luas dengan membentuk ekosistem pembelajaran global yang lebih adil, berbasis kompetensi, dan saling terhubung. Sistem penilaian tradisional sering kali menguntungkan institusi dan wilayah yang memiliki sumber daya memadai, sedangkan alat berbasis AI yang berbiaya rendah dan dapat ditingkatkan skalanya seperti sistem penilaian otomatis dan platform adaptif memperluas akses terhadap evaluasi canggih dan umpan balik yang dipersonalisasi bagi peserta didik di lingkungan yang kurang terlayani.

Yang tak kalah penting, alat diagnostik dan analitik prediktif berbasis AI tidak hanya sekadar mengidentifikasi kesenjangan pembelajaran, tetapi juga mendukung sistem adaptif yang responsif terhadap kebutuhan peserta didik dan terus berkembang seiring waktu; hal ini memungkinkan pendidik merancang jalur pembelajaran individual serta mengalihkan fokus dukungan dari sekadar perbaikan (remediasi) menuju pengembangan kapasitas yang berkelanjutan. Pada tingkat sistemik, AI memfasilitasi terbentuknya jaringan pembelajaran lintas negara dengan menyelaraskan metrik penilaian dan mengintegrasikan analitik yang multibahasa serta responsif secara budaya, sehingga mengubah kolaborasi internasional dari sekadar penyelarasan kebijakan menjadi inovasi pedagogis bersama dan penciptaan pengetahuan secara kolaboratif.

Lebih mendasar lagi, AI memungkinkan pergeseran paradigma dari sekadar menilai pembelajaran menjadi merancang pembelajaran melalui proses penilaian itu sendiri. Evaluasi beralih dari peristiwa sumatif yang terpisah-pisah menjadi proses adaptif yang terintegrasi, yang secara dinamis menyesuaikan konten instruksional, tingkat kesulitan, dan umpan balik bagi setiap peserta didik. Melalui lingkungan penilaian yang imersif dan multimodal, analitik berbasis AI mendukung pengukuran kompetensi kompleks seperti pemecahan masalah dan kolaborasi, sekaligus mengurangi hambatan geografis dan memungkinkan interaksi pembelajaran lintas batas. Secara keseluruhan, perkembangan ini menandai transisi dari sistem pengujian yang kaku dan terstandarisasi menuju infrastruktur penilaian yang fleksibel, dapat saling beroperasi (*interoperabel*), dan berpusat pada peserta didik.

Bab ini bertujuan menyajikan analisis global mengenai lanskap AI yang terus bertransformasi dalam bidang penilaian pendidikan. Terdapat tiga tujuan utama: (1) memberikan gambaran umum tentang kondisi terkini penilaian berbasis AI, termasuk kemajuan teknologi utama dan penerapannya di dunia nyata; (2) mengkaji bagaimana AI diterapkan di berbagai wilayah, negara, dan konteks pendidikan; serta (3) mengidentifikasi praktik-praktik baru yang muncul, tantangan utama, dan tren masa depan. Kontribusi khas bab ini terletak pada sintesis beragam penerapan AI dalam penilaian di tingkat internasional, perumusan pelajaran praktis dari studi kasus global, serta penyusunan rekomendasi yang berorientasi ke masa depan. Secara kolektif, rekomendasi ini menawarkan pertimbangan implementasi sekaligus panduan tata kelola untuk membantu sistem pendidikan dalam memanfaatkan AI secara etis dan efektif guna mentransformasi penilaian dan pembelajaran dalam skala global.

11.2 KONDISI TERKINI DAN PENERAPAN AI DALAM PENILAIAN PENDIDIKAN

Sistem Penilaian Otomatis dalam Penilaian Berskala Besar

Teknologi Utama dan Pendekatan Metodologis untuk Sistem Penilaian Otomatis

Sistem penilaian otomatis telah berkembang dari pengukuran kemiripan berbasis aturan menjadi kerangka kerja canggih yang didukung AI, yang mampu menangani respons kompleks dan multimodal. Metode awal, seperti *Levenshtein Distance* dan *Cosine Similarity*, berfokus pada tumpang tindih tekstual, sementara teknik NLP (Pemrosesan Bahasa Alami) berikutnya yang menggabungkan *word embeddings*, penandaan kelas kata (*part-of-speech tagging*), dan penguraian sintaksis memungkinkan evaluasi semantik yang lebih presisi.

Kemajuan terkini sangat bergantung pada Deep Learning (DL), khususnya model berbasis *transformer* seperti BERT (*Bidirectional Encoder Representations from Transformers*), serta arsitektur termasuk *Long Short-Term Memory* (LSTM) dan *capsule networks*, untuk meningkatkan akurasi penilaian bagi respons teks bebas. Selain teks, sistem kontemporer semakin mendukung input visual dan multimodal (seperti teks tulisan tangan, diagram, dan gambar) menggunakan *Optical Character Recognition* (OCR), metrik visi komputer, model berbasis graf, dan pengenalan simbol berbasis *Convolutional Neural Network* (CNN). *Large Language Models* (LLM) semakin memperluas kemampuan penilaian otomatis dengan mengevaluasi koherensi, argumentasi, dan bahasa spesifik-domain, bukan sekadar fitur permukaan, sementara varian *transformer* yang lebih ringan menyeimbangkan akurasi dan efisiensi komputasi dalam penerapan berskala besar. Secara keseluruhan, perkembangan ini memungkinkan sistem penilaian otomatis beroperasi secara andal dalam skala besar, menempatkannya bukan sekadar sebagai solusi teknis untuk efisiensi pemberian nilai, melainkan sebagai komponen penting dari infrastruktur penilaian berskala besar dan lintas negara.

Penerapan Global Sistem Penilaian Otomatis: Perspektif Berbasis Fitur

Di berbagai negara dan konteks penerapan, perspektif berbasis fitur menawarkan pemahaman yang lebih sistematis mengenai sistem penilaian otomatis dibandingkan deskripsi kasus yang terisolasi. Dimensi perbandingan utama mencakup pemrosesan multimodal dan

kemampuan adaptasi terhadap mata pelajaran, yang menentukan apakah sistem dapat menilai teks, diagram, dan konten visual di berbagai bidang seperti seni bahasa dan STEM; serta granularitas penilaian dan kekayaan umpan balik, yang mencerminkan pergeseran dari skor numerik menuju umpan balik berbasis rubrik dan diagnostik yang selaras dengan tujuan instruksional; ...serta kemampuan untuk memberikan penjelasan (*explainability*) yang dipadukan dengan tata kelola yang melibatkan peran manusia (*human-in-the-loop*); hal ini meningkatkan transparansi, kepercayaan, dan keadilan melalui alur kerja penilaian hibrida.

Implementasi berskala besar yang efektif juga bergantung pada upaya menyeimbangkan skalabilitas dengan adaptasi lokal, yang menuntut adanya pertimbangan kompromi antara akurasi, biaya, dan kegunaan, sembari menyelaraskan model dengan kurikulum, bahasa, dan praktik pedagogis setempat. Secara keseluruhan, dimensi-dimensi perbandingan ini mencerminkan tidak hanya pilihan desain teknis, tetapi juga perbedaan prioritas institusional terkait validitas, tata kelola, dan keadilan dalam konteks penilaian berskala besar.

Studi Kasus Nasional

Studi kasus berikut ini mengilustrasikan bagaimana sistem penilaian otomatis beroperasi dalam konteks pendidikan dan budaya yang berbeda. Ringkasan perbandingan kasus-kasus tersebut disajikan dalam Tabel 11.1. Di Amerika Serikat, E-rater®—yang dikembangkan oleh Educational Testing Service untuk ujian seperti GRE® dan TOEFL®—merupakan sistem acuan dalam menyeimbangkan persyaratan tata kelola dengan skalabilitas pada penilaian yang memiliki konsekuensi tinggi (*high-stakes assessment*). Agar dapat beroperasi dalam skala masif sekaligus beradaptasi dengan berbagai konteks, E-rater® memanfaatkan ratusan fitur linguistik berbasis teknik NLP, bukan pendekatan *deep learning* (DL) *end-to-end*.

Tabel 11.1 Analisis Perbandingan Studi Kasus Sistem Penilaian Otomatis Global

Dimensi	Studi Kasus 1: E-rater® (Amerika Serikat)	Studi Kasus 2: EICTS (Tiongkok)
Konteks dan tujuan	Penilaian berisiko tinggi (*high-stakes assessment*): Digunakan untuk ujian seperti GRE® dan TOEFL®; mengutamakan tata kelola, skalabilitas, dan validitas dalam konteks berskala besar.	Sistem tutor cerdas: Dirancang untuk pembelajaran menulis karangan bahasa Mandarin; mengutamakan umpan balik instruksional dan tingkat kedalaman (*granulartitas*) diagnosis.
Teknologi utama	Memanfaatkan ratusan fitur linguistik berbasis NLP—alih-alih *deep learning* (DL) *end-to-end*—untuk memastikan skalabilitas operasional sekaligus menjaga keterjelasan konstruk.	Mengintegrasikan OCR yang dilatih secara khusus untuk memproses tulisan tangan serta BERT untuk analisis semantik.

Kekuatan dan fokus psikometris	Validitas dan tata kelola yang melibatkan manusia (*human-in-the-loop*): 1. Memastikan validitas skor melalui model verifikasi hibrida manusia-mesin. 2. Memberikan umpan balik yang dapat ditindaklanjuti terkait tata bahasa dan gaya penulisan (melalui Criterion®).	Multimodalitas dan tingkat kedalaman diagnosis: 1. Memproses input tulisan tangan (multimodal). 2. Menerapkan kerangka penilaian empat lapis (bahasa, ekspresi, anomali, wacana) untuk menghasilkan laporan diagnosis yang terperinci.
Keterbatasan dan tantangan	1. Potensi kurangnya representasi konstruk: mungkin memberikan bobot yang tidak proporsional pada fitur organisasi dibandingkan dengan penilai manusia. 2. Kinerja sangat bergantung pada kualitas data pelatihan.	1. Batasan validitas: Efektif untuk keterampilan dasar namun kurang andal untuk analisis tematis tingkat lanjut. 2. Pertukaran (*trade-off*) antara skalabilitas dan adaptasi instruksional lokal.

Meskipun tingkat kesepakatan antara manusia dan AI secara konsisten tinggi di berbagai mesin penilaian, kinerja sistem sering kali dipengaruhi oleh karakteristik data pelatihan dan fitur linguistik tertentu. Demi menjaga validitas, sistem ini mengutamakan aspek keterjelasan (*explainability*) dan desain yang melibatkan manusia (*human-in-the-loop*): baik E-rater® maupun penilai manusia sama-sama memberikan skor untuk setiap esai. Untuk memberikan umpan balik yang dapat ditindaklanjuti, sistem ini (serta perangkat Criterion® yang terkait) menganalisis berbagai dimensi seperti tata bahasa dan gaya penulisan, sembari memberikan panduan langkah demi langkah selama proses pembelajaran berlangsung.

Dalam konteks pendidikan di Tiongkok, *Elion Intelligent Composition Tutoring System*—yang dikembangkan oleh East China Normal University dan Microsoft Research Asia—mengilustrasikan bagaimana multimodalitas dan granularitas penilaian dapat diprioritaskan dalam penilaian menulis yang berorientasi pada kelas. Untuk memproses tugas tulisan tangan mahasiswa, sistem ini menggunakan model OCR yang dilatih secara khusus, sementara metode berbasis BERT digunakan untuk menganalisis fitur linguistik dalam tulisan mahasiswa. Kekuatan sistem ini terletak pada granularitas penilaiannya: sistem ini mengadopsi kerangka penilaian empat lapis (penerapan bahasa, ekspresi, deteksi anomali, dan kualitas wacana) untuk menyajikan laporan diagnostik yang terperinci, bukan sekadar nilai tunggal.

Implikasi dan Arah Masa Depan

Ditinjau dari perspektif sistem, penilaian otomatis telah menjadi bagian integral dari operasional dan tata kelola penilaian berskala besar saat ini, bukan sekadar teknologi penilaian yang berdiri sendiri. Sistem penilaian otomatis telah berkembang dari metode berbasis aturan menjadi kerangka kerja multimodal berbasis AI yang mendukung penilaian berskala besar serta umpan balik instruksional. Kemajuan terkini semakin menyeimbangkan aspek akurasi, skalabilitas, kemampuan penjelasan (*explainability*), dan granularitas umpan balik; hal ini

menempatkan penilaian otomatis tidak hanya sebagai alat efisiensi, tetapi juga sebagai potensi pendukung penilaian formatif jika diselaraskan dengan tujuan pedagogis.

Meskipun demikian, tantangan tetap ada dalam hal penilaian penalaran tingkat tinggi, penjaminan keadilan, dan adaptasi terhadap konteks lokal, yang menegaskan perlunya pengawasan manusia yang berkelanjutan, desain model yang transparan, dan evaluasi yang melampaui sekadar akurasi rata-rata. Kemajuan di masa depan akan bergantung pada kolaborasi erat antara peneliti, pendidik, dan pembuat kebijakan untuk mengembangkan sistem yang tangguh secara teknis sekaligus bermakna secara pedagogis.

Analitik Pembelajaran dan Sistem Diagnostik

Landasan Konseptual dan Kerangka Kerja Teknologi

Analitik pembelajaran dan sistem diagnostik merupakan komponen sentral dalam penilaian pendidikan berbasis AI, yang memberikan wawasan berkelanjutan mengenai perilaku, keterlibatan, dan kinerja pembelajar. Analitik pembelajaran berfokus pada pengumpulan dan analisis data pembelajar secara sistematis untuk memahami serta mengoptimalkan proses dan lingkungan pembelajaran, sementara sistem diagnostik mengidentifikasi miskonsepsi, kesenjangan pembelajaran, dan kekurangan keterampilan guna menghasilkan umpan balik yang dapat ditindaklanjuti bagi pendidik maupun mahasiswa. Kemajuan dalam teknologi AI telah meningkatkan kemampuan sistem-sistem ini dalam menganalisis data yang kompleks, bersifat longitudinal, dan heterogen, sehingga memungkinkan evaluasi yang lebih akurat dan peka terhadap konteks.

Dengan melacak lintasan pembelajaran, mengidentifikasi pola keterlibatan dan risiko, serta mengintegrasikan berbagai sumber bukti mengenai pembelajar, sistem-sistem ini mendukung intervensi yang tepat waktu dan dipersonalisasi. Secara keseluruhan, analitik pembelajaran dan sistem diagnostik mengubah penilaian dari sekadar pengukuran hasil yang bersifat retrospektif menjadi proses yang berkelanjutan, prediktif, dan berpusat pada pembelajar yang mendukung pengembangan.

Aplikasi Global dan Implementasi Platform

Di seluruh dunia, berbagai negara dan institusi sedang menjajaki bagaimana AI dapat meningkatkan sistem analisis pembelajaran untuk memungkinkan penilaian yang berkelanjutan dan berorientasi pada pembelajaran. Alat-alat berkemampuan AI ini menyediakan kerangka kerja adaptif, memberikan umpan balik yang berkelanjutan, dan mendukung perencanaan pembelajaran berdasarkan data.

Platform seperti *Open Learning Analytics* memungkinkan integrasi data lintas platform, memungkinkan pendidik melacak kemajuan peserta didik secara *real-time*. Di Amerika Serikat, universitas seperti Purdue dan Arizona State telah menerapkan platform seperti Course Signals dan Civitas Learning untuk meningkatkan retensi dan kinerja mahasiswa. Di Australia, Universitas Terbuka Australia menggunakan sistem *Personalized Adaptive Study Success* (Keberhasilan Studi Adaptif yang Dipersonalisasi) untuk memberikan perencanaan kurikulum individual bagi mahasiswa yang mengalami kesulitan. Universitas Terbuka di Inggris telah menunjukkan bagaimana strategi pedagogi mempengaruhi keterlibatan jangka panjang melalui studi perbandingan terhadap lebih dari 40 desain pembelajaran. Upaya global ini

menunjukkan bagaimana AI tidak hanya dapat mendukung efisiensi pengajaran tetapi juga kesetaraan pendidikan dengan mengidentifikasi peserta didik yang mungkin terabaikan.

Studi Kasus Nasional

Contoh menarik dari inisiatif analisis pembelajaran nasional dapat ditemukan di Estonia, sebuah negara yang dikenal luas sebagai pemimpin global dalam reformasi pendidikan digital. Didukung oleh dukungan pemerintah yang kuat dan komitmen terhadap kesetaraan, Estonia telah memanfaatkan komitmennya terhadap pembelajaran yang adil dan inovasi digital untuk membangun salah satu sistem pendidikan paling maju di dunia. Inti dari upaya ini adalah platform analisis pembelajaran berskala nasional yang mengintegrasikan data dari sekolah dan sistem manajemen pembelajaran.

Platform ini mengumpulkan informasi tentang kehadiran, penilaian, dan indikator perilaku untuk menghasilkan wawasan real-time di tingkat nasional, sekolah, dan ruang kelas. Analisis yang didukung AI ini memfasilitasi intervensi individu yang tepat waktu dan pembuatan kebijakan sistemik berbasis bukti, yang memberikan contoh bagaimana analisis terintegrasi dapat meningkatkan pembelajaran dan tata kelola yang dipersonalisasi. Model Estonia memberikan contoh bagaimana infrastruktur analitik terintegrasi dapat meningkatkan pembelajaran yang dipersonalisasi dan pembuatan kebijakan berbasis bukti.

Implikasi dan Arah Masa Depan

Singkatnya, analisis pembelajaran dan sistem diagnostik mentransformasikan penilaian pendidikan melalui umpan balik yang dipersonalisasi dan real-time. Mereka memajukan penilaian formatif sambil mendukung kesetaraan dan personalisasi pendidikan. Seperti yang terlihat di negara-negara seperti Estonia, integrasi AI dalam diagnostik menjadi kenyataan global. Namun penerapannya yang efektif memerlukan perhatian terhadap etika, privasi data, dan kesenjangan infrastruktur. Ke depan, kolaborasi di antara para pembuat kebijakan, pendidik, dan ahli teknologi sangat penting untuk menciptakan sistem yang terukur, transparan, dan responsif secara budaya yang melayani beragam pelajar di seluruh dunia.

Pemodelan Prediktif untuk Intervensi Dini

Data, Pendekatan Pemodelan, dan Dasar Pemikiran untuk Prediksi Risiko Dini

Untuk mendukung intervensi yang tepat waktu dan meningkatkan hasil akademik, pemodelan prediktif telah menjadi pendekatan utama dalam mengidentifikasi mahasiswa yang menghadapi risiko akademik di lingkungan pembelajaran berbasis digital. Sistem pendidikan modern menghasilkan data berskala besar yang mencakup informasi profil mahasiswa (misalnya, karakteristik demografis dan kinerja akademik sebelumnya) serta data keterlibatan perilaku yang mencerminkan interaksi peserta didik dengan platform pendidikan.

Dengan menganalisis pola dalam data tersebut, model prediktif berbasis AI dapat mendeteksi sinyal awal penurunan keterlibatan atau kesulitan belajar serta memberikan informasi untuk strategi dukungan yang terarah dan dipersonalisasi. Berbagai pendekatan *Machine Learning* (ML) dan *Deep Learning* (DL), yang dilengkapi dengan *Natural Language Processing* (NLP) untuk data tekstual tidak terstruktur, telah diterapkan dalam prediksi risiko dini. Efektivitas model-model ini sangat bergantung pada karakteristik, ketersediaan, dan

granularitas data yang mendasarinya, sehingga menyoroti pentingnya menyelaraskan pilihan pemodelan dengan konteks pendidikan tertentu.

Penerapan Global Analisis Prediktif dan Sistem Intervensi Dini

Analisis prediktif telah diterapkan secara luas dalam dunia pendidikan di berbagai belahan dunia, yang berdampak pada peningkatan retensi dan kinerja mahasiswa. Di Eropa, universitas-universitas di Jerman telah menunjukkan keberhasilan; Berens dkk. mencapai tingkat akurasi prediksi sebesar 79%–85% pada semester pertama dan 90%–95% pada semester keempat dengan menggunakan metode ML. Di Amerika Utara, Matz menggabungkan data sosio-demografis dan metrik keterlibatan aplikasi untuk memprediksi tingkat putus sekolah (*attrition*) secara lebih akurat. Di Asia, Chung dan Lee mengembangkan sistem peringatan dini putus sekolah bagi Mahasiswa sekolah menengah di Korea. Di Afrika, Masabo dkk. berhasil menerapkan teknik ML untuk deteksi dini Mahasiswa berisiko dalam sistem pendidikan tinggi di Rwanda. Keberhasilan juga tercatat di Amerika Selatan, tepatnya di Uruguay, di mana Queiroga meluncurkan inisiatif analitik pembelajaran tingkat nasional untuk jenjang K–12.

Studi Kasus Nasional

Mengacu pada tren global penggunaan analitik prediktif untuk mendukung mahasiswa yang berisiko putus sekolah, India muncul sebagai contoh nyata penerapan skala besar sistem prediksi putus sekolah berbasis AI. Tingkat putus sekolah Mahasiswa masih menjadi masalah serius di negara-negara berkembang, di mana India menghadapi tantangan signifikan di berbagai jenjang pendidikan. Rata-rata tingkat putus sekolah tahunan untuk sekolah menengah di India mencapai lebih dari 12% pada tahun 2022, sementara pendidikan dasar menunjukkan pola putus sekolah yang lebih rendah namun tetap mengkhawatirkan.

Negara bagian Andhra Pradesh muncul sebagai contoh perintis. Pada bulan Desember 2015, Pemerintah Andhra Pradesh bermitra dengan Microsoft untuk mengembangkan sistem prediksi putus sekolah berbasis AI menggunakan platform *Azure Machine Learning*. Dengan memanfaatkan kemampuan *machine learning* (ML) dan analitik, aplikasi Microsoft tersebut mampu memprediksi kemungkinan seorang Mahasiswa putus sekolah berdasarkan pola yang diidentifikasi dari kumpulan data yang kompleks. Pada tahun 2018–2019, platform ini mengidentifikasi 19.500 mahasiswa yang berpotensi putus sekolah hanya di distrik Vishakhapatnam, dengan faktor-faktor utama yang mencakup kurangnya perabotan dan infrastruktur toilet yang tidak memadai.

Implikasi dan Arah Masa Depan

Pemodelan prediktif telah menjadi komponen kunci intervensi dini dalam pendidikan kontemporer, yang memungkinkan identifikasi mahasiswa berisiko serta penyediaan dukungan yang tepat waktu, personal, dan berskala luas. Dengan mengintegrasikan profil mahasiswa, data perilaku yang rinci, dan teknik ML tingkat lanjut, sistem-sistem ini menunjukkan potensi besar untuk meningkatkan retensi, keberhasilan akademik, dan kesetaraan pendidikan. Namun, tantangan terkait kualitas data, kemampuan interpretasi model, dan potensi bias masih ada; akurasi prediksi saja tidaklah cukup tanpa kapasitas kelembagaan yang memadai, keterlibatan guru, dan tata kelola yang etis. Oleh karena itu, upaya di masa depan harus

memprioritaskan sistem prediktif yang transparan, peka terhadap konteks, dan berlandaskan etika, yang dikembangkan melalui kerja sama erat antara pendidik, pembuat kebijakan, dan ahli teknologi.

Teknologi Penilaian Adaptif dan Personal

Landasan Konseptual dan Teknologi Utama untuk Penilaian Adaptif dan Personal

Penilaian adaptif dan personal berbasis AI merepresentasikan pergeseran dari tes statis menuju evaluasi yang lebih fleksibel dan berpusat pada pembelajar, dengan tujuan meningkatkan presisi pengukuran sekaligus efisiensi penilaian. Perkembangan ini didasarkan pada kemajuan dalam teori psikometrika, khususnya teori respons butir (*item response theory* atau IRT), yang memiliki sifat invariansi parameter yang memungkinkan pemilihan butir secara dinamis lintas bentuk tes dan titik waktu.

Dengan mengacu pada IRT, pengujian adaptif berbasis komputer melakukan personalisasi pemilihan butir berdasarkan respons peserta tes, sehingga meningkatkan akurasi, mempersingkat durasi tes, dan meningkatkan keterlibatan peserta, serta terbukti bermanfaat dalam pendidikan inklusif dan pendidikan luar biasa. Pengembangan lebih lanjut, seperti model IRT yang mempertimbangkan faktor waktu (*time-sensitive IRT models*), semakin menyempurnakan estimasi kemampuan dengan menyertakan informasi mengenai waktu respons. Selain estimasi kemampuan satu dimensi, IRT multidimensi dan model klasifikasi diagnostik mendukung penilaian keterampilan kompleks yang lebih rinci serta memberikan umpan balik diagnostik mengenai penguasaan materi dan miskonsepsi. *Multistage adaptive testing* (atau *multistage tailored testing*, MST) menyeimbangkan kemampuan adaptasi dengan kontrol serta keamanan tes, dan telah diterapkan dalam penilaian berskala besar seperti *Duolingo English Test*.

Dalam cakupan yang lebih luas, psikometrika komputasional (*computational psychometrics*) mengintegrasikan pemodelan psikometrik dengan AI dan sains data untuk menganalisis data proses (misalnya, penekanan tombol dan pola respons) serta mendukung fungsi-fungsi seperti pengawasan jarak jauh (*remote proctoring*). Melengkapi pendekatan ini, analitik pembelajaran dan pembuatan butir soal otomatis (*automated item generation*) semakin memperluas cakupan penilaian adaptif dengan memungkinkan prediksi perilaku, dukungan pembelajaran yang dipersonalisasi, dan pengembangan bank soal yang dapat ditingkatkan skalanya.

Tinjauan Alat AI untuk Pengalaman Pembelajaran Adaptif dan Terpersonalisasi

Kemajuan dalam analitik pembelajaran, NLP, dan penilaian otomatis telah memungkinkan penerapan teknik AI ke dalam alat praktis yang mendukung pembelajaran adaptif dan terpersonalisasi. Di antaranya, sistem tutor cerdas (*intelligent tutoring systems* atau ITS) merupakan salah satu pendekatan yang paling mapan; sistem ini dirancang untuk meniru fungsi utama tutor dengan menganalisis data pembelajar guna memberikan instruksi dan umpan balik yang disesuaikan secara individual. Umumnya disusun berdasarkan model domain, model mahasiswa, dan model tutor, ITS menyesuaikan pembelajaran baik pada tingkat pemilihan tugas maupun pada tahapan langkah demi langkah. Kemajuan terkini dalam AI percakapan dan NLP telah semakin meningkatkan responsivitas sistem ini, disertai bukti

yang terus berkembang bahwa ITS juga dapat mendukung pengembangan keterampilan metakognitif seperti penjelasan mandiri (*self-explanation*).

Selain sistem tutor, platform pembelajaran adaptif memperluas personalisasi ke konteks instruksional yang lebih luas dengan menggunakan *machine learning* (ML) untuk menganalisis profil, perilaku, dan data kinerja pembelajar guna merekomendasikan konten, jalur pembelajaran, dan umpan balik yang disesuaikan. Sistem penilaian dan umpan balik otomatis berperan sebagai pelengkap dengan memungkinkan penilaian formatif yang tepat waktu dan berbasis data, serta mendukung pengambilan keputusan instruksional selama proses pembelajaran berlangsung.

Studi Kasus Nasional

Salah satu contoh yang dapat menggambarkan hal ini adalah *Student Learning Space* (SLS) di Singapura, sebuah platform pembelajaran daring nasional berbasis AI yang dikembangkan oleh Kementerian Pendidikan untuk mendorong pengalaman pembelajaran yang terpersonalisasi dan inklusif. Didukung oleh perangkat AI, platform SLS meningkatkan keterlibatan mahasiswa terhadap materi pembelajaran, menjamin inklusivitas dan aksesibilitas, memperbaiki analisis penilaian, serta meningkatkan produktivitas. Platform ini telah diadopsi secara luas oleh berbagai sekolah di Singapura dan secara signifikan mengubah praktik pembelajaran tradisional di kelas.

Salah satu komponen utamanya adalah sistem pembelajaran adaptif (*adaptive learning system* atau ALS), yang menyesuaikan jalur pembelajaran dan memberikan umpan balik instan berdasarkan tingkat kesiapan mahasiswa. Sejak tahun 2023, kementerian telah memperkenalkan lima perangkat berbasis AI, termasuk tiga "asisten umpan balik pembelajaran" untuk mata pelajaran matematika, bahasa Inggris, dan respons jawaban singkat; ALS; serta "*Authoring Copilot*" yang membantu guru dalam merancang pembelajaran (Kementerian Pendidikan Singapura, 2025). Secara kolektif, perangkat-perangkat ini memberikan umpan balik dan koreksi yang dipersonalisasi bagi mahasiswa, sekaligus memungkinkan dosen untuk merencanakan pembelajaran secara lebih efektif dan efisien.

Contoh pelengkap lainnya dapat ditemukan dalam penerapan tes adaptif di Australia melalui *National Assessment Program – Literacy and Numeracy* (NAPLAN), sebuah asesmen nasional tahunan bagi mahasiswa. Untuk meningkatkan akurasi, NAPLAN beralih dari tes berbasis kertas ke desain tes MST (*Multistage Adaptive Testing*) daring. Sistem ini menyesuaikan tingkat kesulitan soal berdasarkan kinerja masing-masing mahasiswa. Skor akhir mahasiswa ditentukan oleh jumlah serta tingkat kesulitan soal yang dijawab dengan benar.

Pendekatan adaptif ini meningkatkan presisi pengukuran tanpa memperpanjang durasi tes, sekaligus membantu menjaga keterlibatan Mahasiswa. Beberapa tantangan utama yang dihadapi meliputi penyelarasan hasil tes adaptif baru dengan data historis serta pemastian aksesibilitas bagi Mahasiswa penyandang disabilitas. Untuk mengatasi masalah tersebut, NAPLAN memperbarui skala pengukuran dan standar kemampuannya pada tahun 2023. Ringkasan perbandingan kedua kasus tersebut disajikan dalam Tabel 11.2.

Tabel 11.2 Analisis Perbandingan Studi Kasus Penilaian Adaptif

Dimensi	Studi Kasus 1: Student Learning Space (Singapura)	Studi Kasus 2: NAPLAN (Australia)
Konteks dan tujuan	Platform pembelajaran personalisasi berskala nasional: Digunakan untuk kegiatan belajar-mengajar sehari-hari dan penilaian formatif; mengutamakan keterlibatan mahasiswa serta dukungan bagi dosen.	Asesmen nasional berisiko tinggi (high-stakes): Digunakan untuk tes literasi dan numerasi tahunan; mengutamakan presisi dan akurasi pengukuran dalam skala luas.
Teknologi utama	Memanfaatkan sistem pembelajaran adaptif (ALS) dan "asisten umpan balik" berbasis AI (untuk mata pelajaran matematika, bahasa Inggris, dan jawaban singkat) guna menyesuaikan alur pembelajaran.	Menggunakan metode *multistage tailored testing* (MST) untuk menyesuaikan tingkat kesulitan soal berdasarkan kinerja mahasiswa.
Kekuatan dan fokus psikometris	Umpan balik dan personalisasi: 1. Memberikan petunjuk instan dan terarah kepada mahasiswa (bersifat formatif). 2. Mendukung dosen melalui fitur AI "Copilot" dalam penyusunan materi pelajaran. 3. Meningkatkan inklusivitas dan aksesibilitas.	Presisi dan keterlibatan: 1. Meningkatkan presisi pengukuran tanpa menambah durasi tes. 2. Menjaga keterlibatan mahasiswa dengan menyesuaikan tingkat kesulitan soal terhadap kemampuan mahasiswa. 3. Penilaian memperhitungkan tingkat kesulitan butir soal.
Tantangan yang ditangani	Produktivitas dan efisiensi: Meningkatkan produktivitas dosen dalam perencanaan pembelajaran serta menyediakan umpan balik yang dapat diterapkan dalam skala luas di berbagai mata pelajaran.	Keberlanjutan dan standar: Skala pengukuran diperbarui pada tahun 2023 untuk mengatasi tantangan dalam menghubungkan data historis dan memastikan aksesibilitas.

Implikasi dan Arah Masa Depan

Singkatnya, teknologi penilaian yang adaptif dan personal sedang mengubah cara pengukuran pembelajaran melalui evaluasi yang dinamis dan responsif. Kemajuan dalam psikometrika, analitik AI, dan implementasi berskala nasional menggeser fokus penilaian dari sekadar tes statis menjadi dukungan instruksional yang berkelanjutan. Dengan memadukan tes adaptif, umpan balik otomatis, dan analitik pembelajaran, sistem ini meningkatkan presisi, keterlibatan, dan relevansi instruksional. Tantangan utamanya mencakup upaya menjamin keadilan, transparansi, dan aksesibilitas bagi beragam peserta didik, serta menyeimbangkan personalisasi dengan standardisasi dalam situasi penilaian yang berdampak tinggi (seperti ujian penentu kelulusan). Upaya di masa depan harus memprioritaskan model yang dapat diinterpretasikan, tata kelola yang etis, dan kolaborasi antara guru dan AI guna memastikan sistem ini bermakna secara pedagogis dan bertanggung jawab secara sosial.

Penilaian Imersif dan Berbasis Pengalaman Menggunakan VR/AR

Teknologi VR dan AR sebagai Landasan Penilaian Pendidikan yang Imersif

Untuk mentransformasi pendidikan melalui pengalaman belajar yang imersif, VR dan AR hadir sebagai teknologi baru yang melampaui metode tradisional. VR menciptakan lingkungan digital sepenuhnya yang membawa pengguna masuk ke dalamnya melalui *headset* atau perangkat khusus, menggantikan dunia fisik dengan dunia virtual. Sebaliknya, AR memperkaya dunia nyata dengan menampilkan lapisan informasi digital di atas pandangan pengguna menggunakan kacamata, lensa, atau aplikasi seluler. Baik VR maupun AR memungkinkan pembelajaran interaktif dengan materi yang kompleks, sehingga menciptakan pengalaman yang lebih kaya.

Teknologi ini bergantung pada perangkat keras canggih seperti *head-mounted display* (HMD), sensor gerak, dan platform komputasi spasial untuk menciptakan simulasi yang menarik. Selain itu, VR dan AR juga meningkatkan penilaian pendidikan dengan memperbaiki akurasi, keterlibatan, dan interaktivitas; hal ini memungkinkan pendidik merancang penilaian personal dan dinamis yang sesuai dengan gaya belajar individu sesuatu yang sulit dicapai dengan metode tradisional. Dalam konteks yang lebih luas, integrasi VR/AR sebagai sarana pendidikan memiliki potensi besar, terutama dengan tersedianya AI generatif fungsional saat ini yang semakin meningkatkan kemampuan teknologi tersebut.

Keunggulan Pengukuran dalam Penilaian Imersif Berbasis VR/AR

Penilaian berbasis VR/AR telah muncul sebagai instrumen transformatif untuk mengevaluasi kompetensi praktis di berbagai bidang multidisiplin, serta menawarkan keunggulan signifikan dibandingkan metode penilaian tradisional. Teknologi imersif ini memberikan keuntungan khas bagi aplikasi penilaian. VR meningkatkan presisi pengukuran melalui perekaman akurat metrik perilaku, seperti waktu respons dan pola gerakan. Di saat yang sama, AR memperkaya konteks penilaian tradisional dengan menumpangkan informasi digital ke lingkungan dunia nyata, sehingga memungkinkan evaluasi keterampilan yang lebih naturalistik sembari mempertahankan keaslian lingkungan.

Kedua teknologi ini terbukti meningkatkan keterlibatan peserta dan mengurangi kecemasan saat tes melalui pengalaman imersif dan interaktif yang meminimalkan batasan artifisial yang sering ditemukan dalam penilaian berbasis kertas atau laboratorium. Lebih lanjut, kemampuan VR untuk mereplikasi skenario terstandarisasi dengan fidelitas tinggi serta kemampuan AR untuk menempatkan penilaian dalam konteks situasi nyata secara bersama-sama meningkatkan validitas ekologis; hal ini pada gilirannya meningkatkan generalisasi dan validitas prediktif dari hasil penilaian untuk berbagai penerapan di dunia nyata.

Studi Kasus Penerapan Sistem Penilaian VR/AR

Seiring dengan semakin matangnya teknologi VR dan AR, penerapannya dalam bidang pengajaran, pelatihan keterampilan, dan penilaian pun semakin meluas. Sebagai contoh, Wang dkk. mengembangkan kerangka kerja penilaian berbasis VR untuk keterampilan medis menggunakan simulasi prosedur apendektomi. Dengan berpedoman pada model *Substitution-Augmentation-Modification-Redefinition* (SAMR), sistem ini mengintegrasikan evaluasi

kompetensi, tugas penilaian, data kinerja, dan umpan balik di dalam ruang operasi virtual yang imersif.

Mahasiswa melakukan prosedur bedah menggunakan *Head-Mounted Display* (HMD), sementara sistem secara otomatis merekam indikator perilaku seperti durasi operasi, akurasi spasial, dan penggunaan instruksi bantuan untuk memberikan umpan balik formatif maupun sumatif. Hasil penelitian menunjukkan tingkat fidelitas penilaian yang tinggi, cakupan hasil pembelajaran yang komprehensif, serta penerimaan yang baik dari mahasiswa terhadap evaluasi berbasis VR.

Demikian pula, Dinc mengusulkan kerangka kerja penilaian berbasis AR untuk pendidikan kimia melalui laboratorium virtual ARChem. Platform ini memadukan eksperimen AR dengan penilaian berbasis kuis yang terintegrasi untuk mengevaluasi pemahaman teoretis dan keterampilan praktis mahasiswa. Fitur penilaian otomatis, umpan balik seketika, dan mode permainan (*gamification*) memungkinkan penilaian tingkat kemahiran yang bertingkat sekaligus meningkatkan keterlibatan mahasiswa. Pendekatan berbasis AR ini mendukung penilaian yang dinamis, interaktif, dan dipersonalisasi di berbagai tingkat kompetensi peserta didik.

Implikasi dan Arah Masa Depan

Secara keseluruhan, penilaian imersif berbasis VR dan AR menandakan pergeseran menuju pengukuran pendidikan yang lebih autentik dan berorientasi pada keterampilan dengan memungkinkan tugas-tugas yang kaya konteks, perekaman perilaku yang mendetail, serta keselarasan yang lebih erat antara tujuan pembelajaran dan evaluasi. Bukti empiris menunjukkan bahwa lingkungan ini dapat meningkatkan validitas ekologis dan keterlibatan peserta didik, khususnya untuk menilai kompetensi dunia nyata yang kompleks.

Meskipun demikian, adopsi berskala besar masih terkendala oleh kebutuhan infrastruktur, biaya, aksesibilitas, dan perlunya validasi yang kuat di berbagai konteks yang beragam. Kemajuan di masa depan akan bergantung pada integrasi teknologi imersif dengan analitik berbasis AI dan umpan balik adaptif, serta pengembangan kerangka kerja validitas yang ketat dan strategi implementasi yang etis dan inklusif untuk mendukung penggunaannya secara berkelanjutan dalam sistem pendidikan arus utama.

11.3 AI MULTIMODAL DALAM EVALUASI

Integrasi AI Multimodal untuk Evaluasi Holistik

Evaluasi kelas tradisional terutama mengandalkan pengamatan dosen terhadap perilaku, emosi, dan penyelesaian tugas mahasiswa, yang membatasi penilaian pada hasil yang dapat diamati dan perspektif evaluatif tunggal. Pendekatan semacam itu sering kali mengabaikan proses pembelajaran, memunculkan bias subjektif, dan memiliki koherensi yang lemah antar-kriteria evaluasi, sehingga membatasi validitas dan reliabilitasnya. Sebagai tanggapan, kemajuan dalam AI dan teknologi pendidikan telah mendorong pergeseran menuju lingkungan kelas cerdas dan penilaian berbasis AI yang bersifat multimodal. Dengan mengintegrasikan data multimodal dan kerangka kerja pedagogis, pendekatan ini mengatasi sifat unidimensi dari evaluasi tradisional, memungkinkan pemahaman yang lebih holistik dan

berorientasi pada proses mengenai dinamika pembelajaran, serta mendukung pengambilan keputusan instruksional yang lebih tepat sasaran.

Analitik Pembelajaran Multimodal: Kerangka Kerja dan Sumber Data

Analitik pembelajaran multimodal meningkatkan pemahaman mengenai proses pembelajaran dengan mengintegrasikan aliran data heterogen yang sulit ditafsirkan jika dilihat secara terpisah. Pada intinya, pendekatan ini bergantung pada fusi data multimodal yang didefinisikan sebagai integrasi komputasional dari berbagai input sensorik, seperti sinyal linguistik, visual, perilaku, dan fisiologis untuk membangun representasi dengan fidelitas tinggi mengenai kondisi kognitif, afektif, dan perilaku peserta didik. Dengan menangkap interaksi antar-dimensi tersebut, kerangka kerja fusi multimodal menghasilkan validitas prediktif yang lebih kuat dibandingkan analisis unimodal.

Dalam evaluasi pembelajaran berbasis kelas, data multimodal biasanya mencakup empat kategori: (1) data perilaku, seperti gestur dan gerakan tubuh yang diekstraksi melalui teknik visi komputer (*computer vision*) seperti pengenalan titik kunci kerangka tubuh (*skeletal keypoint recognition*) menggunakan OpenPose; (2) data afektif yang diperoleh dari ekspresi wajah dan fitur bicara, termasuk kecepatan bicara serta intonasi; (3) data kognitif dan fisiologis, seperti indikator pelacakan mata (eye-tracking) untuk distribusi perhatian, variabilitas detak jantung, dan aktivitas elektrodermal; serta (4) data interaksi, termasuk akurasi respons dan pola partisipasi dalam kegiatan kelas. Secara keseluruhan, sumber-sumber data multimodal ini bila diintegrasikan melalui teknik fusi yang tepat memungkinkan evaluasi pembelajaran yang lebih holistik dan berorientasi pada proses dibandingkan pendekatan tradisional yang berbasis pada hasil akhir.

Tantangan, Pertimbangan Etis, dan Arah Masa Depan

Meskipun penilaian multimodal menawarkan potensi yang signifikan, penerapannya dihadapkan pada beberapa tantangan krusial. Privasi data tetap menjadi perhatian utama, mengingat pengumpulan informasi sensitive seperti data wajah dan fisiologis harus mematuhi norma-norma etika. Selain itu, aspek keterpahaman (*interpretability*) algoritma menjadi hal yang esensial, karena logika pengambilan keputusan dari model fusi yang kompleks memerlukan transparansi guna memfasilitasi pemahaman guru terhadap hasil evaluasi.

Pada akhirnya, pencapaian kemampuan adaptasi teknis memerlukan keseimbangan antara kelayakan penerapan teknologi dan efektivitas tugas multimodal dalam penilaian pembelajaran. Upaya di masa mendatang harus berfokus pada optimalisasi model yang ringan guna mendorong adopsi luas penilaian multimodal di berbagai skenario pendidikan, serta benar-benar mewujudkan siklus tertutup "penilaian sebagai pembelajaran" (*assessment as learning*).

Integrasi Blockchain untuk Manajemen Data yang Aman

Landasan Pemikiran dan Dasar Teknologi

Seiring dengan percepatan transformasi digital dalam penilaian pendidikan, kebutuhan akan pengelolaan data penilaian dan kredensial yang aman, tahan terhadap manipulasi, serta dapat diverifikasi menjadi sangat krusial. Sistem data tradisional sering kali kurang memadai dalam mengatasi masalah seperti pemalsuan kredensial, interoperabilitas data, dan privasi

mahasiswa. Untuk mengatasi masalah-masalah tersebut, teknologi blockchain muncul sebagai solusi yang menjanjikan, dengan menawarkan sistem buku besar (*ledger*) yang terdesentralisasi, transparan, dan tidak dapat diubah (*immutable*), yang mampu meningkatkan integritas serta keandalan pengelolaan data pendidikan.

Salah satu kekuatan utama blockchain terletak pada arsitektur terdesentralisasinya, yang menghilangkan kebutuhan akan otoritas pusat untuk mengelola dan memverifikasi data. Struktur ini mendistribusikan salinan identik dari buku besar ke seluruh jaringan *node*, sehingga memastikan tidak ada satu pihak pun yang dapat mengubah informasi secara sepihak tanpa adanya konsensus. Desentralisasi semacam ini secara signifikan mengurangi risiko manipulasi data, akses tidak sah, atau kegagalan pada satu titik pusat (*single point of failure*) yang sering menjadi kelemahan sistem terpusat.

Selain itu, blockchain menggunakan fungsi *hash* kriptografis untuk mengamankan blok data. Setiap blok memuat *hash* dari blok sebelumnya, sehingga membentuk rantai informasi yang tidak dapat diubah dan dapat dilacak. Perubahan kecil sekalipun pada data akan mengubah nilai *hash*, yang secara langsung mengindikasikan adanya upaya manipulasi. *Smart contract* (kontrak pintar) yaitu kode yang dapat berjalan secara otomatis dan tersimpan di dalam blockchain memberikan lapisan keamanan dan otomatisasi tambahan. Kontrak ini dapat menegakkan kontrol akses, mengotomatisasi penerbitan kredensial, serta memastikan bahwa transaksi data mengikuti aturan yang telah ditetapkan sebelumnya, tanpa intervensi atau bias manusia. Secara keseluruhan, berbagai fitur teknis ini menjadikan *blockchain* sangat tepat untuk menyimpan dan memverifikasi catatan penilaian, pencapaian akademik, serta kredensial digital.

Aplikasi Utama dan Proposisi Nilai

Blockchain menawarkan landasan yang kokoh untuk meninjau kembali cara data penilaian pendidikan diverifikasi, dibagikan, dan dilindungi. Salah satu kekuatan utamanya terletak pada kemampuannya memfasilitasi verifikasi kredensial dan transkrip secara otomatis dan tepercaya, yang sangat bermanfaat bagi mahasiswa internasional dan pengakuan akademik lintas batas. Berdasarkan desainnya, *blockchain* menjamin kekekalan data (*immutability*), sehingga sangat efektif dalam mencegah pemalsuan ijazah dan memperkuat kepercayaan antar-institusi pendidikan.

Keandalan ini juga mendukung pelacakan jangka panjang terhadap lintasan pembelajaran individu, serta menyediakan catatan historis aman yang dapat menjadi acuan bagi perencanaan pendidikan yang dipersonalisasi. Pada saat yang sama, dukungan blockchain terhadap kerangka kerja identitas yang dikendalikan secara mandiri (*self-sovereign identity*) memungkinkan pembelajar untuk membagikan data mereka lintas platform sembari tetap menjaga privasi dan kepemilikan data.

Bagi pendidik dan pembuat kebijakan, *blockchain* menawarkan keyakinan lebih besar terhadap keaslian dan portabilitas hasil penilaian, hal yang sangat relevan dalam konteks sistem pendidikan yang terglobalisasi dan lingkungan pembelajaran daring. Secara keseluruhan, berbagai kemampuan ini menjadikan blockchain sebagai faktor pendukung yang

menjanjikan bagi ekosistem penilaian yang transparan, aman, dan berpusat pada pembelajar di era digital.

Implikasi dan Arah Masa Depan

Singkatnya, integrasi teknologi *blockchain* dalam penilaian pendidikan menandai langkah signifikan menuju pengelolaan data yang aman, transparan, dan memiliki interoperabilitas global. Meskipun masih terdapat tantangan terkait skalabilitas teknis, adopsi institusional, dan penyelarasan regulasi, peningkatan jumlah program percontohan dan inisiatif yang didukung pemerintah menunjukkan arah perkembangan yang menjanjikan. Seiring dengan semakin pentingnya aspek kepercayaan, privasi, dan kemampuan verifikasi dalam pendidikan digital, blockchain hadir sebagai faktor kunci yang memungkinkan inovasi yang aman dan berpusat pada pembelajar.

Personalisasi dan Penilaian Pembelajaran Sepanjang Hayat

Mengacu pada aplikasi yang dibahas, tren yang muncul mengarah pada personalisasi yang lebih mendalam dalam pembelajaran sepanjang hayat. AI memainkan peran sentral dengan menganalisis data dalam jumlah besar untuk menyesuaikan konten dan kecepatan belajar demi pengembangan keterampilan yang berkelanjutan. Salah satu tren utama dalam penilaian pembelajaran sepanjang hayat adalah meningkatnya penggunaan AI untuk mendukung jalur pembelajaran yang sangat personal, di mana penyampaian materi, kecepatan belajar, dan penilaian disesuaikan secara dinamis dengan profil masing-masing pembelajar. Untuk menghadirkan pengalaman pendidikan yang semakin canggih, sistem berbasis AI di masa depan akan menganalisis profil pembelajar guna merekomendasikan konten yang sangat relevan. Selain pembelajaran adaptif, AI akan menyediakan umpan balik waktu nyata (*real-time*), penilaian otomatis, dan panduan karier yang dipersonalisasi. Alat AI generatif, seperti *chatbot*, akan menawarkan panduan yang disesuaikan berdasarkan kemajuan individu.

Personalisasi akan semakin ditingkatkan melalui teknologi baru yang sedang berkembang, termasuk VR dan AR, komputasi afektif, kurasi konten berbasis AI, serta "agen pembelajaran" cerdas. Secara keseluruhan, perkembangan ini menandakan pergeseran dalam metode penilaian dari evaluasi berbasis tahapan yang terikat pada sistem sekolah formal menuju mekanisme berkelanjutan dan personal yang mendampingi pembelajar di sepanjang rangkaian pembelajaran seumur hidup.

Kredensial, Penyelarasan Tenaga Kerja, dan Tata Kelola Data

Seiring meluasnya personalisasi berbasis AI dalam pembelajaran sepanjang hayat, sistem penilaian semakin dibentuk oleh tata kelola data, pengakuan kredensial, dan penyelarasan tenaga kerja, melampaui sekadar desain instruksional. Persyaratan utama bagi penilaian pembelajaran sepanjang hayat yang dapat ditingkatkan skalanya adalah tata kelola data yang berpusat pada pembelajar; hal ini menekankan kendali individu, transparansi, dan portabilitas data yang aman di berbagai konteks pendidikan dan pekerjaan. Inisiatif MyData di Finlandia menjadi contoh pendekatan ini dengan memungkinkan kendali data oleh individu serta aliran data yang lancar untuk mendukung layanan yang dipersonalisasi.

Secara bersamaan, praktik penilaian beralih menuju pengakuan kompetensi yang lebih rinci melalui *micro-credential* dan lencana digital (*digital badges*), yang memungkinkan pendokumentasian keterampilan secara berkelanjutan dan dapat diverifikasi di sepanjang perjalanan pembelajaran dan karier. Mekanisme semacam itu mendukung profil keterampilan dinamis yang dapat diperbarui dan diakui lintas institusi serta pasar tenaga kerja, sebagaimana ditunjukkan oleh inisiatif seperti lencana terbuka digital (*digital open badges*) dan program Paspor Keterampilan AI. Dalam konteks ini, penilaian berfungsi bukan sekadar sebagai mekanisme instruksional, melainkan sebagai antarmuka penghubung antara sistem pendidikan, pasar tenaga kerja, dan kerangka kerja tata kelola melalui data, kredensial, dan representasi kompetensi yang digunakan bersama.

Di luar pendidikan formal, penilaian berbasis AI semakin mendukung peningkatan keterampilan (*upskilling*) dan pelatihan ulang (*reskilling*) tenaga kerja dengan menyelaraskan kompetensi individu terhadap tuntutan pasar tenaga kerja yang terus berkembang. Pemerintah dan organisasi menerapkan inisiatif pelacakan keterampilan dan pelatihan berbasis AI untuk mengatasi kekurangan tenaga kerja serta membina talenta yang siap menghadapi masa depan; strategi nasional seperti Strategi AI Nasional UEA 2031 menunjukkan integrasi sistemik antara penilaian, pemberian kredensial, dan pengembangan tenaga kerja. Seiring semakin luasnya penggunaan AI dalam pendidikan dan penilaian, tata kelola yang etis menjadi hal yang krusial, dengan penekanan yang terus meningkat pada transparansi, akuntabilitas, dan upaya memastikan bahwa AI memperkuat bukan menggantikan penilaian manusia. Pengembangan literasi AI, yang memadukan kompetensi teknis dengan kesadaran etis dan tanggung jawab sosial, telah ditetapkan sebagai prioritas oleh *UNESCO* dan *World Economic Forum*.

Pada tingkat regulasi, kebijakan seperti *EU AI Act* serta pedoman nasional semakin memprioritaskan perlindungan data, keselamatan, dan pencegahan penyalahgunaan. Menyikapi kekhawatiran mengenai integritas akademik terkait AI generatif, metode penilaian kini beralih ke format terbuka dan berorientasi pada proses yang mengutamakan pemikiran kritis serta penalaran.

Tantangan dan Arah Masa Depan

Meskipun personalisasi berbasis AI dalam penilaian pembelajaran sepanjang hayat memiliki potensi besar, penerapannya secara luas menghadirkan masalah etika dan praktis yang signifikan. Tantangan utama mencakup bias algoritmik, keadilan, dan transparansi dalam pengambilan keputusan otomatis, terutama ketika hasilnya memengaruhi jalur pendidikan atau karier. Pengumpulan data dalam skala besar semakin meningkatkan risiko privasi dan akuntabilitas, sehingga menuntut adanya kerangka tata kelola yang kuat serta peran aktif dan otonomi pembelajar yang bermakna.

Peningkatan literasi AI juga sangat penting untuk menafsirkan umpan balik secara kritis dan mempertahankan pengawasan manusia. Upaya di masa depan harus memprioritaskan penilaian autentik yang berorientasi pada proses dan mendorong pemikiran kritis, sembari membina kolaborasi antar pemangku kepentingan untuk mengembangkan sistem yang transparan, inklusif, dan tepercaya. Menangani berbagai kekhawatiran ini merupakan kunci

dalam membangun kerangka penilaian yang adil dan berpusat pada manusia, yang secara efektif mendukung pembelajaran sepanjang hayat dalam skala luas.

Standar Lintas Negara dan Tantangan Global

Untuk mencerminkan semakin pentingnya peran AI dalam pendidikan, organisasi internasional seperti *United Nations Educational, Scientific and Cultural Organization* (UNESCO) dan *Organisation for Economic Co-operation and Development* (OECD) telah mulai mengintegrasikan AI ke dalam agenda penilaian pendidikan mereka. Bagi OECD, PISA 2029 akan memperkenalkan bidang penilaian baru, yaitu "Literasi Media dan Kecerdasan Buatan" (*Media and Artificial Intelligence Literacy*).

Tes ini akan menyimulasikan lingkungan tempat Mahasiswa berinteraksi dengan AI melalui platform seperti internet, media sosial, dan alat AI generatif, guna menilai literasi digital dan literasi AI mereka (OECD, n.d.). Dikembangkan bersama oleh Uni Eropa dan OECD, dengan dukungan dari Code.org dan para ahli, kerangka kerja ini menegaskan pentingnya AI yang terus meningkat dalam evaluasi pendidikan berskala besar. Akibatnya, standar terkait terus berkembang seiring waktu untuk mengimbangi kemajuan teknologi dan perubahan kebutuhan pendidikan.

Selain itu, *International Association for the Evaluation of Educational Achievement* (IEA), sebuah organisasi penilaian internasional terkemuka, telah merilis standar teknis baru untuk tahun 2025. Ini merupakan revisi komprehensif sejak tahun 1999 yang menyoroti perubahan signifikan dalam bidang penilaian internasional. Pembaruan utama mencakup penilaian berbasis komputer dengan elemen-elemen seperti pengumpulan data proses, integrasi penilaian (*scoring*) berbasis AI, dan desain adaptif. Secara khusus, standar rinci telah ditetapkan untuk penilaian otomatis berbasis AI. Hanya dengan menetapkan standar operasional yang menyeluruh bagi program tersebut, keandalan dan efisiensi tinggi AI dapat terjamin, sehingga memungkinkannya untuk lebih mendukung proses penilaian manual.

Lanskap Global dan Langkah Menuju Penyelarasan Lintas Negara

Meskipun belum ada standar global untuk penilaian pendidikan berbasis AI yang setara dengan PISA yang dipimpin oleh satu organisasi tunggal—berbagai organisasi internasional dan lembaga standardisasi saat ini aktif menjajaki serta merumuskan pedoman yang relevan. Dapat diamati bahwa mulai dari panduan kebijakan dan kerja sama multilateral lintas negara hingga pedoman teknis, semuanya sedang meletakkan landasan bagi penilaian pendidikan berbasis AI. Di masa mendatang, seiring dengan terus bertambahnya pengalaman dari proyek percontohan dan contoh kasus berkualitas tinggi serta semakin kuatnya konsensus internasional, kita diharapkan akan melihat standar penilaian lintas negara dan rencana praktis yang lebih sistematis serta bersifat universal.

Risiko Kognitif Penggunaan AI dalam Penilaian dan Respons Berorientasi Kapabilitas

Dampak AI terhadap penilaian pendidikan pada dasarnya beragam, sehingga menuntut kewaspadaan terhadap risiko kognitifnya, khususnya konten yang keliru dan bias algoritmik. Model Bahasa Besar (LLM) rentan mengalami halusinasi menghasilkan keluaran yang menyimpang dari pengetahuan yang sudah mapan atau mengarang fakta yang dapat menyesatkan pengguna jika tidak ada mekanisme evaluasi yang efektif. Selain itu, sistem AI

dapat mereproduksi atau memperkuat bias layaknya manusia terkait ras dan gender, yang berpotensi memperparah ketimpangan. Bukti menunjukkan bahwa interaksi dengan algoritma yang bias bahkan dapat memperkuat bias manusia dalam persepsi, emosi, dan penilaian sosial—efek yang tidak ditemukan dalam interaksi antarmanusia. Terlebih lagi, temuan terkini mengindikasikan bahwa LLM dapat menunjukkan "preferensi diri" dengan lebih mengutamakan deskripsi produk buatan AI, sehingga mencerminkan bentuk bias yang merugikan manusia. Jika tidak diatur, kekurangan kognitif ini dapat menimbulkan dampak buruk yang lebih luas secara budaya dan etika.

Ringkasan dan Arah Masa Depan

Secara keseluruhan, AI mulai diintegrasikan secara bertahap ke dalam penilaian pendidikan internasional melalui inisiatif berskala besar, standar teknis yang terus berkembang, dan panduan kebijakan yang baru muncul. Meskipun organisasi seperti OECD dan IEA secara aktif mengembangkan kerangka kerja penilaian berbasis AI, praktik global masih belum merata dan terfragmentasi, yang mencerminkan perbedaan dalam lingkungan regulasi, kapasitas teknis, dan prioritas pendidikan. Pada saat yang sama, kekhawatiran terkait bias algoritmik, halusinasi, dan pengaruh kognitif menegaskan pentingnya memandang AI bukan sekadar sebagai alat teknis, melainkan juga sebagai aktor sosio-edukasional. Ke depannya, upaya berkelanjutan menuju koordinasi lintas negara, standar operasional yang lebih jelas, dan pendidikan AI yang berorientasi pada kapabilitas akan menjadi hal krusial untuk mendukung integrasi AI yang bertanggung jawab dan efektif ke dalam sistem penilaian global.

11.4 PENILAIAN PENDIDIKAN BERBASIS AI

Sebagaimana dibahas dalam bab ini, penilaian pendidikan berbasis AI sedang mengubah cara pembelajaran dievaluasi dan didukung sebagai respons terhadap tantangan global yang mendesak, seperti kesetaraan, pengembangan keterampilan, dan komparabilitas lintas negara. Dalam lingkungan pendidikan yang semakin mengglobal, kemampuan untuk menilai hasil pembelajaran di berbagai konteks yang beragam bukan lagi sekadar pilihan, melainkan prasyarat bagi perumusan kebijakan yang tepat dan praktik yang efektif. Berdasarkan kasus dan analisis yang disajikan, kolaborasi, kustomisasi, dan kontroversi muncul sebagai elemen kunci dalam memahami arah transformasi ini di tingkat internasional.

Kolaborasi menjadi landasan bagi penggunaan penilaian berbasis AI yang dapat ditingkatkan skalanya, sebagaimana dicontohkan oleh MOOC internasional dan platform lintas batas yang mengintegrasikan kumpulan data pembelajar yang besar dan beragam untuk keperluan tolok ukur, pembelajaran, dan pengembangan kapasitas. Kustomisasi mencerminkan kemampuan AI dalam mendukung penilaian yang berpusat pada pembelajar dan peka terhadap konteks, sehingga memungkinkan penyelarasan dengan kurikulum lokal, bahasa, dan profil individu pembelajar melalui sistem adaptif dan diagnostik.

Pada saat yang sama, kontroversi tetap menjadi bagian tak terpisahkan dari transformasi ini, karena isu-isu yang belum terselesaikan terkait privasi data, keamanan, tata kelola etis, dan ketimpangan infrastruktur terus menghambat penerapan berskala besar dan interoperabel. Meskipun teknologi seperti *blockchain* menawarkan dukungan menjanjikan

untuk kredensial yang aman, standar global dan kerangka kerja etis bersama masih belum merata dan lambat terbentuk.

Menatap masa depan, konvergensi antara AI multimodal, kredensial yang diamankan dengan *blockchain*, dan protokol evaluasi yang diakui secara internasional menawarkan jalan menuju interoperabilitas, transparansi, dan kepercayaan yang lebih baik. Mewujudkan potensi ini memerlukan kolaborasi berkelanjutan, kepekaan terhadap konteks lokal, serta pendekatan yang seimbang antara inovasi dan etika, guna memastikan bahwa penilaian berbasis AI mampu mendorong keunggulan sekaligus kesetaraan di seluruh dunia. Seiring dengan semakin terintegrasinya AI ke dalam penilaian pendidikan, berikut adalah rekomendasi praktis bagi para pemangku kepentingan utama:

Bagi Pendidik dan Pelatih Pengajar:

- **Mengembangkan Literasi Kritis terhadap Penilaian AI**: Mengevaluasi secara kritis desain, keterbatasan, dan potensi bias dari perangkat penilaian berbasis AI, serta mengintegrasikan prinsip keadilan algoritmik ke dalam pedagogi digital.
- **Mengedepankan Model *Human-in-the-Loop* (Manusia dalam Sistem)**: Memanfaatkan hasil diagnostik dan umpan balik yang dihasilkan AI sebagai bahan pertimbangan—bukan pengganti bagi penilaian profesional dan dukungan yang dipersonalisasi bagi mahasiswa.

Bagi Administrator Pendidikan dan Pimpinan Institusi:

- **Menerapkan Uji Coba Bertahap dan Audit Etis:** Mengadopsi penerapan AI secara bertahap yang disertai dengan analisis dampak sistematis, audit bias, serta kebijakan institusional yang jelas mengenai tata kelola data dan akuntabilitas. Berinvestasi dalam Infrastruktur dan Peningkatan Kapasitas: Mendukung integrasi AI yang adil melalui investasi berkelanjutan pada infrastruktur digital dan pengembangan staf yang berfokus pada penggunaan data secara etis.

Bagi Ahli Psikometri dan Pakar Asesmen:

- **Memelopori Kerangka Kerja Validitas untuk AI:** Mengembangkan dan memvalidasi standar psikometris untuk asesmen berbasis AI yang mencakup aspek keadilan, kemampuan untuk dijelaskan (*explainability*), dan generalisasi lintas konteks.
- **Mendorong Sains Terbuka dan Tolok Ukur (*Benchmarking*):** Mempromosikan tolok ukur terbuka yang menjaga privasi guna memungkinkan evaluasi sistem asesmen AI yang transparan dan lintas budaya.

Bagi Pembuat Kebijakan dan Organisasi Internasional:

- **Membangun Tata Kelola yang Adaptif dan Berbasis Prinsip:** Memajukan pedoman internasional yang fleksibel untuk asesmen AI yang etis, yang menjamin kedaulatan data, keadilan, dan pengakuan lintas batas.
- **Mendorong Ekosistem Inovasi yang Adil:** Mendanai infrastruktur asesmen AI yang terbuka, modular, dan berbiaya rendah untuk mendukung sistem dengan sumber daya terbatas serta mengurangi kesenjangan asesmen global.

Melalui tindakan terkoordinasi dan spesifik sesuai peran yang berlandaskan komitmen etis serta semangat kolaboratif, komunitas pendidikan global dapat mengarahkan pengembangan

asesmen berbasis AI. Tujuan utamanya adalah memanfaatkan kekuatan transformatif ini bukan sekadar untuk efisiensi pengukuran, melainkan untuk mewujudkan masa depan pembelajaran yang lebih autentik, adil, dan berpusat pada manusia di seluruh dunia.

Daftar Singkatan

AI	kecerdasan buatan (*artificial intelligence*)
ALS	sistem pembelajaran adaptif (*adaptive learning system*)
AR	realitas tertambah (*augmented reality*)
BERT	*bidirectional encoder representations from transformers*
DL	pembelajaran mendalam (*deep learning*)
EICTS	*Elion Intelligent Composition Tutoring System*
HMD	perangkat tampilan yang dikenakan di kepala (*head-mounted displays*)
IEA	*International Association for the Evaluation of Educational Achievement*
IRT	teori respons butir soal (*item response theory*)
ITS	sistem tutor cerdas (*intelligent tutoring systems*)
LLM	model bahasa berskala besar (*large language models*)
ML	pembelajaran mesin (*machine learning*)
MOOC	kursus daring terbuka masif (*massive open online courses*)
MST	*multistage tailored*
NAPLAN	*National Assessment Program – Literacy and Numeracy*
NLP	pemrosesan bahasa alami (*natural language processing*)
OCR	pengenalan karakter optis (*optical character recognition*)
OECD	*Organisation for Economic Co-operation and Development*
PISA	*Programme for International Student Assessment*
SLS	*Student Learning Space*
VR	realitas virtual (*virtual reality*)

BAB 12
MENILAI DAN MENDUKUNG KOLABORASI AI

12.1 INTEGRASI AI UNTUK MENILAI DAN MENDUKUNG KINERJA KOLABORATIF

Kolaborasi merupakan keterampilan mendasar, baik dalam dunia pendidikan maupun dunia kerja yang terus berkembang. Pentingnya kolaborasi bersumber dari perannya dalam mendorong pemikiran tingkat tinggi, meningkatkan hasil kerja, serta mempersiapkan individu menghadapi lingkungan kerja yang kompleks dan saling terhubung. Kolaborasi didefinisikan sebagai "aktivitas sinkron yang terkoordinasi, yang merupakan hasil dari upaya berkelanjutan untuk membangun dan mempertahankan pemahaman bersama mengenai suatu masalah". Pada dasarnya, kolaborasi lebih dari sekadar bekerja bersama; kolaborasi melibatkan penciptaan "ruang masalah bersama" (*joint problem space*) untuk mencapai solusi bersama.

Meningkatnya otomatisasi, tim lintas fungsi, dan pekerjaan berbasis proyek dalam ekonomi global telah memperbesar kebutuhan akan kemampuan kolaborasi yang kuat. Pemberi kerja semakin sering melaporkan bahwa "keterampilan lunak" (*soft skills*) seperti kolaborasi, pemecahan masalah, dan kemampuan beradaptasi sama pentingnya dengan keahlian teknis demi kemajuan karier, ketangguhan karier, dan. Menjelang tahun 2030, diperkirakan hingga 39% keterampilan penting tenaga kerja akan mengalami pergeseran, dengan kolaborasi, pemikiran kreatif, dan pengaruh sosial tetap menjadi prioritas utama.

Dalam dunia pendidikan, pembelajaran kolaboratif menghadapkan Mahasiswa pada beragam perspektif, mengajarkan mereka untuk memecahkan masalah secara kolektif, serta memupuk keterampilan sosial. Analisis meta terhadap berbagai studi mengenai pemecahan masalah secara kolaboratif (*collaborative problem-solving* atau CPS) di kelas menunjukkan bahwa metode ini membantu mahasiswa mengembangkan keterampilan berpikir kritis, komunikasi, dan kepemimpinan, sekaligus meningkatkan kepercayaan diri dan keterlibatan mereka. Dengan demikian, kolaborasi tidak hanya menjadi landasan bagi pembelajaran yang efektif; Hal ini juga mempersiapkan mahasiswa dan pekerja untuk masa depan, membekali mereka dalam menghadapi tuntutan dunia kerja yang berubah dengan cepat serta memastikan kemampuan mereka untuk berkembang di lingkungan yang beragam dan kaya akan teknologi.

Namun, untuk memberikan penekanan lebih besar pada kolaborasi, diperlukan definisi yang disepakati mengenai apa saja yang mencakup keterampilan kolaborasi, serta pendekatan yang kokoh untuk menilai, memberikan umpan balik, dan mengajarkan keterampilan tersebut. Institusi pendidikan dan berbagai negara telah merespons kebutuhan akan standar, pelatihan, dan penilaian yang lebih baik ini dengan mengintegrasikan keterampilan tersebut ke dalam kurikulum, mengembangkan kerangka kerja penilaian baru, dan menyesuaikan kebijakan pendidikan. Kerangka kerja dan kebijakan ini telah memengaruhi praktik pendidikan serta mendorong penerapan pendekatan penilaian terstandarisasi, seperti penilaian CPS dalam *Programme for International Student Assessment* (PISA) 2015 dari OECD. Meskipun terdapat kesepakatan umum mengenai pentingnya keterampilan kolaborasi, penilaian dan pemberian

umpan balik terkait kolaborasi masih merupakan bidang yang sedang berkembang. Kurangnya perkembangan di bidang ini disebabkan oleh beberapa faktor.

- **Kolaborasi merupakan konstruk yang kompleks dan memiliki banyak dimensi**. Kolaborasi memadukan komponen kognitif (pemecahan masalah, penalaran), sosial (bergantian giliran bicara, memengaruhi, penyelesaian konflik), dan emosional (empati, membangun kepercayaan). Berbagai disiplin ilmu mengoperasionalkan definisi dan konstruk kolaborasi dengan cara yang berbeda, sehingga sulit untuk menyelaraskan alat ukurnya. Norma spesifik bidang yang dianggap memadai dalam satu konteks pembelajaran atau profesional mungkin tidak dapat diterapkan pada konteks lainnya.
- **Perilaku kolaborasi sangat bergantung pada konteks.** Dalam kelompok perencanaan mahaMahasiswa tingkat sarjana yang bekerja secara daring dan asinkron, anggota mungkin lebih fokus untuk menjaga keterlibatan dan menafsirkan maksud satu sama lain. Sebaliknya, kelompok Mahasiswa sekolah menengah (tingkat SMP) di bidang STEM yang bekerja secara tatap muka mungkin perlu lebih fokus membangun pemahaman bersama mengenai ruang lingkup masalah. Tuntutan spesifik tugas juga dapat memunculkan pola kolaborasi yang berbeda. Aktivitas jangka pendek mungkin berfokus pada pencapaian solusi dengan cepat, sementara aktivitas jangka panjang mungkin lebih menekankan pada pembangunan hubungan dan koneksi antaranggota kelompok.
- **Kolaborasi bersifat dinamis dan berkembang seiring waktu.** Oleh karena itu, penilaian sesaat (*snapshot*) atau penilaian *post hoc* (setelah kegiatan selesai) mungkin melewatkan pola-pola penting. Pola-pola ini juga dapat menunjukkan kemajuan tim dengan cara yang tidak linier. Tim mungkin mengalami fase konflik, keheningan, atau perbedaan pendapat yang justru diperlukan untuk mencapai keberhasilan pada akhirnya. Namun, fase-fase tersebut bisa terlihat sebagai kinerja yang buruk jika diamati secara terpisah.
- **Kolaborasi pada dasarnya bersifat multimodal dan melibatkan banyak pihak.** Bukti adanya kolaborasi dapat muncul melalui aktivitas berbicara, menulis, ekspresi wajah, gestur, atau input ke dalam ruang kerja digital. Pola-pola halus dalam fitur multimodal dapat mengindikasikan faktor-faktor penting seperti bagaimana seorang individu memengaruhi kelompok yang pada gilirannya dapat memengaruhi hasil solusi kelompok tersebut.

Secara keseluruhan, faktor-faktor ini menunjukkan bahwa untuk menilai kolaborasi secara efektif, diperlukan pergeseran dari pengukuran satu dimensi yang sederhana menuju model dinamis yang memperhitungkan nuansa interaksi antarpribadi, komunikasi verbal dan nonverbal, serta norma tim yang terus berkembang. Otomatisasi pengukuran menggunakan metode berbasis AI menawarkan potensi untuk mengatasi kompleksitas tersebut. Dengan memadukan kerangka kerja kolaborasi yang dirancang dengan baik dan teknik penilaian berbasis AI yang tangguh, kita dapat memperoleh gambaran yang mendalam mengenai kinerja kolaboratif serta mengembangkan pendekatan inovatif untuk mendukung mahasiswa dan pengajar. Selain penerapan kerangka kerja yang tepat dan teknologi AI yang akurat, terdapat

berbagai faktor krusial lain yang berperan dalam perancangan aplikasi pendidikan untuk kolaborasi yang realistis dan berkelanjutan. Penerapan teknologi penilaian semacam itu dalam situasi nyata dapat menyulitkan proses pengumpulan data, yang secara langsung memengaruhi jenis tugas dan interaksi yang dapat digunakan. Hal ini juga menuntut penggunaan model penilaian kompleks yang mampu menghasilkan inferensi yang andal dan valid mengenai kinerja individu maupun kelompok.

Oleh karena itu, kami mendukung integrasi AI yang seimbang dan berpusat pada manusia dalam mengembangkan pendekatan penilaian kolaboratif. Bab ini menyajikan taksonomi faktor-faktor yang perlu dipertimbangkan agar penerapan solusi berbasis AI untuk mendukung kolaborasi dapat berhasil. Selanjutnya, bab ini memberikan contoh penerapan di dunia nyata serta menjelaskan bagaimana penerapan tersebut selaras dengan faktor-faktor yang telah disebutkan.

12.2 MENGARAKTERISASI KETERAMPILAN KOLABORASI

Untuk menilai kolaborasi, pertama-tama kita perlu mengidentifikasi keterampilan yang menjadi sasaran. Kerangka kerja keterampilan menyediakan kriteria yang jelas, objektif, dan terstandarisasi untuk mengevaluasi kompetensi individu maupun tim, serta memastikan bahwa penilaian keterampilan bersifat adil, andal, dan selaras dengan tujuan pengalaman pendidikan atau lingkungan kerja. Beberapa kerangka kerja telah dikembangkan untuk mengukur CPS (*Collaborative Problem Solving* atau Pemecahan Masalah secara Kolaboratif), yang masing-masing menekankan aspek dan konteks yang berbeda. Kerangka kerja ini bertujuan untuk menangkap keterampilan kognitif dan sosial yang penting bagi individu dan tim agar dapat bekerja sama dalam memecahkan masalah yang kompleks secara efektif. Namun, beberapa kerangka kerja terutama berfokus pada penilaian kelompok dalam dunia pendidikan, sementara yang lain berfokus pada kinerja tim di dunia kerja. Di sini, kami akan membahas beberapa kerangka kerja utama.

Kerangka Kerja Kolaborasi

Proyek *Assessment and Teaching of 21st Century Skills* (ATC21S) mengembangkan konsep untuk menilai dan mengajarkan keterampilan CPS, dengan fokus pada keterampilan yang dapat diajarkan. Kerangka kerja teoretisnya mencakup:

- **Partisipasi:** Mencakup tingkat keterlibatan secara umum, interaksi dengan orang lain, dan responsivitas terhadap orang lain, serta aspek motivasi yang diperlukan untuk penyelesaian tugas.
- **Pengambilan perspektif (*perspective taking*):** Melibatkan pemahaman dan respons terhadap orang lain, menempatkan ucapan dalam konteks, mengintegrasikan kontribusi, dan menyesuaikan komunikasi seseorang (penyesuaian dengan *audiens*).
- **Regulasi sosial:** Membahas bagaimana kelompok bekerja secara produktif dengan adanya keragaman anggota. Keterlibatan ini mencakup metamemori (pengetahuan tentang diri sendiri), memori transaktif (pengetahuan tentang kekuatan dan kelemahan rekan kolaborasi), keterampilan negosiasi (untuk menyelesaikan konflik), dan keterampilan inisiatif (tanggung jawab terhadap kemajuan kelompok).

Kerangka kerja ATC21S berfokus pada interaksi antarmanusia dalam situasi pemecahan masalah dan interaksi yang autentik dan tidak terlalu kaku (tidak mengikuti naskah yang ketat). Contohnya adalah Mahasiswa yang bekerja dalam latihan kelompok kecil di kelas. Meskipun pendekatan ini memberikan validitas ekologis yang lebih tinggi, penilaian kinerja kelompok tidak mudah untuk diotomatisasi karena bergantung pada deskriptor keterampilan yang sulit dikuantifikasi (misalnya, "Mahasiswa menggunakan komunikasi yang sangat terorganisasi dan berbasis bukti untuk menjaga tim tetap fokus pada tujuan").

Kondisi ini berarti kerangka kerja tersebut memerlukan proses anotasi keterampilan mahasiswa oleh manusia yang memakan banyak waktu dan tenaga. Kerangka kerja PISA dari OECD dirancang untuk menilai keterampilan CPS (pemecahan masalah kolaboratif) dengan mensurvei mahasiswa berusia 15 tahun di 52 negara dan wilayah ekonomi. Kerangka ini mendefinisikan kompetensi CPS sebagai kapasitas individu untuk terlibat secara efektif dalam suatu proses di mana dua atau lebih agen berupaya memecahkan masalah dengan berbagi pemahaman dan upaya yang diperlukan untuk mencapai solusi, serta menggabungkan pengetahuan, keterampilan, dan usaha mereka untuk mencapai solusi tersebut. Kerangka kerja ini disusun berdasarkan dua dimensi utama. Dimensi Kognitif (Pekerjaan Tugas) berfokus pada kompetensi pemecahan masalah individu yang diadaptasi dari PISA 2012, yang mencakup:

- **Eksplorasi dan pemahaman:** Menafsirkan informasi awal maupun informasi baru yang ditemukan terkait suatu masalah.
- **Representasi dan perumusan:** Mengidentifikasi pendekatan, strategi, dan artefak yang relevan untuk memecahkan masalah.
- **Perencanaan dan pelaksanaan:** Menyusun dan menjalankan rencana serta tindakan (baik yang bersifat fisik, sosial, maupun verbal).
- **Pemantauan dan refleksi:** Memantau kemajuan menuju tujuan serta merefleksikan kualitas kemajuan atau solusi yang dihasilkan.

Dimensi Kolaboratif (Kerja Tim) berfokus pada proses sosial yang terlibat dalam kolaborasi:

- **Membangun dan memelihara pemahaman bersama:** Memantau pengetahuan dan perspektif anggota tim serta visi bersama mengenai kondisi masalah dan aktivitas yang dilakukan.
- **Mengambil tindakan yang tepat untuk memecahkan masalah:** Melakukan tindakan fisik dan komunikatif yang mendorong tercapainya solusi.
- **Membangun dan memelihara organisasi kelompok:** Membantu mengorganisasi atau menata ulang kelompok berdasarkan pengetahuan, keterampilan, kemampuan, dan sumber daya anggota, termasuk pembagian peran dan penanganan hambatan.

Kedua dimensi ini saling beririsan membentuk matriks 12 keterampilan untuk penilaian CPS. Penilaian PISA dirancang agar dapat dinilai secara otomatis, di mana mahasiswa berinteraksi dengan dua atau lebih agen berbasis komputer dalam situasi yang dibatasi; dalam situasi ini, mahasiswa dapat memilih respons wacana yang telah ditetapkan sebelumnya dalam percakapan yang mengikuti skrip. Keterbatasan dalam pendekatan penilaian ini disebabkan oleh perlunya menyajikan penilaian yang setara di berbagai negara, memastikan kinerja

mahasiswa tidak bergantung pada mahasiswa lain serta memastikan semua mahasiswa menghadapi situasi yang sama guna menilai seluruh cakupan keterampilan yang ada dalam kerangka kerja tersebut.

Kerangka kerja CPS yang diperkenalkan oleh Sun Dan rekan-rekan memperluas kerangka kerja PISA dengan memberikan fokus lebih besar pada tugas-tugas seperti analisis bersama, negosiasi, dan konstruksi pengetahuan. Komponen kunci dari kerangka kerja ini adalah mengidentifikasi komponen wacana yang muncul selama interaksi dan berfungsi sebagai indikator keterampilan tersebut. Indikator-indikator ini berfokus pada ucapan (apa yang dikatakan) dan tindakan (apa yang dilakukan) yang spesifik dan dapat diukur dalam lingkungan digital (misalnya, menyepakati cara tim berkomunikasi atau giliran kerja sebelum memulai tugas).

Dengan berfokus pada penilaian interaksi antarmanusia melalui indikator bahasa, kerangka kerja ini mendukung percakapan bahasa yang terbuka dan alami dalam lingkungan digital yang tidak dibatasi. Strategi ini memungkinkan analisis terhadap cakupan langkah wacana dan pola interaksi yang jauh lebih luas, sehingga memungkinkan pengukuran keterampilan CPS yang lebih komprehensif dibandingkan dengan pengaturan komunikasi yang terbatas pada PISA. Kerangka kerja lainnya lebih berfokus pada kerja tim, dengan menekankan penerapannya dalam mengukur kinerja tenaga kerja. Kerangka kerja O'Neil dan rekan-rekan mengacu pada riset mengenai kerja tim untuk mengidentifikasi fitur-fitur yang diperlukan guna mengoptimalkan hasil kinerja tim. Keterampilan utamanya meliputi hal-hal berikut:

- **Adaptabilitas**: Memantau tim dan tugas, serta merespons masalah yang muncul.
- **Koordinasi**: Menyinkronkan dan mengintegrasikan aktivitas kelompok untuk memastikan penyelesaian tugas tepat waktu.
- **Keterampilan interpersonal:** Bekerja sama dan mengakomodasi anggota kelompok lainnya.
- **Kepemimpinan**: Memberikan arahan bagi kelompok.
- **Keterampilan komunikasi:** Bertukar informasi secara jelas dan akurat.

Keterampilan-keterampilan ini dipadukan dengan strategi pemecahan masalah (strategi kognitif tradisional) dan metakognisi (komponen pengaturan diri seperti perencanaan dan pemantauan). O'Neil dan rekan-rekan) lebih lanjut mengidentifikasi berbagai pendekatan untuk analisis otomatis terhadap keterampilan-keterampilan ini, termasuk penerapan jaringan Bayesian untuk mendeteksi pengaruh kausal terhadap kinerja tim. Demikian pula, Salas dan rekan-rekan. merumuskan konsep "Lima Besar" (*Big Five*) keterampilan kerja sama tim. Mereka mengidentifikasi lima komponen inti kepemimpinan tim, pemantauan kinerja bersama, perilaku saling mendukung (*backup behavior*), kemampuan beradaptasi, dan orientasi tim sebagai elemen penting dalam kerja sama tim.

Kelima keterampilan ini didukung oleh tiga mekanisme koordinasi mendasar: model mental bersama, komunikasi *closed-loop* (komunikasi dengan umpan balik terkonfirmasi), dan saling percaya. Model ini telah memberikan pengaruh signifikan terhadap perancangan penilaian dan pelatihan tenaga kerja, misalnya di bidang layanan kesehatan dan kepolisian. Meskipun terdapat beragam kerangka kerja, ada kesamaan mendasar dalam cara menilai

kolaborasi. Sebagai contoh, kolaborasi yang sukses memerlukan pelaksanaan tugas (strategi pemecahan masalah, perencanaan, pemantauan) dan kerja tim (komunikasi, koordinasi, pengelolaan hubungan).

Sebagian besar kerangka kerja mengintegrasikan aspek pemrosesan interpersonal dan metakognitif bersama dengan perilaku pemecahan masalah. Baik disebut sebagai "pemantauan dan refleksi" (PISA), "regulasi sosial" (ATC21S), maupun "pemantauan kinerja bersama" (*Big Five*), terdapat kesepakatan bahwa kesadaran akan kemajuan dan kemampuan menyesuaikan strategi merupakan keterampilan yang krusial. Semua kerangka kerja berakar pada konteks kolaboratif dunia nyata (misalnya, pendidikan, dunia kerja, atau keduanya), meskipun metode pelaksanaan penilaiannya (misalnya, dialog terstruktur versus dialog terbuka) bervariasi karena alasan praktis. Terakhir, semuanya mengoperasionalkan kolaborasi melalui indikator perilaku—seperti bergantian giliran bicara, mengambil perspektif orang lain, kemampuan beradaptasi, dan tindakan kepemimpinan—sehingga keterampilan tersebut dapat dinilai melalui observasi atau analisis otomatis. Namun, beberapa kerangka kerja lebih cocok untuk penilaian otomatis.

Sebagai contoh, kerangka kerja Sun dan rekan-rekan dapat diterapkan dengan mudah menggunakan pemrosesan bahasa alami (*natural language processing*) terhadap langkah-langkah wacana kolaboratif. Penerapan ini dimungkinkan karena adanya keterkaitan kerangka kerja tersebut dengan indikator keterampilan yang dapat diwujudkan secara komputasional. Sebaliknya, dimensi *Big Five* didefinisikan secara sangat luas, sehingga biasanya memerlukan analisis berdasarkan observasi manusia.

12.3 MENGHUBUNGKAN KERANGKA KERJA DENGAN PENGEMBANGAN PENILAIAN

Meskipun kerangka kerja memberikan gambaran mengenai keterampilan yang ada, diperlukan pendekatan yang lugas dan berlandaskan prinsip untuk memetakan keterampilan tersebut ke dalam bukti-bukti nyata. *Evidence-centered design* (ECD) atau desain berbasis bukti menyediakan kerangka kerja komprehensif untuk merancang dan melaksanakan penilaian berbasis AI yang bersifat multimodal dan melibatkan banyak pihak, dengan menguraikan apa yang diukur (model mahasiswa), di mana hal itu diukur (model tugas), dan bagaimana hal itu dihitung (model bukti).

ECD sangat relevan untuk mengembangkan metode penilaian keterampilan kompleks seperti kolaborasi, yang sering kali melibatkan data log atau data verbal dalam jumlah besar dan sangat rinci dari lingkungan digital yang perlu diubah menjadi ukuran-ukuran tertentu. Saat diterapkan pada kerangka kerja CPS (*Collaborative Problem Solving*), ECD membantu mengonseptualisasikan keterampilan CPS yang kompleks dan mengekstrak bukti dari catatan rinci mengenai tindakan serta wacana individu.

ECD membantu memperjelas hubungan antara apa yang diukur dan klaim spesifik mengenai fitur-fitur tersebut. Proses ini dimulai dengan model mahasiswa (klaim yang ingin Anda buat mengenai pengetahuan, keterampilan, atau kemampuan), kemudian bergerak mundur untuk menentukan bukti yang diperlukan serta tugas-tugas yang dapat memunculkan bukti tersebut. Dalam konteks penilaian berbasis AI, pendekatan ini mencegah masalah umum

berupa pengukuran hal-hal yang mudah ditangkap (misalnya, jumlah kata atau tingkat klik) alih-alih mengukur aspek yang sebenarnya penting bagi konstruk tersebut (misalnya, indikator verbal dalam negosiasi atau transfer pengetahuan spesifik dari satu orang ke orang lain).

Dengan demikian, ECD menjaga analisis tetap berfokus pada fitur-fitur yang relevan dengan konstruk, bukan pada korelasi semu. Mengingat data kolaboratif pada dasarnya bersifat multimodal (misalnya, ucapan, teks, gestur, dan log tindakan), penetapan konstruk sejak awal memungkinkan penentuan modalitas yang relevan dalam berbagai situasi, fitur mana yang perlu diekstraksi, serta bagaimana fitur-fitur tersebut mendukung inferensi mengenai konstruk terkait. Selain itu, ECD membantu menjadikan penilaian berbasis AI lebih transparan dan dapat dipertanggungjawabkan.

Keharusan untuk menyertakan penalaran eksplisit mengenai alasan mengapa perilaku, respons, atau pola interaksi tertentu dapat menjadi bukti bagi suatu konstruk akan mengurangi sifat "kotak hitam" (ketidakjelasan proses internal) pada sistem penilaian AI. Saat AI mendeteksi penanda perilaku individu, penanda tersebut dapat dipetakan secara eksplisit kembali ke bukti spesifik bagi konstruk yang bersangkutan, sehingga memberikan gambaran proses analisis yang lebih transparan. Transparansi ini sangat penting untuk membangun kepercayaan pemangku kepentingan (peserta didik, pendidik, dan pembuat kebijakan) serta memperjelas bagaimana AI diterapkan pada data multimodal yang kompleks. Dengan demikian, penerapan ECD dalam perancangan penilaian menghadirkan pendekatan yang terarah, berbasis bukti, dan dapat dipertanggungjawabkan untuk mengembangkan tolok ukur kinerja kolaboratif yang valid, adil, dan bermakna.

Sebagai contoh penerapan eksplisit ECD dalam kerangka kerja CPS, kerangka kerja penilaian dalam tugas atau *in-task assessment framework* (I-TAF) dirancang khusus untuk mendukung penilaian keterampilan kompleks, seperti CPS, dalam lingkungan digital yang menghasilkan data dalam jumlah besar, seperti log klik *mouse* dan pesan obrolan. Fitur utama I-TAF adalah penggunaan ontologi CPS (misalnya, model mahasiswa), yang menyediakan representasi hierarkis berbasis teori mengenai konstruk CPS dan keterampilan terkait. Ontologi ini memungkinkan kerangka kerja tersebut mendefinisikan ruang konstruk pada berbagai tingkat granularitas, mulai dari keterampilan tingkat tinggi hingga perilaku teramati tingkat rendah. Selanjutnya, kerangka kerja ini menghubungkan bukti perilaku tersebut dengan lingkungan tugas tertentu (model tugas).

Sebagai contoh, bukti perilaku dapat didasarkan pada data log tindakan mahasiswa atau pada jenis komunikasi tertentu dalam simulasi elektronik. Aturan sistematis yang menghubungkan bukti perilaku dengan keterampilan tingkat tinggi (model bukti) menyediakan sarana untuk mengekstrak bukti keterampilan CPS dari data yang sangat rinci (*fine-grained*). Pendekatan sistematis ini memungkinkan analisis otomatis terhadap interaksi dan komunikasi tim, serta memberikan penilaian mendalam yang didasari teori; pendekatan ini mampu menangani kumpulan data besar dan memberikan umpan balik tepat waktu, yang sangat penting untuk melampaui metode anotasi data kolaboratif oleh manusia yang memakan banyak waktu. Contoh penerapan ECD yang sangat baik dalam platform kolaboratif adalah platform EPCAL (*Educational Testing Service Platform for Collaborative Assessment and*

Learning). EPCAL dibangun untuk memungkinkan pengendalian dan manipulasi variabel kelompok, sehingga memungkinkan studi kolaborasi dalam skala besar sekaligus memastikan bahwa data proses yang kaya informasi selaras dengan model bukti. Dengan demikian, penggabungan kerangka kerja kolaborasi dengan pemetaan eksplisit antara bukti perilaku dan keterampilan membuka jalan bagi otomatisasi penilaian.

12.4 DIMENSI PENILAIAN KOLABORATIF

Mengidentifikasi keterampilan kolaborasi dan memetakannya ke bukti perilaku merupakan komponen penting dalam pengembangan sistem penilaian dan dukungan kolaboratif berbasis komputer. Namun, pengembangan teknologi AI untuk penilaian dan dukungan keterampilan kolaborasi tidak dapat ditinjau secara terpisah dari konteks penerapannya. Secara keseluruhan, teknologi telah menjadikan penilaian keterampilan kolaborasi lebih dapat dilaksanakan dalam situasi yang lebih luas dan lebih realistis. Oleh karena itu, sangat penting untuk meninjau secara luas berbagai peran teknologi dalam mendukung kolaborasi di segala aspeknya. Sebagai contoh, teknologi dapat:

(1) digunakan untuk menyajikan dan mengendalikan situasi tugas,
(2) menyediakan rekan tim virtual yang merespons pengguna berdasarkan analisis situasi,
(3) mengumpulkan bukti seperti ucapan, arah pandangan mata, gerakan menunjuk, hasil kerja, dan catatan tindakan,
(4) menerapkan model untuk menganalisis bukti tersebut secara otomatis,
(5) memberikan umpan balik kepada rekan tim, serta
(6) menghasilkan laporan atau dasbor bagi pengamat eksternal atau pengajar.

Lebih lanjut, kita perlu mempertimbangkan berbagai dimensi potensial dari situasi kolaboratif, termasuk jenis konstruk kolaborasi yang dinilai, lingkungan pengujian, jenis dan jumlah rekan tim (baik AI maupun manusia), jenis data yang dikumpulkan, indikator kolaborasi (verbal, arah pandangan mata, dan lain lain.), jenis model AI yang diterapkan, etika pengumpulan data, serta bagaimana dan kapan hasil penilaian digunakan untuk memberikan umpan balik mengenai kinerja peserta. Dimensi-dimensi ini menghadirkan pertukaran kepentingan (*trade-off*) terkait akurasi pengukuran (*fidelity*), relevansi dengan situasi dunia nyata, kemudahan pelaksanaan, pemanfaatan hasil, serta skalabilitas untuk penggunaan di kelas berskala besar atau penilaian yang berdampak tinggi (*high-stakes assessment*).

Tabel 12.1 merinci dimensi-dimensi utama yang perlu dipertimbangkan dalam otomatisasi penilaian dan dukungan kolaborasi, diikuti dengan ulasan mengenai dimensi-dimensi tersebut. Dimensi-dimensi ini dikelompokkan ke dalam lima ranah pertimbangan utama. Selain itu, tabel tersebut menyajikan pemetaan dimensi-dimensi ini terhadap lapisan-lapisan ECD, serta menunjukkan apakah dimensi tersebut merupakan pertimbangan krusial bagi implementasi teknis. Namun, perlu dicatat bahwa banyak aspek yang dianggap sebagai "implementasi" juga sangat berkaitan erat dengan lapisan "model tugas" (*task model*) dalam ECD.

Tabel 12.1 Dimensi Utama untuk Mempertimbangkan Implementasi Penilaian dan Dukungan Kolaboratif Otomatis

Domain Pengembangan	**Dimensi Penilaian Kolaboratif**	**Kategori (Tidak Menyeluruh)**	**Penyelarasan dengan Lapisan/Faktor Implementasi ECD**
Umpan balik dan penggunaan	Sifat umpan balik AI	Langkah kolaboratif, proses tim, perancah domain (*domain scaffolding*), laporan, insiden kritis, penilaian	Model tugas
	Sasaran umpan balik	Individu, tim, pengajar, evaluator	Model tugas
Cakupan dan skala	Skalabilitas penilaian	Uji coba, skala kecil, skala besar	Implementasi
	Sensitivitas terhadap konteks	Bersifat umum (lintas domain), spesifik domain, disesuaikan dengan konteks, sesuai secara budaya	Implementasi

Konteks dan Pengaturan

Ranah konteks dan pengaturan membahas kondisi dasar tempat penilaian kolaboratif berbasis AI dirancang dan dilaksanakan. Konteks ini dimulai dengan memperjelas tujuan penilaian: formatif (untuk memandu pembelajaran yang sedang berlangsung), sumatif (untuk mengevaluasi pembelajaran di akhir), atau diagnostik (untuk mengidentifikasi kekuatan dan kebutuhan sebelum instruksi). Setiap tujuan membawa implikasi terhadap bagaimana penilaian disusun, tingkat kepentingan hasilnya, dan jenis umpan balik yang diberikan. Sebagai contoh, pengembangan aplikasi formatif mengharuskan pemetaan keterampilan yang teridentifikasi ke dalam umpan balik yang spesifik dan efektif secara pedagogis.

Lingkungan interaksi baik simulasi, langsung (*live),* maupun hibrida membentuk realisme sekaligus tingkat kendali atas skenario kolaborasi. Misalnya, interaksi langsung dapat menangkap dinamika yang autentik namun menimbulkan variabilitas yang jauh lebih besar. Ruang kelas dapat memiliki sumber gangguan yang tidak terduga, kebisingan latar belakang tambahan, dan lebih sedikit batasan terhadap perilaku di luar tugas (*off-task*). Pengaturan simulasi menawarkan kendali yang lebih ketat namun dapat membatasi perilaku alami dan kemudahan penerapan. Demikian pula, struktur dan kendali tugas dapat bervariasi mulai dari tugas dengan skrip ketat hingga pemecahan masalah yang bersifat terbuka (*open-ended*); masing-masing memengaruhi tingkat spontanitas dan jenis kompetensi yang dapat diamati.

Pilihan pengumpulan data (misalnya, audio lisan, video, berkas log, produk akhir) bergantung pada kelayakan dan kekayaan bukti yang diinginkan. Audio lisan, video, dan berkas log menyediakan metode pengumpulan data yang kaya dan tidak mengganggu (*unobtrusive*).

Namun, metode ini memerlukan pemetaan yang presisi antara konstruk dan sumber-sumber bukti tersebut. Produk akhir dapat memberikan indikator yang efektif mengenai hasil kelompok, namun mungkin melewatkan proses yang memungkinkan kelompok tersebut mencapai hasil itu. Jenis data juga memunculkan pertimbangan privasi. Pengumpulan data video dan audio dalam situasi dunia nyata, seperti ruang kelas, dapat menuntut tingkat persetujuan yang lebih tinggi, terutama jika melibatkan anak di bawah umur.

Terakhir, perwujudan rekan satu tim baik AI berinteraksi sebagai rekan tim virtual, manusia berinteraksi satu sama lain, mode *human-in-the-loop* bertindak sebagai agen berbasis AI, atau sekadar penyediaan umpan balik secara grafis dapat memengaruhi keterlibatan, kepercayaan, dan pola interaksi peserta. Dengan demikian, perancangan rekan tim virtual berbasis AI memerlukan pertimbangan cermat mengenai bagaimana umpan balik atau respons akan dihasilkan, misalnya melalui pembuatan bahasa alami atau respons berbasis aturan yang telah ditetapkan sebelumnya.

Pemodelan dan Bukti

Ranah pemodelan dan bukti berfokus pada apa yang diukur serta bagaimana bukti diperoleh, diproses, dan ditafsirkan. Sebagaimana dijelaskan dalam bagian mengenai kerangka kerja keterampilan dan ECD, ranah ini melibatkan identifikasi kompetensi yang ditargetkan, yang dapat mencakup keterampilan CPS (pemecahan masalah kolaboratif), keterampilan sosial-emosional, tindakan pedagogis, atau pengetahuan spesifik bidang tertentu seperti STEM. Keterampilan-keterampilan ini dapat didefinisikan berdasarkan tingkat perincian analisisnya (misalnya, apakah pengukuran keterampilan dilakukan pada tingkat individu atau kelompok), karena hal ini memengaruhi kebutuhan data maupun metode analisis. Ranah ini selanjutnya mengidentifikasi sumber-sumber bukti, yang dapat mencakup data verbal dan nonverbal (misalnya, giliran bicara, arah pandangan, gerakan menunjuk), artefak hasil kerja (misalnya, produk kelompok), atau penilaian oleh pengamat. Pemilihan sumber-sumber ini bergantung pada apa yang dapat diukur dalam situasi kolaboratif tersebut serta keberhasilan pemetaannya terhadap keterampilan yang dituju.

Model analitik mencakup berbagai pendekatan komputasional dan statistik yang digunakan untuk mengekstrak bukti dari data kolaborasi. Pada tingkat paling sederhana, model ini dapat berupa penghitungan frekuensi atau statistik deskriptif dasar untuk menangkap tingkat partisipasi, giliran bicara, atau pola kontribusi. Namun, model berbasis AI telah menjadi semakin canggih dalam mengenali ucapan, menganalisis dan memberikan anotasi pada bahasa, mengukur kinerja kolaboratif dari berkas log, mengenali gestur, memproses arah pandangan dan emosi dari fitur wajah, serta menggabungkan data multimodal yang melibatkan banyak pihak. Mengulas seluruh pendekatan tersebut berada di luar cakupan bab ini. Von Davier dan rekan-rekan menyajikan pemantauan komprehensif mengenai metode komputasional untuk menilai kolaborasi di berbagai konteks dan penerapan metode psikometrik untuk validasi, sembari mengkaji penerapannya baik di dunia kerja maupun lingkungan pendidikan. Ouyang dan Zhang serta Tan dan rekan-rekan. menyajikan pemantauan mengenai analitik berbasis AI untuk pembelajaran kooperatif yang didukung

komputer. Giannakos dan rekan-rekan. dan Ochoa dan rekan-rekan. memberikan gambaran umum mengenai analitik pembelajaran multimodal (MMLA) dan pendekatan fusi data.

Komunikasi antaranggota tim merupakan salah satu sumber bukti terkaya mengenai proses dan kognisi tim. Langkah-langkah wacana yang berkaitan dengan aspek kolaborasi seperti negosiasi, pemberian pengakuan, pembangunan pengetahuan bersama, dan pembentukan pemahaman bersama dapat dipetakan ke dalam indikator keterampilan kolaborasi. Analisis terhadap fitur semantik, struktural, dan sekuensial dari wacana tersebut memungkinkan interpretasi atas makna yang disampaikan.

Pendekatan ini telah banyak digunakan bersama berbagai metode pemrosesan bahasa alami, seperti *latent semantic analysis*, pengklasifikasi *Naïve Bayes* berbasis *n-gram*, *conditional random fields* untuk urutan label, model *word embedding* seperti BERT dan Word2Vec, serta yang terbaru, penggunaan model bahasa berskala besar atau *large language models* (LLM). Pendekatan-pendekatan ini telah menunjukkan tingkat akurasi yang selaras dengan anotasi manusia terhadap kategori wacana, sehingga secara signifikan mempercepat dan mengotomatisasi karakterisasi kondisi kolaborasi melalui bahasa. Kendati demikian, wacana lisan harus terlebih dahulu diproses oleh sistem pengenalan ucapan otomatis atau *automatic speech recognition* (ASR). Meskipun ASR dapat mencapai tingkat akurasi dengan *word error rate* (WER) kurang dari 5% dalam kondisi terkendali yang senyap, kinerjanya di lingkungan dunia nyata yang bising dan saat melibatkan partisipan anak-anak menunjukkan angka WER antara 20% hingga 80%; hal ini berarti sebagian besar wacana mungkin hilang atau dihasilkan secara keliru.

Kolaborasi lebih menitikberatkan pada bagaimana kelompok mencapai hasil tertentu daripada hasil itu sendiri. Log data proses menyajikan informasi yang kaya mengenai cara individu dan kelompok melakukan tindakan, mengoordinasikan tugas, berbagi pengetahuan, serta beralih antara tahapan pemecahan masalah dan koordinasi. Fitur-fitur untuk data proses dapat mencakup:

(1) informasi temporal seperti latensi respons dan waktu jeda (*idle time*), yang mengindikasikan tingkat keterlibatan atau koordinasi;
(2) urutan (*sekuens*) seperti n-gram tindakan dan transisi antar-tahapan, yang menunjukkan pola alur kerja kolaboratif;
(3) partisipasi seperti frekuensi giliran bicara dan rasio dominansi, yang mencerminkan keseimbangan dalam kelompok serta efektivitas pergantian giliran; dan
(4) penanda berbasis wacana, yang menunjukkan transisi dari satu jenis langkah wacana ke langkah lainnya serta perubahan sentimen.

Berbagai pendekatan pemodelan telah digunakan untuk mengkarakterisasi data proses semacam itu. Sebagai contoh, He serta He dan rekan-rekan, menganalisis proses yang mendasari tindakan mahasiswa dalam penilaian CPS PISA 2015, yang memberikan pemetaan langsung terhadap keterampilan CPS yang menjadi target penilaian. Analisis jejaring sosial telah digunakan untuk merepresentasikan kolaborasi sebagai jaringan hubungan antara partisipan dan tindakan.

Pendekatan ini memungkinkan pengukuran aspek sentralitas, kepadatan, dan kohesi sebagai indikator partisipasi yang setara, pengaruh, dan koordinasi. Model Markov tersembunyi (*Hidden Markov models*) memungkinkan deteksi tahapan kolaborasi laten serta waktu terjadinya transisi antar-tahapan tersebut. Seiring dengan meningkatnya kemampuan bidang ini dalam mengumpulkan beragam aliran data seperti video, audio, dan log Tindakan muncul kebutuhan yang lebih besar bagi MMLA untuk mengintegrasikan data tersebut dan memberikan pemahaman yang lebih holistik serta bernuansa mengenai dinamika kelompok. Pendekatan MMLA memadukan informasi dari berbagai modalitas, seperti komunikasi verbal dan tertulis, ekspresi wajah, gestur, arah pandangan mata untuk atensi visual bersama, sensor fisiologis, dan berkas log tindakan, sehingga memungkinkan analisis yang lebih mendalam terhadap perilaku, kondisi afektif, dan proses kognitif selama pelaksanaan tugas kolaboratif.

Sebagai contoh, menggabungkan aspek prosodik dan transkrip ucapan dapat memberikan representasi bahasa yang lebih komprehensif, sementara integrasi fitur dari bahasa alami, video, dan audio dapat meningkatkan kinerja prediksi secara signifikan dalam konteks kolaboratif. Berbagai studi menunjukkan bahwa sistem multimodal secara konsisten lebih akurat dibandingkan sistem unimodal terbaik yang sejenis, dengan peningkatan rata-rata hampir 10%. Dengan demikian, pendekatan MMLA ini memungkinkan transisi dari analisis aliran data secara terpisah menuju model terpadu yang menghasilkan interpretasi lebih kaya, namun memerlukan alur kerja (*pipeline*) yang lebih canggih.

Waktu dan Umpan Balik

Sifat umpan balik AI bisa sangat beragam, mulai dari mengidentifikasi langkah-langkah kolaboratif dan mendiagnosis proses tim hingga memberikan perancah (*scaffolding*) khusus bidang tertentu, menyusun laporan insiden, atau menilai kinerja. Hasil-hasil ini dapat mendukung pembelajaran formatif, evaluasi sumatif, atau bahkan analitik prediktif untuk perencanaan intervensi. Namun, hal yang krusial adalah memastikan bahwa umpan balik tersebut dikaitkan kembali dengan konstruk CPS serta disajikan dengan cara yang dapat dipahami oleh pengguna akhir. AI generatif menawarkan sarana untuk menafsirkan analisis CPS dan mengubahnya menjadi umpan balik. Akan tetapi, model LLM perlu dilatih dan dinilai secara cermat untuk memastikan akurasi umpan balik tersebut. Sebagai contoh, Wang dan Demszky menemukan bahwa umpan balik bagi pengajar mengenai metode pembinaan (*coaching*) mereka meskipun relevan untuk meningkatkan kualitas pengajaran sering kali tidak menawarkan hal baru atau wawasan yang mendalam.

Sasaran umpan balik—baik itu individu, tim, maupun pendidik/pengamat menentukan nada, tingkat perincian, dan kualitas umpan balik yang dapat ditindaklanjuti. Umpan balik untuk individu mungkin menyoroti kontribusi spesifik dan area pengembangan diri, sementara umpan balik tingkat tim dapat berfokus pada koordinasi, negosiasi, dan pengambilan keputusan bersama. Jika ditujukan kepada pengajar atau evaluator, umpan balik AI dapat menjadi dasar bagi perencanaan pembelajaran, pemberian dukungan yang terarah, dan alokasi sumber daya. Perancangan yang disesuaikan dengan pengguna tujuan menuntut keseimbangan antara kemudahan interpretasi, spesifisitas, dan keselarasan dengan tujuan pembelajaran yang lebih luas.

Pertimbangan mengenai waktu sangatlah penting, baik dalam tahap pengumpulan data maupun penggunaannya dalam penilaian berbasis AI. Pengumpulan dan analisis data dapat dilakukan mulai dari metode *streaming* waktu nyata hingga analisis yang mendekati waktu nyata atau analisis *post hoc* (setelah kegiatan selesai). Namun, setiap pilihan memengaruhi kecepatan penyampaian umpan balik, skalabilitas, dan kebutuhan komputasi. Situasi dengan konsekuensi tinggi mungkin memerlukan analisis berbasis kelompok demi akurasi, sedangkan konteks yang mendukung pembelajaran mungkin lebih mengutamakan umpan balik cepat, meskipun harus mengorbankan tingkat keandalan model.

Umpan balik yang disampaikan bersamaan dengan pelaksanaan tugas kelompok dapat membentuk pola kolaborasi yang sedang berlangsung, namun harus dirancang sedemikian rupa agar tidak mengganggu interaksi alami atau menyebabkan beban kognitif berlebih. Umpan balik *post hoc* memungkinkan analisis dan refleksi yang lebih mendalam, namun kehilangan kesempatan untuk membentuk perilaku pada saat interaksi itu sendiri terjadi. Oleh karena itu, dalam konteks formatif, sangatlah penting untuk tidak hanya mengukur akurasi model dalam menggambarkan keterampilan CPS, tetapi juga menilai seberapa efektif umpan balik yang dihasilkan dari pengukuran tersebut dalam mendorong perubahan positif pada kolaborasi. Model dengan tingkat fidelitas lebih rendah yang dapat dijalankan secara *real-time* bisa jadi sama efektifnya atau bahkan lebih efektif dibandingkan model dengan fidelitas tinggi yang membutuhkan biaya komputasi lebih besar atau pemrosesan tambahan setelahnya (*post hoc*). Menyeimbangkan berbagai pertimbangan ini mengharuskan adanya keselarasan antara aspek waktu dengan tujuan penilaian, tingkat kepentingan hasil penilaian, serta kebutuhan peserta didik.

Cakupan dan Skala

Cakupan dan skala berkaitan dengan pertimbangan implementasi yang lebih luas, guna memastikan bahwa penilaian dapat berjalan efektif di lingkungan yang dituju serta menjamin validitas dan kepraktisan evaluasi tersebut. Skalabilitas penilaian mencakup perencanaan untuk tahap percontohan (*pilot*), uji coba berskala kecil, hingga penerapan berskala besar pada akhirnya; setiap tahapan ini memiliki kebutuhan infrastruktur, pelatihan, dan penjaminan mutu yang berbeda. Keberhasilan dalam peningkatan skala tidak hanya bergantung pada sistem teknis yang tangguh, tetapi juga pada dukungan organisasi, kepercayaan pemangku kepentingan, serta kemampuan beradaptasi dengan berbagai konteks yang berbeda. Pertimbangan utama dalam menerapkan sistem berbasis CPS mencakup cara pengumpulan data yang etis dan tidak mengganggu aktivitas pengguna. Pengambilan data audio dan video di ruang kelas atau tempat kerja bisa jadi rumit secara logistik, terutama jika instruktur atau peserta didik harus menyiapkan dan mengoperasikan sistem tersebut. Sebagai contoh, jika fitur wajah atau gestur digunakan, bagaimana sistem memastikan penempatan kamera yang tepat untuk menangkap citra pengguna? Jika mikrofon digunakan, apakah setiap peserta didik memerlukan mikrofon terpisah, atau cukup satu mikrofon yang ditempatkan di tengah kelompok? Seberapa besar tingkat kebisingan latar belakang yang masih dapat ditoleransi agar sistem tetap berfungsi secara efektif?

Cakupan implementasi juga harus mempertimbangkan bagaimana penilaian dan pelatihan CPS selaras dengan tujuan organisasi. Sebagai contoh, pengembangan profesional seperti apa yang akan diberikan terkait konstruk CPS dan bagaimana hal tersebut dikaitkan dengan kurikulum organisasi? Bagaimana hasil penilaian atau umpan balik dari sistem CPS akan diintegrasikan ke dalam alur kerja penilaian pengajar atau rencana pembelajaran? Selain itu, organisasi pendidikan tentu menginginkan sistem CPS yang selaras dengan kurikulum serta standar khusus mereka, yang mungkin berbeda-beda di setiap negara bagian dan distrik.

Penilaian kolaboratif berbasis AI dapat dirancang agar bersifat umum lintas domain (dapat diterapkan pada berbagai mata pelajaran), spesifik domain (disesuaikan dengan disiplin ilmu tertentu), atau disesuaikan dengan konteks (diatur menurut kebutuhan, budaya, dan kendala lokal). Saat menerapkan sistem, seberapa besar upaya yang diperlukan untuk menilai kemampuan transfer, komparabilitas hasil, dan lokalisasi di berbagai konteks yang berbeda? Lebih spesifik lagi, dalam menentukan cakupan penilaian, penting untuk menyadari bahwa kolaborasi yang efektif memiliki bentuk yang sangat berbeda di berbagai konteks budaya. Apa yang dianggap sebagai partisipasi produktif, gaya komunikasi yang tepat, atau kepemimpinan yang efektif sangat bervariasi antara budaya individualistis dan kolektivistis, lingkungan komunikasi berkonteks tinggi dan rendah, serta masyarakat dengan norma jarak kekuasaan yang berbeda.

Sebagai contoh, dalam beberapa konteks pendidikan di Asia Timur, sikap diam dan pengamatan cermat dapat menandakan keterlibatan mendalam dan rasa hormat, bukan ketidakterlibatan. Di sisi lain, perbedaan pendapat secara langsung mungkin dianggap sebagai tindakan yang menjatuhkan harga diri (mengancam *face* atau citra diri) alih-alih sebagai perdebatan yang produktif. Demikian pula, pola giliran berbicara, peran hierarki dalam pengambilan keputusan, serta ekspektasi mengenai pembentukan konsensus versus akuntabilitas individu semuanya memiliki makna budaya yang dapat disalahartikan oleh sistem AI yang dilatih terutama menggunakan data dari populasi Barat yang berpendidikan, terindustrialisasi, kaya, dan demokratis. Sistem AI yang mengutamakan frekuensi kontribusi verbal atau tantangan langsung sebagai penanda kualitas kolaborasi dapat secara sistematis merugikan mahasiswa dari budaya yang menganggap perilaku tersebut tidak pantas atau tidak sopan.

Implementasi praktis penilaian kolaborasi yang responsif secara budaya memerlukan pendekatan teknis sekaligus partisipatif. ECD dapat membantu dengan menyusun kerangka kerja penilaian yang responsif secara budaya, yaitu dengan menjabarkan secara eksplisit klaim, bukti, dan dasar pemikiran (*warrants*) yang menghubungkan perilaku yang dapat diamati dengan kompetensi kolaboratif yang mendasarinya untuk budaya tertentu. Model AI dapat dikembangkan lebih lanjut menggunakan beragam data pelatihan yang merepresentasikan berbagai konteks budaya. Model-model ini dapat menerapkan teknik seperti adaptasi domain atau *transfer learning* untuk mengakomodasi perbedaan populasi, serta menyertakan kuantifikasi ketidakpastian guna mendeteksi saat model penilaian beroperasi di luar konteks yang telah divalidasi.

Namun, solusi teknis saja tidaklah cukup. Akomodasi yang bermakna terhadap variasi sosial-budaya menuntut kolaborasi berkelanjutan dengan para pendidik, mahasiswa, dan pakar budaya dari komunitas sasaran di sepanjang proses perancangan mulai dari pemodelan kompetensi awal hingga studi validasi. Pendekatan partisipatif ini membantu memastikan bahwa kerangka kerja yang dihasilkan tidak sekadar memaksakan norma budaya yang dominan, melainkan benar-benar mengakui beragam bentuk keunggulan kolaboratif, sembari mengakui pula bahwa beberapa kompetensi kolaboratif (seperti berkoordinasi demi tujuan bersama atau mengembangkan gagasan orang lain) mungkin memiliki validitas lintas budaya, meskipun manifestasi perilakunya bisa saja berbeda-beda.

12.5 COBI UNTUK PEMBELAJARAN KOLABORATIF

Untuk mengilustrasikan dimensi-dimensi yang diuraikan dalam Tabel 12.1 melalui implementasi nyata dari sistem kolaboratif, kami memaparkan pengembangan *Community Builder* (CoBi) serta menyertakan tautan ke pertimbangan desain utama yang tercantum dalam tabel tersebut. CoBi adalah mitra AI multipihak yang utamanya bertujuan membantu mahasiswa sekolah menengah mengembangkan keterampilan belajar kolaboratif dalam kelompok kecil secara lebih baik. Sistem ini dirancang untuk meningkatkan keterampilan belajar kolaboratif mahasiswa, dengan fokus khusus pada dimensi hubungan dalam kolaborasi. Tujuannya adalah memberdayakan pengajar dan mahasiswa dengan menyediakan wawasan berharga yang memfasilitasi pengalaman belajar kolaboratif yang lebih bermakna dan menarik. CoBi secara khusus menyasar dimensi hubungan dalam kolaborasi, mendorong dinamika tim yang positif, serta memandang kolaborasi sebagai tujuan yang bernilai pada dirinya sendiri.

Selama sesi kolaborasi, CoBi secara otomatis mengidentifikasi dan memvisualisasikan ungkapan-ungkapan yang mencerminkan "wacana yang membangun secara sosial" dalam percakapan kelompok mahasiswa, yang berkaitan dengan *Community Agreements* (CA) atau Kesepakatan Komunitas tertentu. Kesepakatan ini berasal dari materi sains OpenSciEd untuk jenjang K-12 yang selaras dengan standar nasional (*OpenSciEd, t.thn*.). CA tersebut merupakan norma perilaku yang disepakati bersama melalui proses negosiasi antara mahasiswa dan pengajar untuk memandu kolaborasi di kelas. Saat ini, CoBi berfokus mendukung tiga CA: membangun komunitas (atau berkomitmen terhadap komunitas kita), bergerak dan berpikir maju, serta bersikap saling menghargai. Aspek kunci dari CoBi adalah ekstraksi dan penyajian "temuan" (*noticings*) yaitu contoh spesifik penerapan positif dari CA yang muncul dalam percakapan mahasiswa. Umpan balik kualitatif ini dirancang untuk mendorong refleksi dan transparansi, serta membantu mahasiswa memahami bagaimana percakapan mereka berkontribusi pada pembangunan komunitas dalam situasi kolaboratif.

Implementasi desain CoBi di kelas berakar pada serangkaian lokakarya desain bersama (*co-design*) yang melibatkan para remaja; mereka menyatakan keinginan agar AI dapat membantu mereka membangun komunitas yang kuat di dalam kelas. Sistem ini beroperasi melalui empat tahapan berbeda di kelas:

1. **Negosiasi Bersama terkait Kesepakatan**. Dengan bimbingan pengajar, mahasiswa bekerja dalam kelompok kecil untuk merumuskan dan memasukkan contoh-contoh CA versi mereka sendiri ke dalam antarmuka CoBi. Langkah awal ini memastikan bahwa norma-norma yang ditetapkan telah disepakati bersama dan relevan dengan konteks mahasiswa.
2. **Analisis dan Visualisasi Wacana secara *Real-Time***. Selanjutnya, mahasiswa terlibat dalam tugas-tugas pembelajaran kolaboratif. Saat mereka mengerjakan tugas, CoBi secara aktif menganalisis wacana lisan mereka untuk mendeteksi adanya kesepakatan-kesepakatan tersebut, yang disebut sebagai "*noticings*" (penandaan momen kesepakatan). Audio direkam menggunakan mikrofon *omnidirectional* (*Yeti Blue*) yang diletakkan di meja setiap kelompok. Mikrofon ini dipilih secara khusus karena harganya yang terjangkau dan kemudahan penggunaannya dalam menangkap percakapan antara dua hingga empat mahasiswa. Audio dialirkan dalam potongan berdurasi 10 detik, diproses menggunakan *Whisper* untuk *Automatic Speech Recognition* (ASR), lalu dianalisis oleh model bahasa RoBERTa yang telah disesuaikan secara khusus (*fine-tuned*). Untuk menjaga privasi mahasiswa, hasil analisis digabungkan dari berbagai kelompok mahasiswa dan divisualisasikan pada tingkat kelas. Pengajar dapat memantau visualisasi ini secara *real-time*; visualisasi tersebut dapat berupa diagram radar untuk ringkasan kuantitatif atau animasi pohon yang tumbuh dengan bunga-bunga yang bermekaran sebagai penanda berbagai kesepakatan yang berbeda (lihat Gambar 12.1).
3. **Refleksi Kolaboratif**. Setelah tugas pembelajaran kolaboratif selesai, pengajar menggunakan visualisasi CoBi untuk memandu Mahasiswa merefleksikan sejauh mana wacana kelompok mereka selaras dengan kesepakatan kolaborasi (CA) yang telah mereka sepakati bersama. Proses ini dirancang untuk menumbuhkan wawasan yang lebih mendalam mengenai dinamika kolaborasi dan meningkatkan pengalaman belajar secara keseluruhan.
4. **Membangun Literasi AI**. Pengajar memiliki opsi untuk menampilkan "*noticings*" dengan peringkat teratas (contoh spesifik dari penerapan positif CA) yang diidentifikasi oleh CoBi selama sesi berlangsung. Transparansi ini sangat penting karena mendorong refleksi dan diskusi yang lebih mendalam mengenai kekuatan serta keterbatasan sistem AI. Sebagai contoh, mahasiswa dapat mendiskusikan potensi kesalahan kategorisasi yang disebabkan oleh gangguan suara (*noise*) atau isyarat kontekstual yang terlewatkan, sehingga sekaligus mengembangkan keterampilan literasi AI yang kritis dan relevan dengan abad ke-21.

Dalam konteks dimensi pada Tabel 12.1, CoBi berfungsi sebagai alat penilaian formatif yang beroperasi secara *real-time* di lingkungan kelas yang dinamis, di mana struktur tugas atau kendali yang ketat sering kali tidak tersedia. CoBi dapat diterapkan pada berbagai jenis tugas kolaboratif dan mampu memproses komunikasi mahasiswa yang bersifat terbuka (*open-ended*). Mitra AI ini diwujudkan melalui serangkaian visualisasi yang menunjukkan tingkat kepatuhan terhadap Kesepakatan Kolaborasi (CA) serta contoh-contoh pengamatan perilaku.

Keputusan untuk menampilkan visualisasi gabungan kepada mahasiswa dan pengajar diambil secara sengaja guna meminimalkan gangguan bagi mahasiswa, sehingga mereka dapat lebih fokus pada proses kolaborasi itu sendiri daripada sekadar hasil akhirnya. CoBi ditampilkan pada layar besar di dalam kelas, memungkinkan mahasiswa dan pengajar untuk melihat hasil upaya kolaborasi mereka secara *real-time* selama tugas berlangsung. Pengajar dapat memanfaatkan visualisasi tersebut untuk memandu diskusi setelah tugas kolaborasi selesai, guna mendorong mahasiswa berpikir lebih mendalam mengenai kualitas kolaborasi yang telah mereka lakukan.



Gambar 12.1 Kesepakatan mahasiswa CoBi dan representasi pohon mengenai kepatuhan terhadap kesepakatan tersebut.

Pemodelan dan Bukti dalam CoBi

Untuk menilai wacana kolaboratif, potongan audio yang masuk diproses menggunakan Whisper sebuah model ASR akses terbuka mutakhir yang telah disesuaikan (*fine-tuned*) untuk mengenali ucapan anak-anak guna mentranskripsikan ucapan individu. Tantangan signifikan dalam tahap ini adalah kondisi lingkungan kelas yang bising, yang mencakup suara percakapan mahasiswa yang saling tumpang tindih, kebisingan latar belakang, serta variabilitas alami dan pola bicara anak-anak yang cenderung kurang jelas dibandingkan orang dewasa. Kondisi lingkungan seperti ini menyebabkan tingkat kesalahan kata (*Word Error Rate* atau WER) yang tinggi, berkisar antara 60% hingga 70% pada berbagai kumpulan data tingkat sekolah menengah pertama.

Namun, ketika keluaran ASR tersebut digunakan dalam tugas klasifikasi berbasis NLP tahap lanjutan untuk mengidentifikasi ucapan terkait CPS (*Collaborative Problem Solving*), dampak dari tingginya WER dapat diredam secara. Sebagai contoh, penggunaan ucapan dengan tingkat kepercayaan (*confidence*) ASR yang tinggi dapat mengurangi kesalahan klasifikasi, sehingga memastikan bahwa agen hanya menganalisis ucapan mahasiswa yang tingkat akurasinya cukup meyakinkan. Perlu dicatat pula bahwa bias rasial dalam sistem ASR merupakan hal yang patut diwaspadai. Berbagai pembelajaran menunjukkan tingkat akurasi yang lebih rendah bagi penutur berkulit hitam dibandingkan penutur berkulit putih akibat

perbedaan pola akustik. Kesalahan-kesalahan ini dapat berdampak berantai dan memengaruhi klasifikasi wacana tahap selanjutnya secara negatif jika tidak ditangani.

CoBi terutama menggunakan model RoBERTa yang telah disesuaikan (*fine-tuned*) untuk mengklasifikasikan tuturan mahasiswa. Model-model tersebut disesuaikan secara individual menggunakan dataset beranotasi manusia dengan label biner untuk setiap CA. Label CA dan pendekatan anotasi tersebut diadopsi agar selaras dengan konstruk dari kerangka kerja CPS karya Sun dan rekan-rekan. Model-model tersebut kemudian disesuaikan lebih lanjut dan diuji menggunakan data beranotasi dari tugas dan konteks kolaboratif yang sepenuhnya berbeda guna mengembangkan model yang lebih dapat digeneralisasi.

Untuk meningkatkan ketahanan terhadap kesalahan ASR serta memperbaiki kemampuan generalisasi, CoBi menerapkan augmentasi ASR selama proses penyesuaian model. Teknik ini melibatkan pelatihan model menggunakan tuturan hasil transkripsi manusia maupun transkripsi ASR, yang secara efektif melipatgandakan data pelatihan sekaligus memaparkan model pada jenis-jenis kesalahan yang ditemui dalam situasi dunia nyata. Metode ini telah menunjukkan peningkatan akurasi yang signifikan serta mengurangi kesenjangan kinerja antara transkrip manusia dan transkrip ASR. Model menghasilkan nilai probabilitas antara 0 dan 1 untuk setiap tuturan, yang menunjukkan kemungkinan tuturan tersebut mencerminkan CA tertentu, di mana probabilitas di atas 0,5 menandakan kecocokan positif. CoBi juga menghasilkan "*noticings*" (temuan pengamatan), yaitu contoh spesifik dan positif dari CA yang teridentifikasi dalam tuturan mahasiswa.

Perlu dicatat bahwa model-model lain juga sedang dieksplorasi, termasuk LLM seperti Mistral 7B. Model-model LLM ini digunakan dalam konfigurasi seperti *few-shot prompting* atau dikombinasikan dengan pengklasifikasi *support vector machine* (SVM) untuk *embedding* yang dihasilkannya, guna meningkatkan kemampuan generalisasi di berbagai konteks dan kurikulum. Ditinjau dari aspek pemodelan, bukti, dan umpan balik pada Tabel 12.1, model komputasional tersebut selaras secara langsung dengan konstruk CA melalui proses *fine-tuning* pada data yang dianotasi manusia, di mana tuturan diberi label berdasarkan ada atau tidaknya CA spesifik tersebut. "Temuan" (*noticings*) yang diekstraksi dan disajikan sebagai umpan balik merupakan contoh konkret dari perilaku kolaboratif ini, yang secara langsung menggambarkan bagaimana tuturan Mahasiswa mencerminkan CA tersebut.

Visualisasi umpan balik ini selanjutnya memandu Mahasiswa dan pengajar dalam merefleksikan sejauh mana wacana kolaboratif mereka selaras dengan kesepakatan yang telah dinegosiasikan bersama, sehingga mendukung pengembangan keterampilan kolaborasi kritis dan literasi AI. Fokus ini sejalan dengan konsep validitas kriteria, yang memastikan bahwa model selaras dengan konstruk yang bermakna bagi para pemangku kepentingan dan hasil pendidikan, seperti apakah mahasiswa sedang mengembangkan pemikiran mereka atau menunjukkan sikap saling menghargai.

Cakupan dan Skala CoBi

Hingga saat ini, CoBi telah digunakan oleh ratusan kelompok mahasiswa dari berbagai distrik sekolah. Dari segi skala, CoBi dibangun di atas arsitektur modular yang dapat ditingkatkan skalanya (*scalable*) dan di-hosting di *Amazon Web Services*, sehingga menjamin

keamanan akses serta penyimpanan data. Pengumpulan data di dalam kelas hanya memerlukan *Chromebook* dan mikrofon Yeti untuk setiap kelompok, karena seluruh pemrosesan AI dilakukan di sisi *backend*. CoBi diintegrasikan ke dalam rangkaian penganalisis cerdas multimodal yang lebih besar, yang mencakup berbagai alat analisis AI dan multimodal untuk pembelajaran yang lebih luas mengenai interaksi di kelas.

Meskipun pada awalnya CoBi terutama mengandalkan aspek bahasa, pengembangan yang sedang berlangsung mencakup integrasi isyarat non-verbal (misalnya, gestur, postur tubuh, dan arah pandangan mata) untuk menciptakan model kolaborasi yang lebih komprehensif dan tangguh, sehingga dapat mengakomodasi beragam gaya komunikasi. Secara keseluruhan, CoBi bertujuan untuk menumbuhkan keterampilan sosial dan literasi AI yang krusial melalui analisis data multipihak dan multimodal dalam konteks pendidikan yang autentik.

Pendekatan ini melampaui metode tradisional dalam menilai dan mengajarkan kolaborasi dengan menyajikan umpan balik teragregasi yang membantu mahasiswa serta pengajar untuk merefleksikan dan memahami bagaimana komunikasi mereka berkaitan dengan proses kolaboratif yang menjadi bagian dari kesepakatan kolaborasi (CA). Dengan demikian, pendekatan ini mengubah narasi: alih-alih berperan sebagai pengawas perilaku kolaboratif, AI justru mendukung mahasiswa untuk mematuhi kesepakatan kolaborasi yang telah disepakati bersama. Pendekatan ini dirancang untuk mengubah secara mendasar peran AI dalam situasi kolaboratif dengan memberdayakan mahasiswa untuk membangun hubungan kolaborasi dan keterampilan sosial yang lebih baik, alih-alih sekadar memperluas model tradisional bimbingan satu-lawan-satu atau sistem yang melakukan intervensi intrusif terhadap proses kolaborasi.

12.6 MENGUKUR DAN MENDUKUNG KETERAMPILAN KOLABORASI

Mengukur dan mendukung keterampilan kolaborasi merupakan hal yang sulit. Selama 20 tahun terakhir, telah terjadi kemajuan signifikan dalam pengembangan kerangka kerja dan tercapainya kesepakatan mengenai apa saja yang mencakup keterampilan kolaborasi. Dengan semakin besarnya integrasi teknologi digital di ruang kelas, terdapat lebih banyak peluang untuk menangkap informasi kontekstual yang kaya mengenai kolaborasi mahasiswa. Teknologi AI telah berkembang dengan laju eksponensial, di mana model bahasa besar (LLM) mampu memproses data multimodal yang kompleks secara waktu nyata (*real-time*). Namun, meskipun keterampilan kolaborasi, kerangka kerja, dan berbagai teknik AI telah banyak diteliti dalam riset mengenai penilaian dan pembelajaran kooperatif berbasis komputer, hanya sedikit yang telah diterapkan dalam penggunaan skala besar di dunia nyata untuk pengukuran kinerja atau pembelajaran. Penilaian CPS PISA 2015 merupakan pengecualian karena penilaian tersebut dilaksanakan lintas negara terhadap sampel yang melibatkan lebih dari 100.000 mahasiswa berusia 15 tahun. Akan tetapi, penilaian tersebut hanya dilakukan satu kali pada populasi terpilih. Lebih lanjut, kritik terhadap pendekatan PISA mempertanyakan sejauh mana pendekatan pengukuran berbasis komputer tersebut setia pada konstruk yang dituju, serta

sejauh mana penilaian individu mahasiswa terhadap tugas tersebut mencerminkan kolaborasi dalam situasi dunia nyata.

Dengan demikian, saat ini tidak ada penilaian kolaborasi berskala besar yang sedang digunakan, ataupun perangkat lunak pendidikan yang diadopsi secara luas dan dirancang secara khusus untuk mendukung atau menilai kolaborasi. Salah satu alasan kurangnya adopsi secara luas di ruang kelas adalah bahwa pengumpulan data multimodal menghadirkan tantangan besar terkait privasi dan persetujuan yang harus ditangani dengan cermat. Pengumpulan data berupa ucapan, video, ekspresi wajah, pola tatapan mata, dan data biometrik lainnya dari mahasiswa menimbulkan kekhawatiran yang sangat sensitif; data tersebut dapat mengungkap informasi pribadi mengenai kondisi emosional, pola perhatian, dan bahkan kondisi kesehatan mahasiswa yang melampaui apa yang relevan secara pedagogis.

Oleh karena itu, agar pengumpulan data semacam itu dapat terwujud, bidang ini perlu terlebih dahulu menunjukkan kepada mahasiswa, orang tua, pengajar, dan administrator bahwa penyediaan data mereka menghasilkan nilai pendidikan melalui perangkat kolaborasi tersebut. Peningkatan ini dapat dicapai dengan menunjukkan bagaimana AI dapat bermitra dengan mahasiswa dan pengajar untuk mendukung pembelajaran melalui cara-cara yang transparan dan terbuka untuk dikritisi, seperti fitur *noticings* (pengamatan/catatan) pada CoBi. Kedua, kita perlu mengubah konsep persetujuan berdasarkan informasi (*informed consent*) dari sekadar keputusan biner untuk ikut serta atau tidak (*opt-in/opt-out*) menjadi proses berkelanjutan dan terukur yang menjaga transparansi serta akuntabilitas selama pengumpulan data berlangsung.

Kompleksitas regulasi memperumit tantangan privasi ini, karena berbagai undang-undang mengatur aspek yang berbeda terkait data mahasiswa: FERPA membatasi pengungkapan catatan pendidikan, COPPA mewajibkan persetujuan orang tua untuk pengumpulan informasi dari anak-anak di bawah usia 13 tahun, dan PPRA mengatur survei yang membahas topik sensitif; hal ini menciptakan serangkaian persyaratan beragam yang harus dipenuhi oleh pendidik dan peneliti sembari tetap menjamin perlindungan hak serta martabat mahasiswa. Jika kita meninjau konteks di mana kolaborasi diukur dalam skala besar, kita perlu melihat dunia kerja, tempat aturan mengenai persetujuan tidak seketat di dunia pendidikan. Sebagai contoh, di bidang medis, survei, observasi, dan kuesioner terstruktur secara rutin digunakan untuk menilai persepsi, sikap, komunikasi, kepemimpinan, dan interaksi.

Sebuah meta-analisis terhadap studi kerja sama tim menelaah 51 artikel pembelajaran yang mencakup 72 intervensi unik dan 8.439 partisipan. Hasilnya menunjukkan bahwa studi-studi tersebut terutama menggunakan ukuran hasil kinerja, observasi manusia, dan hasil survei, alih-alih ukuran keterampilan otomatis. Meskipun dunia kerja kini sangat terdigitalisasi dengan penggunaan surel, pesan instan, log komunikasi, dan pengukuran kinerja analisis otomatis terhadap kerja sama tim belum diadopsi secara luas. Dengan demikian, bidang ini telah mencapai titik di mana kita perlu mempertimbangkan apa yang diperlukan untuk memajukan kolaborasi melampaui eksperimen laboratorium, demonstrasi, penerapan di kelas berskala kecil, dan observasi manusia. Kerangka kerja dan model AI yang mampu menilai

kolaborasi secara akurat hanyalah sebagian kecil dari apa yang dibutuhkan untuk berhasil menerapkan, memvalidasi, dan mengimplementasikan kolaborasi dalam skala luas. Yang diperlukan adalah pemikiran ulang mengenai bagaimana kolaborasi dinilai dan diajarkan.

Meskipun terdapat standar untuk mengajarkan dan menilai keterampilan ini, kolaborasi bukanlah sebuah kurikulum tersendiri. Keterampilan kolaborasi perlu diintegrasikan ke dalam berbagai konteks (misalnya, Sains, Matematika, Desain) di mana negosiasi, koordinasi, dan pertukaran pengetahuan sangat penting dalam pelaksanaan tugas dan proses pembelajaran. Untuk mencapai hal ini, teknologi kolaboratif harus dapat diterapkan secara umum di berbagai konteks. Hal ini akan mempermudah pengembangan dan integrasinya ke dalam beragam kurikulum, serta membuat perangkat pembelajaran dan metode penilaian lebih dapat diterima oleh sekolah dan praktik pengajaran yang sudah ada. Evolusi AI yang berkelanjutan akan semakin memungkinkan analisis yang lebih canggih terhadap data interaksi kolaboratif. Model Bahasa Besar (LLM) dengan cepat mengubah karakteristik data pembelajaran yang dapat dianalisis, menunjukkan bahwa penilaian keterampilan berpikir tingkat tinggi serta pemberian umpan balik yang terarah menjadi lebih dapat dilakukan.

Sebagai contoh, laporan *Future of Teaching and Learning* dari Kantor Teknologi Pendidikan menyatakan, "Ke depannya, kita harus mengembangkan kemampuan AI yang selaras dengan prinsip-prinsip pembelajaran kolaboratif dan sosial, serta menghargai mahasiswa tidak hanya dari aspek kognitifnya, tetapi juga dari keseluruhan rangkaian keterampilan manusiawi yang dimilikinya". Bidang ini dengan cepat mengadopsi pembelajaran mengenai pendekatan integrasi komunikasi non-verbal (gerak tubuh, postur, arah pandangan mata) demi meningkatkan akurasi dan memperoleh pemahaman yang lebih holistik tentang kolaborasi. Namun, terdapat pula kebutuhan mendesak akan kerangka kerja yang transparan, etis, dan adil bagi integrasi AI, baik di ruang kelas maupun dalam penilaian yang berdampak besar (*high-stakes assessments*).

Oleh karena itu, perlu adanya peningkatan upaya mendorong inovasi yang bertanggung jawab dalam pengembangan dan penerapan teknologi-teknologi ini. Manfaat pembelajaran perlu diseimbangkan dengan tantangan terkait hilangnya privasi dan kekhawatiran mengenai akurasi pengukuran. Untuk mewujudkan visi ini, bidang tersebut perlu menggali wawasan dari para pendidik, spesialis pengukuran, ilmuwan kognitif, inovator AI, ahli etika, dan spesialis interaksi manusia-komputer guna membangun pendekatan seimbang yang berpusat pada manusia dalam mengintegrasikan AI untuk mendukung kolaborasi.

BAB 13
DAMPAK AI TERHADAP VALIDITAS DAN VALIDASI TES

13.1 VALIDITAS DALAM PENGUJIAN BERBASIS AI

Validitas telah digambarkan sebagai salah satu konsep terpenting dalam pengujian pendidikan dan. Pentingnya konsep ini berakar pada fakta bahwa, dalam konteks tes pendidikan dan psikologi, validitas mengacu pada ketepatan dasar teoretis, logis, dan empiris yang mendukung klaim suatu program pengujian. Dengan kata lain, validitas mengacu pada sejauh mana bukti dan teori mendukung penggunaan suatu tes untuk tujuan tertentu. Sebelum membahas lebih dalam mengenai dampak kecerdasan buatan (AI) terhadap validitas, kita perlu menguraikan berbagai lapisan makna yang terkandung dalam definisi yang tampak sederhana ini. Istilah-istilah seperti "bukti", "teori", "penggunaan", dan "tujuan" memerlukan penjelasan mengenai maknanya dalam konteks pengujian, khususnya ketika AI digunakan dalam berbagai aspek pengembangan tes, pelaksanaan tes, pemberian skor, pelaporan skor, dan interpretasi skor. Dalam bab ini, pertama-tama kami menyajikan gambaran umum mengenai konsep-konsep tersebut. Kemudian, kami menguraikan bagaimana konsep-konsep itu dipengaruhi oleh penggunaan AI.

Pada saat penulisan ini dibuat, AI sedang mengubah seluruh aspek pengujian pendidikan dan psikologi, mulai dari penentuan apa yang dapat diukur hingga pembuatan konten tes, pemberian skor atas respons, dan pelaporan skor. Dalam bab ini, kami mengeksplorasi bagaimana perubahan-perubahan tersebut memengaruhi sejauh mana bukti dan teori mendukung penggunaan suatu tes. Sebagaimana akan terlihat jelas, aspek-aspek paling mendasar dari validitas dan validasi tetap bertahan di era AI ini; namun, cara kita menyikapi aspek-aspek tersebut dan melakukan validasi terhadap program pengujian memerlukan perspektif baru.

Memahami Validitas

Mengingat bahwa validitas mengacu pada sejauh mana bukti dan teori mendukung penggunaan tes untuk tujuan tertentu, lembaga penyelenggara tes harus mendokumentasikan dasar pemikiran mengenai mengapa tes tersebut diperlukan (yaitu, teori yang mendasari penggunaan tes) serta bukti yang mereka peroleh untuk mendukung penggunaan tes tersebut. Dalam hal ini, kita memahami validitas bukan sebagai sifat tetap yang melekat pada tes yang memungkinkan pengguna membuat interpretasi apa pun atau membenarkan penggunaan apa pun melainkan sebagai sesuatu yang bergantung pada konteks penggunaan tertentu. Artinya, sebuah tes tidak lantas "memiliki" atau "tidak memiliki" validitas, karena validitas mengacu pada pembenaran atas interpretasi, keputusan, dan tindakan yang diambil berdasarkan skor tes. Di era kecerdasan buatan (AI) ini, pembenaran semacam itu menjadi semakin penting karena apa yang diukur oleh tes, bagaimana pengukuran dilakukan, dan makna skor tersebut bisa jadi kurang transparan dibandingkan dengan tes tradisional. Dua langkah pertama dalam pengembangan tes yang bertanggung jawab adalah merumuskan tujuan program tes dan mendeskripsikan apa yang diukur oleh tes tersebut. Agar tes dapat memenuhi tujuan yang

dimaksudkan—termasuk interpretasi hasil tes yang tepat—kedua langkah ini merupakan hal yang mutlak diperlukan. Sifat mutlak dari langkah-langkah ini menjadi semakin jelas ketika kita menyadari bahwa keduanya menjadi landasan bagi validasi tes. Berikut ini kami uraikan pentingnya masing-masing langkah tersebut.

Merumuskan Tujuan dan Penggunaan Tes

Tes dikembangkan dengan tujuan-tujuan tertentu. Hal ini merupakan pengamatan yang didasarkan pada praktik pengembangan tes selama berpuluh-puluh tahun sekaligus pengakuan bahwa tes adalah sebuah teknologi; sebagaimana teknologi pada umumnya, terdapat niat eksplisit dalam pembuatan tes tersebut. Pembuat tes (atau lembaga penyelenggara tes) berkewajiban untuk tidak hanya mendefinisikan tujuan-tujuan tersebut secara jelas, tetapi juga mendefinisikan penggunaan yang dimaksudkan. Tujuan dan penggunaan yang dimaksudkan dari suatu program tes menentukan sasaran bagi perancangan, pengembangan, dan validasi tes. Karena validitas mengacu pada sejauh mana bukti dan teori mendukung penggunaan tes untuk tujuan tertentu, maka tujuan-tujuan (yang dimaksudkan) itulah yang menentukan inferensi dan tindakan yang perlu divalidasi.

Penting untuk dicatat bahwa tes dapat digunakan untuk tujuan selain tujuan awal perancangannya. Dalam situasi seperti itu, pengguna tes berkewajiban untuk memberikan pembenaran atas penggunaan spesifik tersebut melalui teori dan bukti. Oleh karena itu, untuk setiap penggunaan tes, bukti validitas harus disajikan guna membenarkan penggunaan tersebut. Bukti semacam itu menunjukkan kepada peserta tes, pengguna, dan masyarakat umum bahwa klaim mengenai makna hasil tes memiliki nilai, serta bahwa tindakan yang diambil berdasarkan skor tes tersebut dapat dipertanggungjawabkan secara empiris.

AI kemungkinan besar akan memperluas cakupan tujuan yang dapat dicapai oleh tes pendidikan dan psikologis. Sebagai contoh, jika tujuan tes adalah menyediakan informasi diagnostik bagi Mahasiswa secara individu maupun kelompok kelas, AI dapat membantu mewujudkan tujuan tersebut dengan merangkum kinerja peserta tes secara lebih baik, baik untuk peserta tes secara perorangan maupun kelompok. Perkembangan terkini di bidang ini mencakup penggunaan AI untuk menyusun deskripsi naratif yang merangkum hasil dari kelompok mahasiswa yang besar atau dari berbagai asesmen bagi seorang mahasiswa.

Demikian pula, AI kini digunakan untuk meningkatkan kualitas umpan balik yang diberikan kepada mahasiswa setelah mereka mengikuti tes. Berbagai keunggulan baru dari program tes ini perlu divalidasi untuk memastikan apakah informasi yang dihasilkan benar-benar bermanfaat. Dengan demikian, kebutuhan untuk memvalidasi setiap tujuan program tes bukanlah hal baru; namun, tujuan dan klaim baru yang dimungkinkan oleh AI merupakan hal baru yang akan terus berkembang di masa depan. Tujuan-tujuan yang didukung oleh AI ini kemungkinan akan memperluas cakupan aspek yang perlu divalidasi dalam program tes kontemporer.

Menjelaskan Apa yang Diukur oleh Tes

Teori-teori yang mendasari dan menjelaskan apa yang diukur oleh suatu tes berasal dari berbagai sumber, bergantung pada konteks pengujiannya. Dalam banyak konteks pendidikan, teori pembelajaran dan standar kurikulum menentukan pengetahuan dan

keterampilan yang menjadi sasaran tes, serta bagaimana respons Mahasiswa terhadap butir-butir tes mencerminkan pengetahuan dan keterampilan tersebut. Dalam pengujian lisensi dan sertifikasi seperti saat kandidat diuji untuk mendapatkan lisensi di bidang penataan rambut, akuntansi, keperawatan, dan sebagainya analisis mengenai pengetahuan, keterampilan, dan kemampuan yang diperlukan untuk praktik yang aman dan berhasil biasanya digunakan untuk menjelaskan apa yang diukur oleh tes tersebut. Dalam konteks psikologis seperti identifikasi gangguan kejiwaan, pengukuran kecemasan, kepribadian, dan fenomena psikologis lainnya teori mengenai perilaku, kepribadian, sikap, neuropsikologi, dan patologi juga digunakan untuk mendefinisikan apa yang diukur.

Terdapat begitu banyak atribut berbeda yang diukur oleh tes pendidikan dan psikologis sehingga istilah *konstruc* diperkenalkan untuk menjelaskan apa yang sedang diukur oleh suatu tes. Saat memperkenalkan konsep ini, Cronbach dan Meehl mendefinisikan konstruk sebagai suatu atribut yang dipostulatkan pada diri seseorang, yang diasumsikan tercermin dalam kinerja tes... Kita mengharapkan seseorang pada waktu tertentu memiliki atau tidak memiliki atribut kualitatif... atau memiliki tingkat tertentu dari atribut kuantitatif...

Para ahli psikometri (mereka yang mengembangkan dan mengevaluasi tes) menganggap gagasan tentang konstruk ini bermanfaat dalam beberapa hal, termasuk dalam mendefinisikan apa yang diukur oleh tes dan menyediakan kerangka kerja untuk pembelajaran validasi. Sebagai contoh, istilah representasi konstruk digunakan untuk menggambarkan sejauh mana suatu tes merepresentasikan ranah pengetahuan, keterampilan, sikap, kepribadian, dan sebagainya yang hendak diukurnya. Jika tes mengukur atribut atau perilaku lain yang tidak tercakup dalam definisi konstruk, faktor-faktor "pengganggu" tersebut disebut sebagai hal yang tidak relevan dengan konstruk (*construct irrelevant*).

Dengan demikian, validitas dalam pengujian didukung dengan penyediaan bukti mengenai representasi konstruk, dan ketiadaan ketidakrelevanan konstruk. Berfokus pada aspek ketidakvalidan, Messick menulis, "Tes merupakan alat ukur konstruk yang tidak sempurna karena tes tersebut bisa jadi mengabaikan hal yang seharusnya disertakan... atau justru menyertakan hal yang seharusnya ditiadakan, atau bahkan keduanya". Dengan demikian, konsep representasi konstruk dan ketidakrelevanan konstruk menjadi penting dalam mengevaluasi sejauh mana suatu tes memenuhi tujuan yang ditetapkan. Kami meyakini bahwa pendefinisian mengenai apa yang diukur oleh suatu tes akan menjadi semakin penting di era pengujian yang diperkuat oleh kecerdasan buatan (AI), mengingat sering kali sulit untuk memahami bagaimana AI menghasilkan keluarannya serta sulit pula untuk mempertanggungjawabkan proses tersebut dalam kerangka relevansi konstruk.

Sebagai contoh, jika pengembang tes dan peneliti tidak dapat mempertahankan penggunaan AI dalam penyusunan konten tes dari perspektif teoretis, maka konten yang dihasilkan harus dapat dipertanggungjawabkan berdasarkan penilaian ahli mengenai sejauh mana konten tersebut relevan dan representatif terhadap konstruk yang dituju. Ketika asesmen didasarkan pada analisis data yang sudah ada mengenai seseorang, atau dilengkapi dengan data semacam itu, perlu disertakan landasan teori yang kuat mengenai bagaimana data tersebut mendukung interpretasi konstruk yang dimaksud. Dalam istilah non-psikometrik,

pihak yang mengembangkan dan menggunakan tes wajib memberikan bukti bahwa asesmen tersebut benar-benar mengukur apa yang hendak diukur; klaim semacam itu harus diawali dengan definisi yang memadai mengenai tujuan pengukuran tes serta penggunaan hasil tes yang tepat.

Singkatnya, pengembang tes bertanggung jawab untuk menjabarkan apa yang diukur oleh suatu asesmen serta tujuan penggunaan hasil tes tersebut. "Definisi konstruk" dan "pernyataan tujuan penggunaan" yang menyertainya ini membangun pemahaman bersama mengenai objek pengukuran di antara lembaga penyelenggara tes dan seluruh pemangku kepentingan (peserta tes, pengguna tes, pembuat kebijakan, dan lain lain.). Selain itu, hal ini memungkinkan terciptanya pemahaman bersama mengenai nilai yang diharapkan dapat diperoleh dari penggunaan hasil tes tersebut.

Validasi Tes dan Bukti Validitas

Pada bagian sebelumnya, kami telah mendefinisikan secara singkat konsep validitas tes beserta beberapa istilah penting untuk memahaminya. Pada bagian ini, kami menguraikan validasi tes, yaitu proses mengevaluasi tes berdasarkan (a) bagaimana tes tersebut dikembangkan, dan (b) penyelidikan terhadap kualitas serta kegunaannya. Baik (a) maupun (b) melibatkan analisis bukti dan menekankan pentingnya menyertakan prosedur pengendalian mutu serta mengikuti praktik terbaik saat mengembangkan tes, sekaligus melakukan studi tambahan untuk memastikan tes tersebut memenuhi tujuan khususnya.

Proses validasi tes pada dasarnya merupakan sebuah evaluasi, yang menyiratkan adanya penilaian akhir bahwa penggunaan tes untuk tujuan tertentu dapat dibenarkan berdasarkan bukti yang telah dievaluasi. Dengan demikian, tes dikatakan "divalidasi" untuk tujuan tertentu apabila bukti yang terkumpul mendukung kesimpulan tersebut. Sebagaimana akan kami uraikan pada bagian-bagian selanjutnya dalam bab ini, penggunaan AI dalam asesmen kemungkinan besar memerlukan studi tambahan di luar apa yang selama ini tercantum dalam riset mengenai validitas. Namun, AI juga dapat dimanfaatkan untuk memperkaya studi validitas dengan menyediakan cara-cara baru dalam mengumpulkan bukti validitas. Standar AERA dan rekan-rekan menguraikan lima sumber bukti validitas "yang dapat digunakan dalam mengevaluasi validitas dari interpretasi yang diusulkan atas skor tes untuk penggunaan tertentu". Sumber-sumber tersebut meliputi bukti validitas yang didasarkan pada:

(a) konten tes,
(b) proses respons,
(c) struktur internal,
(d) hubungan dengan variabel lain dan
(e) konsekuensi pengujian.

Pada saat penulisan ini dibuat, standar tersebut sedang direvisi, dan diperkirakan akan ada penambahan sumber bukti validitas keenam yang didasarkan pada interpretasi skor tes. Berbagai sumber bukti ini menggambarkan jenis-jenis studi yang dapat dilakukan untuk menyelidiki aspek-aspek spesifik validitas serta menjawab pertanyaan penting terkait validitas, seperti "Apakah tes tersebut mengukur apa yang diklaimnya?", "Apakah tes tersebut

menghasilkan skor yang sebanding di berbagai format pengujian?", dan "Apakah tes tersebut bias terhadap jenis peserta tes tertentu?". Pada bagian ini, kami menguraikan secara singkat kelima sumber bukti validitas tersebut serta bagaimana bukti-bukti itu dapat digunakan untuk menyusun argumen validitas yakni sintesis dari bukti dan teori yang mendukung penggunaan suatu tes untuk tujuan tertentu.

Bukti validitas yang didasarkan pada konten tes mengacu pada informasi yang mencerminkan sejauh mana konten tes selaras dengan tujuan pengujian yang dimaksudkan. Informasi semacam itu biasanya didasarkan pada analisis independen terhadap konten tes oleh pakar materi pelajaran (*subject-matter experts* atau SME) eksternal; mereka meninjau dan menilai butir soal serta rubrik penilaian berdasarkan sejauh mana elemen-elemen tersebut merepresentasikan konstruk yang ditetapkan oleh pengembang tes secara memadai. Studi validitas konten menyusun tugas peninjauan butir soal dengan berbagai cara dan sering kali melibatkan evaluasi kritis terhadap keselarasan antara tujuan yang diukur oleh butir soal dengan tujuan yang sebenarnya dimaksudkan. Belakangan ini, kecerdasan buatan (AI) telah digunakan untuk menyediakan data keselarasan dengan cara melakukan penilaian kesesuaian antara butir soal dan tujuan serupa dengan penilaian oleh pakar manusia atau dengan menyelaraskan tujuan di berbagai jenis kurikulum maupun tugas pekerjaan.

Bukti validitas yang didasarkan pada proses respons mengacu pada data yang diperoleh dari penyelidikan mengenai apakah proses kognitif dan perilaku yang digunakan peserta tes untuk menyelesaikan masalah dalam suatu asesmen memang merupakan proses dan perilaku yang ditargetkan untuk diukur. Bentuk tradisional dari bukti validitas berbasis proses respons berasal dari berbagai sumber, termasuk studi *think-aloud* (di mana peserta tes menjelaskan penalaran mereka saat mengerjakan butir soal dan tugas asesmen) atau wawancara kognitif (di mana pertanyaan berbasis hipotesis diajukan kepada peserta tes setelah mereka menyelesaikan asesmen).

Pemodelan matematis terhadap karakteristik tugas yang memengaruhi tingkat kesulitan butir soal merupakan bentuk lain dari bukti validitas yang didasarkan pada proses respons. Belakangan ini, "*data log*" yang dihasilkan oleh asesmen berbasis digital telah digunakan untuk memberikan bukti validitas berdasarkan proses respons, dan dalam hal inilah teknik AI terbukti sangat bermanfaat. Sebagai contoh, AI dapat digunakan untuk menelusuri data log dalam jumlah besar guna mengidentifikasi pola respons yang mengungkap strategi atau proses berpikir peserta tes saat menjawab butir soal.

Bukti validitas berdasarkan struktur internal mengacu pada analisis respons peserta tes terhadap butir-butir tes untuk memahami struktur data tersebut dan mengevaluasi apakah struktur yang teramati konsisten dengan struktur yang dijelaskan dalam definisi konstruk (yaitu, apakah struktur data yang teramati sesuai dengan struktur yang dimaksudkan). Jenis bukti ini bersifat statistik dan menggunakan teknik deteksi atau konfirmasi dimensionalitas, seperti analisis faktor eksploratori atau konfirmatori, atau analisis residual model, seperti sejauh mana data tes yang teramati sesuai dengan model teori respons butir. Bukti validitas berdasarkan struktur tes juga dapat digunakan untuk mengevaluasi apakah struktur tes konsisten di berbagai subkelompok peserta tes atau di berbagai moda pelaksanaan asesmen.

Wells menyajikan pemantauan komprehensif mengenai cara melakukan analisis "invariansi pengukuran" semacam itu.

Salah satu penerapan AI yang menarik dan sedang berkembang dalam ranah bukti validitas berbasis struktur internal adalah penggunaan model bahasa berskala besar (*large language models*) untuk menentukan struktur dimensional dari teks penyusun tes berdasarkan fitur-fitur butir, seperti makna semantik, kompleksitas linguistik, dan atribut lainnya. Seperti halnya penerapan validitas lainnya, AI kemungkinan besar akan sangat memperluas ragam perangkat yang tersedia bagi ahli psikometri untuk mengumpulkan dan menganalisis bukti validitas berdasarkan struktur internal.

Bukti validitas berdasarkan hubungan dengan variabel lain diperoleh dari studi pembelajaran yang menyertakan skor tes sebagai salah satu variabel dalam analisis serta data dari sumber lain sebagai variabel tambahan. Studi-studi ini menelaah dan mendokumentasikan hubungan antara skor tes dengan faktor-faktor yang relevan maupun yang tidak relevan dengan konstruk tersebut. Contoh umum dari studi yang relevan dengan konstruk adalah hubungan antara skor tes penerimaan dan kinerja selanjutnya di sekolah menengah atau perpengajaran tinggi berdasarkan indeks prestasi kumulatif (IPK).

Korelasi antara skor tes mahasiswa sebelumnya dan IPK mereka di kemudian hari diharapkan bernilai besar dan positif, mengingat tujuan prediktif dari asesmen semacam itu. Model regresi yang memprediksi IPK berdasarkan skor tes mahaMahasiswa dan kriteria penerimaan lainnya (misalnya, nilai SMA) juga lazim digunakan dalam konteks tes penerimaan serta dapat dimanfaatkan untuk mengevaluasi kekuatan prediksi di berbagai kelompok peserta tes. Contoh analisis validitas terkait faktor yang tidak relevan dengan konstruk (*construct-irrelevant validity*) adalah meneliti hubungan antara kemahiran digital dan skor tes kemampuan membaca, atau antara kompleksitas linguistik dan skor tes matematika. Dalam kedua kasus tersebut, korelasi yang tinggi akan mengindikasikan adanya faktor-faktor yang tidak relevan dengan konstruk yang memengaruhi performa tes, sehingga melemahkan argumen validitas. Kami memperkirakan bahwa AI dapat mempermudah pengumpulan data mengenai variabel eksternal untuk memperoleh bukti validitas.

Bukti validitas yang didasarkan pada konsekuensi pengujian mengacu pada hasil penyelidikan mengenai dampak penggunaan tes terhadap individu, komunitas, organisasi, dan masyarakat secara umum. Studi tradisional umumnya berfokus pada dampak hasil tes terhadap "hak-hak" tertentu, seperti perolehan kredensial atau penerimaan kerja oleh pemberi kerja. Jika tes digunakan sebagai kriteria utama untuk pemberian manfaat semacam itu, tingkat pencapaian yang diraih oleh peserta tes dari berbagai subkelompok perlu dievaluasi untuk mendeteksi potensi bias terhadap subkelompok tersebut.

Sebagai contoh, Komisi Kesempatan Kerja Setara (*Equal Employment Opportunity Commission*) Amerika Serikat menerapkan aturan "empat perlima" (*four-fifths rule*) untuk mengevaluasi dampak merugikan dari prosedur seleksi terhadap kelompok minoritas. Secara khusus, jika tingkat kelulusan kelompok minoritas kurang dari 80% dari tingkat kelulusan kelompok mayoritas, kelompok minoritas tersebut dapat mengajukan gugatan hukum terhadap program pengujian tersebut. Bukti validitas berdasarkan konsekuensi pengujian

mencakup analisis yang lebih komprehensif yang dikaitkan dengan konsekuensi yang diharapkan dari suatu program pengujian. Sebagai contoh, untuk tes yang dirancang guna meningkatkan kualitas pembelajaran, studi dapat dilakukan untuk mengevaluasi perubahan dalam praktik pengajaran yang berkaitan dengan pelaksanaan program pengujian tersebut.

Penyelidikan validitas harus mencakup bukti bahwa tujuan program pengujian telah tercapai, sekaligus memastikan tidak timbulnya dampak negatif yang tidak diinginkan. Kecerdasan Buatan (AI) dapat bermanfaat dalam menyediakan bukti semacam itu melalui studi simulasi yang memodelkan berbagai skenario hasil tes yang mungkin terjadi; namun, tentu saja bukti tersebut perlu dicermati dan diuji secara saksama. Bukti validitas yang didasarkan pada interpretasi skor tes yang tepat merupakan sumber bukti validitas yang diusulkan oleh O'Leary dan rekan-rekan. Bukti ini diperoleh dari studi sistematis mengenai sejauh mana interpretasi dan tindakan yang didasarkan pada hasil tes dilakukan secara tepat oleh pengguna tes. Sebagai contoh, jika pengajar kelas diharapkan menindaklanjuti hasil tes untuk menyesuaikan strategi pembelajaran bagi mahasiswa mereka, mereka harus mampu memahami informasi yang tercantum dalam laporan skor tes sebagaimana dimaksudkan oleh pengembang tes. Demikian pula, jika umpan balik penilaian bertujuan membantu mahasiswa memahami dan memperbaiki kesalahan mereka, mahasiswa harus memahami umpan balik tersebut agar tujuan itu dapat tercapai. Studi mengenai seberapa baik pengguna menginterpretasikan skor tes masih jarang dilakukan, namun hal ini sangat diperlukan. Akan menarik untuk melihat apakah AI dapat bermanfaat bagi para pengembang tes dan peneliti di bidang ini.

Merangkum dan Mendokumentasikan Bukti Validitas: Argumen Validitas

Pada bagian sebelumnya, kami telah menguraikan berbagai sumber bukti validitas dan menyoroti aspek-aspek di mana penerapan AI dapat memberikan manfaat khusus. Pada bagian ini, kami menekankan pentingnya merangkum bukti dan teori validitas agar calon pengguna tes dapat mengambil keputusan yang tepat mengenai kegunaan tes untuk tujuan mereka, serta memastikan bahwa seluruh pemangku kepentingan dapat memercayai informasi yang dihasilkan oleh skor tes. Proses merangkum dan mendokumentasikan seluruh bukti validitas ini dapat dipandang sebagai penyusunan "argumen validitas" sebuah argumen yang menyatakan bahwa akumulasi bukti yang ada mendukung penggunaan tes tersebut untuk interpretasi dan tujuan yang dimaksudkan.

Meskipun argumen validitas merupakan dokumen ringkasan yang memfasilitasi pengambilan keputusan yang tepat terkait penggunaan tes, dokumen ini juga bersifat komprehensif dan merupakan hasil dari perencanaan matang sebelumnya mengenai jenis-jenis bukti yang perlu dikumpulkan serta dianalisis secara mendalam. Analisis ini bertujuan untuk mengevaluasi sejauh mana tes tersebut memenuhi tujuan awalnya dan memastikan tidak timbulnya konsekuensi yang tidak diinginkan. Mengingat potensi AI yang menjanjikan dalam merangkum data dalam jumlah besar, kami memperkirakan AI akan digunakan untuk merangkum hasil studi individual (khususnya studi kualitatif yang melibatkan data wawancara atau diskusi kelompok terarah dalam jumlah besar), serta membantu mengintegrasikan berbagai hasil tersebut menjadi sebuah argumen validitas yang koheren.

13.2 ASESMEN YANG DISEMPURNAKAN OLEH AI DAN DAMPAKNYA

Pada bagian sebelumnya, kami telah menguraikan konsep tradisional mengenai validitas dan validasi tes, serta bagaimana AI dapat dikaitkan dengan pembelajaran dan praktik dalam pengujian pendidikan dan psikologis. Pada bagian ini, pertama-tama kami menjelaskan jenis-jenis asesmen berbasis AI yang sedang berkembang serta bagaimana fitur-fiturnya berdampak pada berbagai topik yang telah dibahas sebelumnya. Asesmen berbasis AI mengacu pada penggunaan teknologi AI termasuk *machine learning* (ML), *deep learning* (DL), dan penerapannya bersamaan dengan *natural language processing* (NLP) untuk merancang, menyelenggarakan, memberikan skor, menafsirkan, dan menyempurnakan asesmen pendidikan dan psikologis. Sistem-sistem ini mampu mengotomatisasi pembuatan butir soal, menyesuaikan konten tes bagi peserta didik secara *real-time*, memberikan umpan balik yang segera dan dipersonalisasi, serta mendukung pendidik dengan wawasan berbasis data untuk keperluan pengajaran dan intervensi.

Meskipun kami mengakui semakin besarnya pengaruh perangkat pendidikan berbasis AI seperti dasbor analitik pembelajaran atau sistem tutor cerdas (*intelligent tutoring systems*; perangkat-perangkat tersebut bukanlah fokus utama bab ini. Fokus kami lebih tertuju pada asesmen berbasis AI yang secara khusus dirancang untuk mengukur atau memberikan sertifikasi atas keterampilan, atau yang digunakan untuk memantau sistem pendidikan, memastikan akuntabilitas, dan memandu proses pengajaran. Asesmen berbasis AI melampaui cakupan pengujian tradisional dengan memungkinkan evaluasi keterampilan individu yang dinamis, dapat ditingkatkan skalanya (*scalable*), dan lebih personal, namun di saat yang sama juga menghadirkan tantangan baru terkait validitas dan keadilan. Tabel 13.1 merangkum bagaimana berbagai aplikasi berbasis AI dalam asesmen pendidikan berkontribusi terhadap validitas sekaligus berpotensi mengancamnya. Tabel tersebut menyoroti peran ganda AI, baik sebagai katalisator inovasi maupun sebagai sumber risiko psikometrik baru. Pada bagian selanjutnya, kami meninjau kembali kerangka kerja validitas untuk membantu memaksimalkan manfaat asesmen berbasis AI sembari meminimalkan risiko yang terkait dengannya. Namun sebelumnya, kami menyajikan ringkasan mengenai manfaat serta risiko terhadap validitas yang ditimbulkan oleh berbagai aplikasi berbasis AI dalam penilaian pendidikan.

Tabel 13.1 Manfaat dan Ancaman terhadap Validitas yang Terkait dengan Aplikasi Penilaian Berbasis AI

Aplikasi Berbasis AI	Manfaat Validitas	Ancaman Validitas	Referensi
Identifikasi dan pengembangan konstruk	Mendukung definisi konstruk berbasis teknologi (misalnya, literasi AI, rekayasa prompt)	Kurang terwakilinya konstruk	Dorsey and Michaels (2022a), Hao et al. (2024), Swiecki et al. (2022) and Yan et al. (2024)
Pembuatan butir otomatis	Meningkatkan cakupan konten, pengembangan butir yang hemat biaya	Risiko menghasilkan butir berkualitas rendah atau bias	LaFlair et al. (2023), Suárez-Álvarez et al. (2026), von Davier

			(2018) and von Davier et al. (2024)
Jenis butir inovatif	Meningkatkan autentisitas dan keterlibatan peserta didik	Dapat menimbulkan varians yang tidak relevan dengan konstruk dan hambatan aksesibilitas	Foster and Piacentini (2023) and Swiecki et al. (2022)
Kalibrasi parameter butir	Mengurangi kebutuhan untuk pengujian awal yang ekstensif	Risiko estimasi parameter yang tidak akurat	Guonole et al. (2024), Suárez-Álvarez et al. (2026), and Tran et al. (2025)
Penyelarasan konten	Mengurangi beban kerja pakar bidang melalui pemetaan berbantuan AI	Potensi ketidaksesuaian antara butir dan standar	Butterfuss and Doran (2024)
Penskoran otomatis	Meningkatkan efisiensi penskoran dan mengurangi bias manusia	Transparansi dan interpretabilitas algoritma penskoran terbatas	Bejar (2017), LaFlair et al. (2023), Lane (2023) and Shermis (2022)
Personalisasi	Memungkinkan adaptasi terhadap kebutuhan peserta didik secara individual	Tantangan untuk standardisasi dan keterbandingan yang ketat, hasil yang stereotip	Arslan et al. (2024), Sireci et al. (2024) and Yan, von Davier, and Weiss (2024)
Keamanan tes	Meningkatkan deteksi kecurangan dalam tes dan pengawasan	Mungkin sulit untuk menetapkan kesetaraan dan keterbandingan butir	Foster (2020)

Identifikasi dan Pengembangan Konstruk

Teknologi AI memungkinkan identifikasi dan penyempurnaan konstruk baru yang semakin relevan dalam lingkungan pembelajaran yang dimediasi secara digital. Konstruk seperti literasi AI, pemikiran komputasional, dan rekayasa *prompt* mulai membentuk lanskap pengukuran. Pengenalan pola berbasis AI pada data berskala besar seperti berkas log, interaksi pembelajar, dan respons terbuka dapat membantu mengonseptualisasikan serta mengoperasionalkan konstruk-konstruk ini dengan cara-cara baru. Namun, ketergantungan pada AI menimbulkan risiko kurangnya representasi konstruk (*construct underrepresentation*). Sistem AI mungkin lebih mengutamakan fitur yang mudah diekstraksi atau dimodelkan, seperti pola leksikal atau perilaku yang tampak di permukaan, sehingga berpotensi mengabaikan dimensi kognitif atau sosio-emosional yang lebih mendalam yang menjadi tujuan penilaian tersebut. Bias terhadap fitur yang dapat dikuantifikasi ini dapat mendistorsi validitas konstruk jika tidak ditangani melalui landasan teoretis yang kuat dan pengawasan manusia.

Pembuatan Butir Soal Otomatis

Pembuatan butir soal otomatis *(Automated Item Generation* atau AIG) memanfaatkan AI untuk menghasilkan butir soal tes dalam jumlah besar dari templat, sehingga memfasilitasi cakupan materi yang luas dan mengurangi biaya pengembangan. Dengan menghasilkan varian butir soal secara sistematis, AIG meningkatkan generalisasi dan representasi konstruk sekaligus mendukung penyajian tes yang adaptif. Namun, AIG juga dapat memperbesar risiko terhadap validitas jika proses pembuatannya tidak disertai dengan pengendalian mutu oleh manusia yang memadai. Model butir soal yang dirancang kurang tepat atau data pelatihan yang bias dapat menghasilkan butir soal yang tidak sesuai secara linguistik atau budaya, tidak akurat tingkat kesulitannya, atau tidak selaras dengan konstruk yang dituju. Risiko-risiko ini menegaskan perlunya validasi iteratif yang melibatkan manusia (*human-in-the-loop*), terutama dalam konteks berisiko tinggi serta lingkungan yang multikultural dan multibahasa.

Jenis Butir Soal Inovatif

Penggunaan AI memungkinkan pengembangan format butir soal inovatif seperti simulasi interaktif, asesmen berbasis skenario, dan dialog berbasis obrolan (*chat*). Jenis-jenis butir soal ini menawarkan representasi yang lebih kaya mengenai keterampilan terapan dan proses pengambilan keputusan, sehingga meningkatkan validitas ekologis dan konsekuensial. Desain semacam itu sangat berharga dalam bidang-bidang yang menuntut penalaran dinamis atau kompetensi interpersonal. Akan tetapi, format inovatif dapat secara tidak sengaja memunculkan varians yang tidak relevan dengan konstruk jika terdapat perbedaan tingkat keakraban peserta tes terhadap antarmuka digital atau jika lingkungan simulasi tidak dapat diakses secara setara. Perbedaan dalam literasi digital, konektivitas, atau kompatibilitas teknologi bantu (*assistive technology*) dapat mengganggu keadilan dan keterbandingan hasil tes.

Kalibrasi Parameter Butir Soal

AI dapat membantu dalam mengestimasi parameter butir soal dengan memanfaatkan metode seperti jaringan Bayesian atau regresi *machine learning* (ML) untuk menyimpulkan parameter butir soal berdasarkan karakteristik butir soal atau informasi awal. Pendekatan-pendekatan ini mengurangi beban logistik dari uji coba awal (*pilot testing*) tradisional sekaligus mempercepat pengembangan bank soal yang telah dikalibrasi. Inovasi ini menawarkan peningkatan dibandingkan model statistik tradisional yang digunakan untuk mengestimasi kompleksitas butir soal.

Kendati demikian, penggunaan parameter butir semu (*pseudo-item parameters*) memerlukan verifikasi empiris untuk memastikan bahwa estimasi parameter butir tidak tidak stabil atau bias. Sejauh mana pendekatan baru dalam kalibrasi butir soal ini mendukung interpretasi skor yang valid dan adil masih memerlukan pembelajaran lebih lanjut. Namun, beberapa program tes telah melaporkan hasil yang menjanjikan dengan menggunakan pendekatan berbasis AI ini.

Keselarasan Konten

Sebagaimana disebutkan sebelumnya, algoritma AI dapat digunakan untuk memetakan butir soal asesmen dengan standar pembelajaran atau kerangka kurikulum,

sehingga membantu keselarasan instruksional dan mengurangi beban kerja para ahli materi pelajaran (SME). Jika dilatih dengan tepat, alat-alat ini dapat meningkatkan transparansi dan efisiensi proses penyelarasan. Akan tetapi, satu hal yang perlu diwaspadai dalam bidang ini adalah bahwa sistem penyelarasan otomatis masih rentan terhadap kesalahan klasifikasi dan tidak memiliki kemampuan penilaian bernuansa yang dimiliki pakar manusia dalam menafsirkan maksud kurikulum. Ketidakselarasan dapat mengarah pada kesimpulan yang menyesatkan mengenai efektivitas instruksional atau kemajuan peserta didik, yang pada akhirnya mengancam aspek validitas, baik dari segi konten maupun konsekuensi.

Penilaian Otomatis

Sistem penilaian otomatis, yang didukung oleh NLP dan DL, memberikan evaluasi yang cepat dan konsisten terhadap respons terbuka, seperti esai, jawaban singkat, dan bahkan interaksi lisan. Sistem ini mengurangi biaya penilaian, meminimalkan bias penilai, dan menawarkan mekanisme umpan balik yang dapat diskalakan. Namun, model penilaian algoritmik sering kali beroperasi sebagai "kotak hitam" (*black box*), sehingga sulit untuk memastikan bahwa skor mencerminkan konstruk yang dimaksudkan. Kurangnya transparansi ini menimbulkan ancaman terhadap validitas interpretatif dan menantang kemampuan pengembang tes untuk mengaudit proses penilaian, terutama dalam lingkungan yang berisiko tinggi serta beragam secara budaya dan bahasa.

Personalisasi

AI mendukung tes adaptif dan jalur pembelajaran yang dipersonalisasi yang merespons karakteristik individu pembelajar. Dengan menyesuaikan konten penilaian terhadap tingkat kemampuan dan preferensi, sistem ini dapat meningkatkan keterlibatan, mengurangi kecemasan, dan meningkatkan presisi pengukuran, khususnya bagi pembelajar yang secara historis kurang terlayani. Akan tetapi, personalisasi juga menghadirkan tantangan terkait komparabilitas. Ketika pembelajar yang berbeda menerima butir soal yang berbeda, sulit untuk memastikan bahwa skor mereka dapat disetarakan atau diinterpretasikan dalam skala yang sama. Kekhawatiran ini sangat relevan dalam konteks yang melibatkan sertifikasi, penempatan, atau akuntabilitas, namun cenderung kurang berdampak signifikan pada penilaian yang dirancang untuk mendukung pembelajaran.

Keamanan Tes

Alat pengawasan (*proctoring*) dan deteksi kecurangan berbasis AI meningkatkan keamanan tes dengan memantau perilaku yang tidak wajar, termasuk pola gerakan mata, aktivitas peramban (*browser*), dan isyarat audio. Alat-alat ini dapat mencegah pelanggaran dan melindungi integritas penilaian yang berisiko tinggi. Namun, menjaga keamanan tes melalui rotasi bentuk tes yang sering dilakukan atau penghapusan butir soal secara reaktif dapat memengaruhi kesetaraan bentuk tes dari waktu ke waktu, sehingga mengurangi stabilitas skor dan keterbandingan longitudinal. Selain itu, ketergantungan yang berlebihan pada pengawasan dapat menimbulkan kekhawatiran etis terkait privasi dan otonomi mahasiswa.

13.3 PEMBELAJARAN VALIDASI UNTUK ASESMEN BERBASIS AI

Karena AI berpotensi mengubah konsep dasar mengenai apa yang dimaksud dengan "tes", hal ini kemungkinan akan berdampak pada praktik validasi dengan memperluas cakupannya guna memastikan definisi konstruk dipahami dengan baik dan penggunaan AI yang tidak transparan (yaitu, model "kotak hitam" atau *black box*) justru mendukung, bukan mengurangi, validitas. Cakupan dan kecanggihan metode komputasi berbasis AI berkembang pesat, sehingga bentuk-bentuk asesmen yang sebelumnya tidak terbayangkan kini menjadi semakin mungkin untuk diterapkan. Sebagai contoh, agen asesmen berbasis percakapan (*conversational assessment agents*) kini digunakan untuk menilai keterampilan komunikasi, begitu pula dengan simulasi yang diperkuat oleh AI untuk keperluan pembelajaran dan asesmen. AI siap mengubah praktik-praktik inti validasi bukan dengan cara meninggalkannya, melainkan dengan mendorong peninjauan kembali terhadap asumsi-asumsi yang telah lama dipegang. Jika kita membandingkan praktik validasi AI (yaitu, bagaimana alat AI dievaluasi di luar konteks tes pendidikan) dengan validasi tes, kita dapat melihat munculnya praktik-praktik baru. Sebagai contoh, studi kapasitas sering kali dilakukan terhadap aplikasi AI untuk memahami kemampuan lain yang dimiliki teknologi AI tersebut di luar tujuan utamanya. Selain itu, sebagaimana telah disebutkan sebelumnya, ketika asesmen diperkuat menggunakan AI dan melibatkan model yang tidak transparan (*opaque models*), model-model tersebut perlu dikuantifikasi dan diinterpretasikan.

Meskipun demikian, bahkan ketika AI digunakan di sepanjang proses asesmen baik untuk pemberian skor, interpretasi, maupun pengambilan Keputusan persyaratan mendasar mengenai validitas tetap berlaku, yaitu mengevaluasi sejauh mana suatu tes memenuhi tujuan yang ditetapkan dan tidak menimbulkan konsekuensi yang tidak diinginkan. Perlu dicatat bahwa banyak sumber bukti yang diusulkan oleh Cronbach dan Meehl, Messick), serta pakar lainnya sebagai bagian dari proses validasi masih tetap relevan untuk digunakan: perbedaan antar-kelompok, analisis korelasi/regresi, analisis faktor, studi mengenai perubahan dari waktu ke waktu, studi tentang proses respons, dan sebagainya. Bahkan ketika sebuah model bersifat tidak transparan (*opaque*), selama model tersebut menghasilkan keluaran yang dapat dikuantifikasi, prosedur-prosedur validasi tersebut tetap dapat diterapkan. Namun, perubahan signifikan muncul ketika sistem AI tidak mengharuskan dihasilkannya satu skor tunggal yang terpadu untuk mendukung pengambilan keputusan. Berbeda dengan penilaian tradisional yang mengandalkan model penskoran satu dimensi, sistem AI dapat memanfaatkan keluaran (*output*) yang kompleks dan multidimensi.

Sebagai contoh, dalam jaringan saraf tiruan (*neural networks*), pengambilan keputusan dapat didasarkan pada vektor berdimensi tinggi yang dihasilkan pada lapisan akhir yang secara efektif berfungsi sebagai skor, meskipun bukan dalam pengertian konvensional. Evolusi ini menantang gagasan konvensional mengenai apa yang kita maksud dengan "skor" dalam pengujian. Dengan demikian, kerangka kerja validitas mungkin perlu diperbarui untuk mengakomodasi sistem di mana representasi numeriknya lebih kompleks daripada sekadar metrik ringkasan tunggal; bahkan ketika diterapkan pada satu instrumen yang mungkin tetap kita sebut sebagai "tes". Cara lain AI menantang teori validitas adalah dengan mendorong

peninjauan kembali terhadap konsep "tes" itu sendiri. Jika kita mengadopsi definisi pengukuran yang luas yaitu pemberian angka atau kategori kepada individu maka banyak model AI (terutama yang digunakan untuk klasifikasi atau prediksi) termasuk dalam cakupan pengukuran. Namun, penyamaan ini mengaburkan perbedaan penting.

Sifat banyak model AI yang tidak transparan dan sulit diinterpretasikan (*opaqueness*) menghadirkan perbedaan kualitatif dibandingkan dengan tes tradisional. Sebuah tes bukan sekadar sistem teknologi yang memberikan keluaran kepada individu; tes merupakan pengalaman terstruktur yang dirancang untuk memancing perilaku atau respons seseorang guna mendukung inferensi tertentu. Tes dirancang secara sengaja untuk memicu respons kognitif atau emosional, sehingga pengalaman mengikuti tes menjadi hal sentral bagi validitas inferensi yang diambil. Tersirat dalam desain ini adalah asumsi mengenai kemampuan untuk dijelaskan (*explainability*) dan kemampuan untuk digeneralisasi (*generalizability*) bahwa cara kerja tes dapat dipahami, dan hasilnya dapat diperluas maknanya melampaui pelaksanaan tes pada saat itu saja. Sebaliknya, banyak aplikasi AI beroperasi tanpa transparansi atau landasan teoretis semacam itu; hal ini menyiratkan bahwa proses validasi perlu mencakup penjelasan transparan mengenai relevansi konstruk dari apa yang disajikan kepada peserta tes serta bagaimana mereka merespons konten tes tersebut. Dorsey dan Michaels menguraikan kompleksitas penggunaan AI dalam membantu pengembangan dan penskoran tes, sembari tetap memastikan representasi konstruk dan interpretasi skor yang valid. Sebagaimana mereka ungkapkan,

Kita mungkin menggunakan metode *Natural Language Processing* (NLP) untuk menilai wawancara, esai, atau bentuk penilaian berbasis teks lainnya, namun sering kali mengabaikan fakta bahwa bahasa atau teks memiliki berbagai tingkat abstraksi yaitu sintaksis, leksikal, semantik, dan pragmatis yang melibatkan serangkaian inferensi berbasis konstruk, yang pada gilirannya melibatkan interaksi kompleks antara budaya dan bahasa. Oleh karena itu, berdasarkan kerangka kerja validitas modern, standar yang lebih ketat mungkin mengharuskan adanya bukti bahwa algoritma berbasis NLP mampu memahami secara akurat kata atau teks yang dihasilkan, mengenali alasan pendukung yang diberikan oleh peserta tes, serta mengevaluasi relevansi dan ketepatan penalaran tersebut. Seiring dengan semakin meluasnya penerapan AI di berbagai bidang, teknologi baru mungkin akan semakin mengambil alih peran yang sebelumnya dipegang oleh tes tradisional dalam sistem kemasyarakatan seperti memandu pengambilan keputusan, mengalokasikan peluang, dan memberikan sertifikasi kompetensi tanpa harus menyerupai penilaian konvensional baik dari segi bentuk maupun fungsinya. Ketika batas antara pengujian dan penerapan AI menjadi semakin kabur, kerangka kerja yang kita gunakan untuk menilai ketepatan inferensi dan penggunaan tes pun harus ikut menyesuaikan diri.

Oleh karena itu, kami mendorong pembaca untuk memanfaatkan AI guna melengkapi, bukan menggantikan, metode validasi tradisional. Meski demikian, perlu dicatat bahwa upaya regulasi yang ditujukan untuk mengatur AI kemungkinan besar akan memberikan dampak yang meluas hingga ke luar ranah AI itu sendiri, serta memengaruhi berbagai sistem hibrida di mana AI berperan sebagai komponen pendukung maupun komponen dasar. Pergeseran yang

lebih luas ini mengisyaratkan bahwa masa depan teori validitas tidak hanya akan dibentuk oleh kemajuan di bidang psikometri, tetapi juga oleh evolusi harapan masyarakat, arsitektur teknologi, dan kerangka kerja hukum.

Bagaimana AI Memengaruhi Representasi dan Interpretasi Konstruk?

Sulit untuk sepenuhnya memprediksi bagaimana AI akan berkembang, mengingat pesatnya kemajuan teknologi ini, besarnya investasi yang masuk, serta beragam harapan dan kepentingan yang melingkupinya. AI tetap menjadi ranah yang dinamis dan eksperimental—sebuah "wadah yang mendidih" penuh inovasi. Namun, di tengah ketidakpastian ini, beberapa dampak yang mungkin terjadi terhadap representasi konstruk dalam penilaian sudah mulai terlihat. Dalam penerapannya yang paling langsung, AI mulai mengubah cara konstruk dioperasionalisasikan, dideteksi, dan disimpulkan. Baik melalui NLP, pelacakan perilaku, maupun integrasi data multimodal, AI menghadirkan cara-cara baru untuk menangkap bukti mengenai pembelajar yang mungkin berbeda dari definisi konstruk tradisional. Pergeseran ini menantang para ahli teori validitas untuk meninjau kembali sejauh mana data yang ditangkap oleh sistem AI selaras dengan konstruk yang dituju, serta apakah sistem tersebut benar-benar mengukur apa yang diklaimnya atau justru memunculkan konstruk baru sama sekali tanpa landasan teoretis yang eksplisit.

Sebagai contoh, penggunaan AI dalam pembuatan butir soal (*item*) memperluas batasan cara kita mendefinisikan "ranah konten" dalam penilaian. Secara tradisional, ranah konten dirumuskan melalui cetak biru, kerangka kerja, atau serangkaian perilaku dan kompetensi yang hendak direpresentasikan oleh tes tersebut. Namun, pembuatan butir soal berbasis AI sering kali mengandalkan masukan eksplisit seperti korpus teks, dokumen kurikulum, atau materi sumber lainnya sebagai panduan dalam menciptakan butir soal.

Akibatnya, materi-materi tersebut semakin berfungsi sebagai definisi *de facto* dari ranah konten itu sendiri. Pergeseran ini menempatkan sumber konten bukan sekadar sebagai referensi, melainkan sebagai komponen aktif dalam representasi konstruk. Lebih jauh lagi, demi menjaga integritas tes dan menghindari masalah butir soal yang mudah ditebak akibat materi persiapan yang beredar luas, pengembang tes kini mungkin memerlukan sumber konten "emas" atau terbatas yakni kumpulan pengetahuan rahasia yang telah dikurasi untuk memandu pembuatan butir soal namun tidak dapat diakses oleh peserta tes. Hal ini menghadirkan lapisan kompleksitas baru dalam representasi konstruk, di mana batas antara konstruk, konten, dan protokol keamanan mulai menjadi kabur.

Dalam konteks ini, AI memperbesar isu lama namun sering kali kurang mendapat perhatian dalam penilaian: adanya kesenjangan antara ranah konten yang dinyatakan dan ranah konten aktual yang digunakan untuk menyusun tes. Meskipun ketidaksesuaian ini sebenarnya sudah ada sejak lama, penggunaan AI dalam pembuatan butir soal membuatnya lebih terlihat jelas dan berpotensi membawa dampak yang lebih signifikan. Dalam praktiknya, kita dapat membayangkan tiga himpunan yang saling tumpang tindih:

(a) ranah konten yang dituju atau bersifat teoretis, sebagaimana ditetapkan oleh standar atau cetak biru;

(b) materi sumber yang digunakan khususnya dalam proses berbasis AI untuk menyusun butir soal; dan
(c) konten aktual yang disajikan kepada peserta tes. Ketidakselarasan di antara ketiga himpunan ini menimbulkan berbagai bentuk ketidaktepatan representasi konstruk.

Sebagai contoh, jika materi sumber yang digunakan oleh model AI memunculkan bias atau melewatkan aspek konten yang krusial, butir soal yang dihasilkan dapat mendistorsi konstruk yang sedang diukur—meskipun tes tersebut tampak mengikuti cetak biru aslinya. Kondisi ini menegaskan perlunya meninjau kembali bagaimana representasi konstruk didefinisikan, dioperasionalkan, dan diaudit di era ketika AI memainkan peran yang semakin besar dalam pengembangan tes. Ketegangan dalam representasi konstruk ini menjadi semakin nyata ketika AI generatif ditugaskan untuk menyusun keseluruhan tes. Persepsi pengembang tes mengenai domain konten dan apa yang sebenarnya digunakan oleh model AI untuk menghasilkan butir soal dapat berbeda secara halus namun signifikan. Bahkan ketika model bahasa besar (LLM) memberikan penjelasan mengenai cara ia menafsirkan perintah (*prompt*) atau mengikuti pedoman, proses internalnya sebagian besar tetap tidak transparan.

Oleh karena itu, mengandalkan jaminan dari sistem tersebut saja tidaklah cukup. Untuk memastikan kesesuaian dengan konstruk yang dituju, sangatlah penting untuk melakukan analisis *post hoc*—yaitu memeriksa butir soal yang dihasilkan dan memetakannya kembali ke cetak biru formal atau spesifikasi konten. Proses verifikasi keselarasan ini dapat menjadi praktik terbaik yang krusial dalam penggunaan AI generatif untuk *Automatic Item Generation* (AIG), yang membantu menjaga representasi konstruk dan mempertahankan validitas interpretasi skor yang dihasilkan. Setiap kali AI digunakan untuk meniru peran ahli materi pelajaran (SME), penilai manusia, atau peran lain yang secara tradisional dijalankan oleh manusia dalam proses penilaian, risiko kesalahan representasi konstruk pun muncul. Penting untuk dicatat bahwa risiko ini tidak serta-merta lebih besar dibandingkan risiko yang timbul dari keterlibatan manusia; bagaimanapun juga, penilaian manusia itu sendiri merupakan sumber ketidakjelasan dan variabilitas dalam pengujian.

Namun, penggantian peran manusia oleh AI memunculkan sorotan baru terhadap isu transparansi, bias, dan kesetiaan pada representasi konstruk. Sama halnya dengan tuntutan akan kemampuan penjelasan (*explainability*) yang semakin tinggi terhadap sistem AI, standar serupa juga wajar diterapkan pada penilai manusia, yang proses kognitifnya tidak kalah tidak transparannya. Kendati demikian, pembelajaran yang ada menunjukkan bahwa LLM tidak meniru peran manusia dengan tingkat akurasi yang seragam. Kinerja mereka dapat sangat bervariasi tergantung pada konteks dan dipengaruhi oleh data pelatihan yang sering kali bersifat eksklusif serta sangat dipengaruhi oleh norma budaya dan linguistik Barat, khususnya Amerika. Hal ini menimbulkan kekhawatiran mengenai distorsi konstruk ketika LLM diterapkan dalam beragam konteks pendidikan atau psikologis, karena interpretasi dan hasil keluarannya mungkin tidak mencerminkan konstruk yang dituju secara responsif terhadap budaya atau sesuai dengan konteksnya.

Perhatian khusus diperlukan ketika LLM digunakan untuk meniru peserta tes. Pembelajaran terkini telah mengeksplorasi penggunaan responden sintetis yaitu representasi

buatan AI yang menggantikan peserta tes sesungguhnya—sebagai metode untuk menentukan karakteristik butir soal atau menyimulasikan pola respons. Pendekatan ini sering kali melibatkan pembuatan profil pembelajar secara rinci, diikuti dengan pemodelan logis atas respons berdasarkan asumsi struktur pengetahuan (misalnya, jika seorang mahasiswa menguasai X dan Y, mereka diprediksi akan menjawab dengan cara tertentu). Meskipun menjanjikan untuk tujuan seperti kalibrasi butir soal atau evaluasi bentuk tes, prosedur ini memerlukan kerangka kerja validasi yang ketat dan spesifik.

Salah satu cara paling substansial untuk memvalidasi pembuatan respons sintetis adalah dengan membandingkan "klon" buatan AI tersebut dengan peserta tes nyata yang memiliki profil serupa. Yang terpenting, validasi semacam itu tidak boleh hanya terbatas pada pencocokan jawaban, tetapi juga harus melibatkan bukti proses respons—seperti membandingkan protokol *think-aloud* (penyampaian alur pikir secara lisan) atau alur penalaran. Jika teknik ini digunakan secara luas namun diterapkan tanpa langkah pengamanan yang memadai, hal ini menimbulkan risiko signifikan berupa kesalahan representasi konstruk, yang berpotensi melemahkan validitas interpretasi skor tes yang diperoleh dari data tersebut.

Ketika AI digunakan untuk pemberian skor, terdapat beberapa risiko krusial yang harus dipertimbangkan secara cermat. Salah satu kekhawatiran utama adalah potensi bias algoritmik (*International Test Commission* [ITC] & *Association of Test Publishers* [ATP]). Bergantung pada bagaimana sistem AI dilatih, data yang mendasarinya mungkin mengandung bias—baik yang bersifat historis, budaya, maupun demografis—yang kemudian melekat dalam model dan memengaruhi hasil pemberian skor. Bias ini dapat muncul secara global, memengaruhi seluruh peserta tes, atau secara lokal, memberikan dampak yang tidak proporsional terhadap subkelompok atau wilayah geografis tertentu. Masalah ini, yang secara luas dikenal sebagai bias algoritmik, telah didokumentasikan secara ekstensif dalam riset mengenai etika dan keadilan AI. Tantangan ini menjadi sangat nyata pada model AI yang bersifat tidak transparan atau disebut sebagai model "kotak hitam" (*black-box*), di mana proses pengambilan keputusan internalnya sulit untuk diinterpretasikan.

Sebagai tanggapan, berbagai metode statistik dan komputasi telah dikembangkan untuk mendeteksi, mengukur, dan memitigasi bias semacam itu, termasuk penggunaan metrik keadilan, audit bias, dan teknik penyesuaian *post hoc*. Meskipun demikian, penggunaan perangkat-perangkat ini memerlukan integrasi yang terencana ke dalam proses validasi guna memastikan bahwa sistem pemberian skor otomatis tidak mengorbankan keadilan dan validitas interpretasi skor tes. Selain itu, kemajuan terkini dalam teknologi AI telah memberikan tekanan yang belum pernah terjadi sebelumnya terhadap keamanan tes. Meskipun AI menawarkan perangkat yang menjanjikan untuk mempersonalisasi asesmen dan meningkatkan desain tes, teknologi ini juga memungkinkan bentuk-bentuk kecurangan baru yang semakin canggih. Hanya dalam dua tahun terakhir, muncul berbagai kasus pelanggaran yang melibatkan bantuan AI, termasuk penggunaan ekstensi peramban yang menghasilkan jawaban secara *real-time*, perangkat yang dapat dikenakan (*wearable devices*) seperti kacamata pintar yang terhubung dengan LLM, hingga penggunaan *deepfake* untuk penipuan identitas.

Dari perspektif validitas, pelanggaran semacam itu bukan sekadar pelanggaran prosedural, melainkan ancaman langsung terhadap kemampuan interpretasi dan dasar pertanggungjawaban skor tes. Lebih jauh lagi, ketika sebagian individu memiliki akses ke perangkat kecurangan berbasis AI sementara yang lain tidak, masalah ini melampaui aspek validitas dan menjadi persoalan keadilan. Dampak ganda terhadap akurasi inferensi maupun kesetaraan kondisi tes menegaskan perlunya mengintegrasikan pertimbangan keamanan tes ke dalam kerangka kerja validitas terkini. Dengan demikian, sejauh mana kecurangan dan ancaman keamanan lainnya terjadi tetap menjadi aspek penting dalam validasi di era kecerdasan buatan (AI) ini.

13.4 METODE EVALUASI PENILAIAN BERBASIS AI

Dalam konteks penilaian tradisional, desain tes sering kali lebih mengutamakan ketepatan psikometris dibandingkan pengalaman peserta tes. Pembelajaran terkini dalam bidang pengukuran pendidikan mengusulkan peninjauan kembali konsep validitas tes melalui sudut pandang pengalaman—sebuah pendekatan yang memandang pengalaman nyata peserta tes sebagai elemen integral bagi keberhasilan dan keadilan penilaian. Pergeseran ini selaras dengan tren yang lebih luas dalam desain yang berpusat pada manusia (*human-centered design*) dan mencerminkan pengakuan yang kian berkembang bahwa desain antarmuka, struktur umpan balik, serta persepsi mengenai keadilan turut memengaruhi keterlibatan maupun kinerja peserta.

Berdasarkan perspektif pengalaman ini, validitas tidak hanya dipandang sebagai aspek teknis dari interpretasi skor, melainkan juga sebagai pencapaian sosio-teknis yang bergantung pada keberhasilan penerapan desain yang dirasakan bermakna, mudah diakses, dan tepat. Dalam kerangka kerja ini, pembelajaran pengalaman pengguna (UX) menjadi relevan bagi validitas: memahami dan mengoptimalkan pengalaman peserta tes turut mendukung legitimasi kesimpulan yang diambil dari hasil penilaian. Oleh karena itu, teknik-teknik seperti protokol *think-aloud* (menyuarakan pikiran), pengujian kegunaan (*usability testing*), dan pelacakan emosi menjadi sarana penting untuk mengumpulkan bukti berbasis pengalaman guna mendukung validitas tes; selain itu, pendekatan desain bersama atau *co-design* (yang melibatkan peserta tes dan pemangku kepentingan lainnya dalam perancangan serta revisi penilaian) akan sangat bermanfaat.

Integrasi AI ke dalam sistem penilaian baik melalui penilaian otomatis, tes adaptif, maupun perancangan butir soal generative telah menghadirkan dinamika baru dalam pengalaman peserta tes. Sistem berbasis AI ini tidak sekadar menyajikan materi, tetapi juga beradaptasi secara *real-time* berdasarkan interaksi pengguna, sehingga memengaruhi konten maupun struktur penilaian itu sendiri. Evolusi ini menuntut perluasan cakupan pertimbangan UX ke ranah pengalaman algoritmik atau *algorithmic experience* (AX), yang mengacu pada pengalaman subjektif dan perilaku saat berinteraksi dengan sistem algoritmik. Dalam penilaian yang diperkuat oleh AI, AX mencakup:

- persepsi mengenai bagaimana mesin adaptif memilih dan menyajikan butir soal

- kemampuan untuk memahami logika (*interpretability*) dan persepsi keadilan terkait penilaian otomatis
- responsivitas dan kredibilitas tutor AI atau agen percakapan
- prediktabilitas dan transparansi perilaku sistem secara keseluruhan

AX negatif—seperti perasaan dinilai secara tidak adil oleh algoritma yang tidak transparan atau kebingungan akibat pemilihan butir soal yang tidak dapat diprediksi—dapat mengikis kepercayaan terhadap sistem dan, pada gilirannya, memengaruhi legitimasi hasil penilaian. Bidang interaksi manusia-AI telah mulai merumuskan kerangka kerja formal untuk mengevaluasi sistem algoritmik dari perspektif yang berpusat pada pengguna. Sebagai contoh, studi oleh Wang dan rekan-rekan mengusulkan kerangka kerja konseptual untuk menilai kepercayaan, persepsi agensi, dan kemampuan penjelasan (*explainability*) dalam interaksi AI. Kerangka kerja semacam itu dapat diterapkan secara langsung pada lingkungan pengujian berbasis AI, di mana persepsi mengenai keadilan, kejelasan, dan kendali dapat membentuk keterlibatan pengguna maupun proses pemberian respons.

Untuk menjamin validitas dalam penilaian yang diperkuat oleh AI, pembelajaran AX perlu diintegrasikan ke dalam proses perancangan tes dan validasi. Pendekatan metode campuran yang menggabungkan log perilaku, wawancara kualitatif, dan data afektif dapat mengungkap wawasan penting mengenai bagaimana peserta tes mengalami penilaian algoritmik dan bagaimana sistem dapat disempurnakan untuk mendukung pengukuran yang adil dan tepercaya.

Data Sintetis untuk Mensimulasikan Peserta Tes dan SME

Kemajuan terkini dalam LLM telah membuka peluang baru untuk menyimulasikan respons peserta tes dan umpan balik dari pakar materi pelajaran (SME) menggunakan data sintetis. Metode ini menawarkan alternatif yang dapat diskalakan dan hemat biaya untuk tahap awal evaluasi butir soal serta pengembangan tes. Dua jalur pembelajaran menunjukkan kelayakan dan tantangan dari pendekatan ini. Liu dan rekan-rekan mengkaji apakah LLM dapat menyimulasikan respons mahaMahasiswa terhadap butir soal aljabar tingkat perpengajaran tinggi dengan fidelitas psikometrik yang memadai. Mereka meminta enam model berbeda—termasuk GPT-3.5, GPT-4, dan LLaMA 3—untuk menghasilkan respons, lalu membandingkannya dengan respons aktual dari mahaMahasiswa sarjana. Hasilnya menunjukkan bahwa beberapa LLM mengungguli manusia dalam hal rata-rata tingkat kemahiran, namun tidak memiliki variabilitas distribusi yang ditemukan pada populasi nyata. Meskipun tidak ada satu model pun yang dapat sepenuhnya meniru distribusi kemahiran manusia, penggabungan beberapa LLM menghasilkan kumpulan responden sintetis dengan keselarasan yang lebih baik terhadap data manusia.

Studi mereka juga memperkenalkan strategi augmentasi yang mencampurkan respons nyata dan sintetis, serta menunjukkan bahwa sampel hibrida (misalnya, 50% manusia, 50% LLM) menghasilkan pemulihan parameter yang lebih baik dalam kerangka teori respons butir soal (*item response theory*). Dengan demikian, LLM dapat berfungsi sebagai "peserta tes virtual" untuk mengurangi beban pengumpulan data sembari tetap menjaga ketepatan metodologi psikometrik. Secara paralel, Lu dan Wang mengusulkan konsep *generative*

students—yaitu pembelajar hasil simulasi LLM yang memiliki profil kekuatan dan kelemahan yang telah ditetapkan sebelumnya. Agen-agen generatif ini digunakan untuk mengevaluasi dan menandai butir soal yang bermasalah selama proses pengembangan soal. Dengan mengontrol pola penguasaan keterampilan, mereka mampu menyimulasikan perilaku pengerjaan tes yang bernuansa serta memvalidasi kesesuaian butir soal dengan konstruk yang dituju.

Pendekatan ini mencerminkan pergeseran dari simulasi respons yang bersifat umum menuju pemodelan kognisi dan kinerja yang terarah serta berbasis profil, sehingga membawa pengembangan tes lebih dekat pada praktik validasi yang berlandaskan teori. Lebih jauh lagi, dengan menyimulasikan perilaku pakar materi pelajaran (SME) bersamaan dengan perilaku Mahasiswa, kerangka kerja mereka memungkinkan validasi paralel terhadap plausibilitas respons maupun koherensi penalaran yang merupakan dimensi kunci dalam desain asesmen modern. Potensi simulasi data berbasis LLM untuk pengembangan dan validasi asesmen sangat menjanjikan. Banyak bentuk bukti validitas terutama yang didasarkan pada analisis butir soal secara empiris, uji coba tes, atau evaluasi proses respon memerlukan pengumpulan data ekstensif dari partisipan manusia, yang memakan biaya besar, menyita waktu, dan rumit secara logistik. Penggunaan strategis model bahasa besar (*large language models*) untuk menyimulasikan atau menambah respons manusia menawarkan alternatif yang dapat ditingkatkan skalanya (*scalable*) guna mengumpulkan bukti validitas berdasarkan proses respons atau struktur internal; hal ini dapat mengurangi beban pembelajaran sembari tetap menjaga ketelitian analisis.

Jika kerangka kerja validasi yang kokoh dapat ditetapkan untuk mendukung penggunaan penuh data sintetis, pendekatan ini tidak hanya dapat menjadi pelengkap, tetapi dalam konteks tertentu, juga menjadi pengganti yang layak bagi pengujian empiris tahap awal. Dalam pengertian ini, simulasi berbasis LLM berpotensi menjadi alat transformatif untuk augmentasi data, yang memungkinkan pengembang mengeksplorasi fungsi butir soal, menyempurnakan cetak biru tes, dan mengungkap varians yang tidak relevan dengan konstruk (*construct-irrelevant variance*) semuanya dilakukan sebelum peserta tes yang sesungguhnya berinteraksi dengan tes tersebut.

Mengevaluasi Sistem Penilaian (*Scoring Engines*)

Dalam pengukuran pendidikan, fungsi penilaian baik yang dilakukan manusia maupun algoritma merupakan komponen sentral dari argumen interpretasi. Interpretasi tes mencerminkan klaim mengenai atribut psikologis yang mendasarinya berdasarkan proses respons yang telah ditetapkan. Validitas inferensi semacam itu sebagian bergantung pada apakah mekanisme penilaian menangkap konstruk yang dituju secara tepat. Dalam kasus sistem penilaian berbasis AI, hal ini mencakup lebih dari sekadar akurasi label keluaran; aspek yang juga perlu diperhatikan meliputi representasi konstruk, stabilitas lintas konteks, dan transparansi pengambilan keputusan. Messick menekankan penilaian sebagai tindakan perancangan—tindakan yang mewujudkan asumsi mengenai apa yang dianggap sebagai bukti pengetahuan atau kemampuan. Kesesuaian model penilaian, terutama yang bersifat otomatis, harus dievaluasi tidak hanya dari segi kinerja kriteria (misalnya, kesesuaian dengan penilai

manusia) tetapi juga dalam kaitannya dengan relevansi konstruk dan kurangnya representasi konstruk (*construct underrepresentation*).

Sebagai contoh, jika sistem penilaian AI lebih mengutamakan panjangnya jawaban atau kompleksitas sintaksis dibandingkan akurasi konseptual, sistem tersebut mungkin secara sistematis memberikan nilai lebih pada fitur-fitur yang tidak relevan, sehingga mengganggu kesetiaan skor terhadap konstruk yang dituju. Standar AERA dan rekan-rekan juga menekankan bahwa validasi sistem penilaian otomatis harus mencakup studi empiris mengenai kesepakatan antarpenilai (*inter-rater agreement*), pola kesalahan, dan generalisabilitas di berbagai subkelompok serta jenis instruksi soal (*prompt types*). Penting juga untuk mengkaji dinamika interaksi manusia-AI dalam sistem penilaian hibrida, di mana penilai manusia dapat dipengaruhi oleh rekomendasi AI—sebuah ranah yang belum banyak diteliti namun memiliki implikasi terhadap validitas maupun keadilan.

Mengevaluasi Potensi Bias dalam Mesin Penilaian

Bias dalam pengujian secara tradisional dikonseptualisasikan sebagai varians yang tidak relevan dengan konstruk (*construct-irrelevant variance*) yaitu kesalahan sistematis yang secara tidak adil menguntungkan atau merugikan kelompok tertentu, terlepas dari konstruk yang sebenarnya hendak diukur. Dalam penilaian yang diperkuat oleh AI, sumber potensi bias menjadi jauh lebih luas. Mesin penilaian dan generator butir soal yang dilatih menggunakan korpus berskala besar dapat mewarisi bias sosial yang terkandung dalam data tersebut, sehingga melanggengkan kesenjangan yang sulit dideteksi tanpa perancangan dan evaluasi yang cermat. Dorsey dan Michaels secara tepat mengemukakan bahwa algoritma penilaian berbasis AI dapat melanggengkan bias manusia, sekaligus memunculkan bias unik yang bersumber dari algoritma itu sendiri atau dari data pelatihannya.

Penekanan Cronbach dan Meehl pada jaringan nomologis yakni jalinan hubungan teoretis dan empiris antar-konstruk memberikan landasan untuk mendeteksi adanya fungsi yang bias. Ketika pola penilaian atau prediksi yang teramati menyimpang dari hubungan yang diharapkan di antara berbagai kelompok, hal ini dapat mengindikasikan bahwa skor tersebut tercemar oleh fitur-fitur yang tidak relevan. Hal ini diformalkan lebih lanjut dalam praktik modern melalui perangkat statistik seperti analisis fungsi butir soal diferensial (*differential item functioning*) dan analisis prediksi diferensial; keduanya direkomendasikan oleh standar AERA dan rekan-rekan sebagai elemen krusial dalam mengevaluasi keadilan.

Untuk sistem berbasis AI, evaluasi bias tidak boleh hanya terbatas pada hasil akhir (*output*), tetapi harus mencakup seluruh alur perancangan: data pelatihan, rekayasa *prompt*, dan mekanisme umpan balik yang membentuk perilaku model. Metode seperti pengujian keadilan kontrafaktual (*counterfactual fairness testing*) dan studi audit interseksional kini semakin banyak digunakan untuk mengevaluasi apakah model penilaian secara sistematis menunjukkan kinerja yang lebih rendah bagi populasi tertentu atau merepresentasikan konstruk yang dituju secara keliru dengan cara yang berkorelasi dengan ras, gender, bahasa, atau atribut demografis lainnya.

Pada akhirnya, evaluasi terhadap mesin penilaian maupun bias harus dikaitkan dengan teori validitas yang memandangnya sebagai tanggung jawab sosial dan etis. Validitas bukan

sekadar masalah koherensi teknis internal, melainkan juga menyangkut konsekuensi penggunaan tes bagi individu maupun kelompok. Seiring dengan semakin meluasnya penggunaan penilaian berbasis algoritma, tanggung jawab evaluatif ini pun turut berkembang—menuntut para ahli psikometri untuk mendalami baik teori dasar maupun metode komputasi yang sedang berkembang.

13.5 KOMPLEKSITAS VALIDASI TES AKIBAT KECERDASAN BUATAN

Dalam bab ini, kita mengawali pembahasan dengan penjelasan mengenai konsep-konsep dasar validitas, seperti penetapan tujuan tes dan konstruk yang akan diuji, serta pengumpulan bukti untuk memastikan bahwa tes berjalan sesuai rencana dan menghasilkan luaran yang diharapkan. Selanjutnya, kami menguraikan bagaimana penerapan AI dalam penilaian menuntut kita untuk berpikir lebih luas mengenai cara mempertahankan kesetiaan terhadap konsep-konsep dasar tersebut, terutama ketika AI membuat aspek-aspek pengembangan, pelaksanaan, pemberian skor, atau interpretasi skor tes menjadi kurang transparan. Hal-hal yang memerlukan perhatian lebih dalam proses validasi mencakup cara pembuatan konten dan sejauh mana konten tersebut merepresentasikan konstruk yang dituju, bagaimana skor atau informasi lain dihasilkan dari interaksi peserta tes dengan tes tersebut, serta penyelidikan terhadap potensi bias yang bersumber dari algoritma *machine learning* (ML) dan data yang digunakan untuk melatihnya.

Saat kita meninjau dampak AI terhadap penilaian, satu kesimpulan menjadi jelas—validasi menjadi semakin penting dan akan membutuhkan upaya lebih besar dibandingkan dengan penilaian tradisional. Tes tradisional kerap dikritik karena kurangnya transparansi terkait proses pengubahan respons mahasiswa menjadi skor skala. Di era AI ini, input dan output tes akan semakin bergantung pada algoritma dan model yang mungkin tidak dipahami dengan baik oleh sebagian besar pemangku kepentingan dalam proses pengujian. Oleh karena itu, memvalidasi relevansi konstruk dan keadilan dari pendekatan-pendekatan ini beserta hasilnya menjadi aspek tambahan yang penting dalam validasi tes.

BAB 14
POTENSI TRANSFORMATIF AI DALAM ASESMEN SKALA BESAR

14.1 MASA DEPAN ASESMEN SKALA BESAR DI ERA AI

Kecerdasan buatan (AI) secara pesat mengubah lanskap asesmen pendidikan. Sistem asesmen yang dulunya hanya bergantung pada keahlian manusia kini semakin didukung atau bahkan digerakkan oleh algoritma cerdas yang mampu melakukan penilaian (skoring), pembuatan, klasifikasi, dan penyusunan konten tes. Kemajuan ini telah merambah ke asesmen skala besar, termasuk program asesmen skala besar internasional (ILSA) seperti *Trends in International Mathematics and Science Study* (TIMSS), *Progress in International Reading Literacy Study* (PIRLS), dan *Programme for International Student Assessment* (PISA), serta asesmen nasional seperti *National Assessment of Educational Progress* (NAEP). Mengingat peran penting program-program ini dalam memantau kemajuan pendidikan, memandu kebijakan, dan menjadi dasar pengambilan keputusan di tingkat sistem, peran AI yang semakin sentral di dalamnya membawa potensi besar sekaligus tanggung jawab yang mendalam.

Selama satu dekade terakhir, teknologi AI telah berkembang dari alat eksperimental menjadi komponen penting dalam sistem asesmen. Inovasi dalam *machine learning* (ML) dan *natural language processing* (NLP) kini memungkinkan penilaian otomatis terhadap jawaban terbuka, pembuatan butir soal yang selaras dengan kerangka kurikulum, serta penyusunan tes adaptif yang disesuaikan dengan kemampuan mahasiswa. Dalam ILSA, di mana keadilan lintas bahasa dan lintas budaya sangatlah krusial, *large language models* (LLM) membuka peluang baru untuk penilaian otomatis lintas bahasa, pembuatan konten yang sesuai secara budaya, serta penyelarasan dengan kerangka kurikulum dan internasional. Inovasi-inovasi ini mendukung pengembangan dan pelaksanaan asesmen yang tidak hanya kokoh secara psikometris tetapi juga berkelanjutan secara operasional dalam skala global.

Terlepas dari kemajuan-kemajuan tersebut, penggunaan AI dalam penilaian bukannya tanpa risiko. Kekhawatiran utama mencakup privasi, bias algoritmik, kurangnya transparansi, keadilan budaya dan linguistik, serta potensi melemahnya argumen validitas ketika model *black box* (kotak hitam) diadopsi tanpa dokumentasi atau kemampuan interpretasi yang memadai. Sebagaimana ditekankan oleh para peneliti, AI dapat melanggengkan ketimpangan historis jika desain dan penerapannya tidak diperiksa secara cermat melalui pemantauan bias yang berkelanjutan, analisis pergeseran skala (*scale drift*), dan studi validasi.

Karena sistem AI beroperasi secara berbeda dari model psikometrik tradisional, penerapannya dapat mengubah beberapa elemen argumen validitas yang komprehensif. Penilaian otomatis dapat memunculkan aturan keputusan hasil pembelajaran mesin yang mungkin tidak selaras dengan konstruk yang dituju, sehingga menimbulkan pertanyaan mengenai struktur internal dan generalisabilitas. Pembuatan butir soal berbasis AI berpotensi mengubah cakupan domain konstruk, yang berdampak pada validitas isi dan representasi konstruk. Demikian pula, alat penyelarasan berbasis *embedding* bergantung pada kemiripan semantik dan memerlukan evaluasi cermat untuk menentukan kapan pendekatan ini secara

bermakna mencerminkan konstruk yang mendasarinya. Sepanjang bab ini, kami menyoroti interaksi tersebut dan mengidentifikasi di mana metodologi yang ada telah memberikan panduan yang memadai serta di mana diperlukan perluasan praktik saat ini yang spesifik untuk AI. Untuk memastikan penerapan yang bertanggung jawab, organisasi-organisasi terkemuka menganjurkan penyelarasan sistem penilaian berbasis AI dengan standar pengukuran yang ada, seperti *Standards for Educational and Psychological Testing* (AERA, APA, dan NCME,), sembari mengembangkan panduan baru yang khusus menangani kemampuan dan risiko unik AI. Rekomendasi yang sering muncul dalam riset adalah penerapan pendekatan *human-in-the-loop* (melibatkan manusia dalam proses sistem), di mana ahli psikometri, pendidik, dan pakar bidang terkait berkolaborasi dengan AI untuk menjamin kualitas, keadilan, dan relevansi kontekstual. Selain itu, masalah privasi harus ditangani saat menggunakan layanan berbasis *cloud*, misalnya dalam penilaian jawaban tertulis mahasiswa yang berpotensi memuat informasi yang dapat mengidentifikasi individu.

Bab ini ditujukan bagi ahli psikometri, pengembang penilaian, pembuat kebijakan, dan peneliti yang sedang mempertimbangkan atau menerapkan proses berbasis AI dalam program penilaian berskala besar. Kami berfokus pada empat kasus penggunaan yang paling berdampak signifikan bagi ILSA (penilaian berskala besar internasional): penilaian otomatis, pembuatan butir soal, perakitan tes, dan penyelarasan butir soal, dengan mengacu pada pembelajaran terkini, implementasi percontohan, dan kerangka kerja etis. Tujuan kami bukanlah menyajikan daftar lengkap aplikasi AI dalam pendidikan, melainkan mengkaji aplikasi yang secara langsung mengubah ekosistem penilaian berskala besar serta mengidentifikasi peluang inovasi yang ditawarkannya maupun batasan-batasan yang harus dipatuhi. Pada akhirnya, bab ini mengusung pendekatan yang memanfaatkan AI untuk memperluas kemampuan manusia sembari tetap mempertahankan peran manusia dalam pengambilan keputusan pada tahapan-tahapan krusial serta menjunjung tinggi prinsip kesetaraan dan keterinterpretasian yang menjadi landasan bagi praktik penilaian yang bertanggung jawab.

14.2 KASUS PENGGUNAAN PENTING AI DALAM ASESMEN BERSKALA BESAR

Integrasi AI ke dalam program asesmen berskala besar paling terlihat pada empat ranah utama: penilaian otomatis, pembuatan butir soal, verifikasi dan penyelarasan butir soal, serta perakitan tes. Kasus-kasus penggunaan ini mewakili fungsi-fungsi inti di mana teknologi AI saat ini sedang diteliti, diuji coba, dan diterapkan. Setiap aplikasi berupaya mengatasi tantangan lama dalam asesmen seperti skalabilitas, konsistensi, ketepatan waktu, dan kebutuhan sumber daya yang besar sembari menghadirkan kemampuan-kemampuan baru.

Penilaian Otomatis

Penilaian otomatis telah digunakan dalam pengukuran pendidikan selama beberapa dekade, dengan memanfaatkan metode NLP (*Natural Language Processing*) dan linguistik komputasi, serta model statistik, untuk mengevaluasi respons peserta tes terhadap pertanyaan terbuka (*open-ended questions*). Ini merupakan salah satu aplikasi ML (*Machine Learning*) dan AI yang paling matang dalam pengukuran pendidikan di berbagai konteks, termasuk asesmen K–12, lisensi, dan program sertifikasi, khususnya untuk mengevaluasi

kemahiran dalam menulis, membaca, dan matematika. Penerapannya sangat krusial dalam asesmen berskala besar, di mana penilaian respons terbuka dalam jumlah masif menghadirkan tantangan logistik, finansial, dan psikometrik yang terus-menerus muncul. Penilaian oleh manusia terhadap butir soal respons konstruksi (*constructed-response* atau CR) tidak hanya mahal dan memakan banyak tenaga, tetapi juga rentan terhadap ketidakkonsistenan akibat variabilitas penilai, kelelahan, atau kesalahan interpretasi lintas budaya.

Tantangan-tantangan ini menjadi lebih berat dalam asesmen multibahasa, di mana respons dari berbagai negara yang sering kali melibatkan 100 versi bahasa atau lebih harus dinilai secara andal dan adil. Sistem penilaian otomatis telah berkembang secara signifikan selama dua dekade terakhir. Sistem-sistem awal, termasuk *Project Essay Grader*, *e-rater*, dan *Intelligent Essay Assessor*, mengandalkan fitur-fitur yang dirancang oleh pakar, seperti tata bahasa, sintaksis, kosakata, dan koherensi. Sistem-sistem ini terutama dikembangkan untuk menilai esai dalam asesmen monolingual yang memiliki konsekuensi tinggi (*high-stakes assessments*).

Perkembangan terkini telah mengintegrasikan ML dan NLP untuk melampaui batasan rangkaian fitur yang telah ditetapkan sebelumnya, sehingga memungkinkan model mempelajari pola penilaian dari korpus besar berisi respons yang telah dinilai oleh manusia. Seiring dengan munculnya teknik *deep learning* dan model bahasa berbasis *transformer*, kemampuan AI dalam memahami dan mengevaluasi respons Mahasiswa yang beragam dan kompleks telah meningkat secara signifikan. Meskipun kemajuan teknis ini sangat berarti, hal tersebut juga memunculkan pertimbangan penting terkait validitas. Sistem penilaian otomatis tradisional yang berbasis pada fitur rancangan manual menawarkan transparansi dalam proses pengambilan keputusan penilaian—penilai manusia dapat melacak fitur linguistik tertentu untuk memahami dasar pemberian skor.

Pendekatan jaringan saraf (*neural network*) modern, khususnya model berbasis *deep learning*, beroperasi layaknya "kotak hitam" (*black box*); hal ini menyulitkan upaya pembuktian representasi konstruk untuk memastikan bahwa model menilai konstruk yang dituju, dan bukan sekadar fitur permukaan atau variabel pengganggu lainnya. Pergeseran ini menuntut protokol validasi yang lebih ketat, termasuk perhatian cermat terhadap perbedaan kinerja antar-kelompok demografis serta pemeriksaan mengenai fitur linguistik dan semantik apa yang mendasari keputusan penilaian tersebut. Dalam konteks ILSA, Jung dan rekan-rekan melakukan studi dasar yang menunjukkan bahwa jaringan saraf tiruan (*artificial neural networks* atau ANN) dapat memberikan skor secara efektif pada butir soal jawaban konstruksi (*constructed-response* atau CR) singkat dalam tes matematika dan sains TIMSS menggunakan data dari Amerika Serikat. Pendekatan mereka berfokus pada empat butir soal dengan sistem penilaian dikotomis; hasilnya menunjukkan bahwa ANN yang dilatih menggunakan data berlabel manusia mencapai tingkat kesepakatan rata-rata yang tinggi dengan penilai manusia untuk keempat butir soal tersebut (r = 0,91).

Selain itu, mereka memperkenalkan metode untuk meningkatkan kualitas penilaian dengan memberikan bobot pada data pelatihan berdasarkan skor estimasi teori respons butir (*item response theory* atau IRT), sehingga mampu menyaring label yang mengandung derau

(*noisy labels*) dan mengurangi varians model. Pendekatan ini menunjukkan potensi tidak hanya dalam mempertahankan akurasi penilaian, tetapi juga dalam meningkatkan reliabilitas psikometrik.

Melanjutkan hasil tersebut, Jung dan rekan rekan. membahas tantangan utama dalam ILSA, yaitu penilaian respons multibahasa. Alih-alih mengembangkan model terpisah untuk setiap bahasa—praktik yang memakan biaya besar dan sering kali tidak praktis mengingat skala operasional ILSA—para penulis mengusulkan alur kerja (*pipeline*) yang menggabungkan penerjemahan mesin (*machine translation* atau MT) dengan penilaian berbasis ANN. Dengan menggunakan respons dalam enam bahasa dari delapan negara yang mencakup aksara Latin maupun non-Latin—respons tersebut diterjemahkan ke dalam bahasa Inggris menggunakan Google Translate dan ChatGPT, kemudian dinilai menggunakan model *bag-of-words* yang dikombinasikan dengan jaringan saraf *feed-forward* yang terhubung sepenuhnya (*fully connected*).

Mengembangkan lebih lanjut lini pembelajaran ini, Jung dan rekan-rekan menggunakan data dari tiga butir soal sains yang mencakup seluruh negara peserta (52 negara dan 42 bahasa) dalam TIMSS 2023, serta membandingkan berbagai metode penilaian otomatis, termasuk model *supervised* (terawasi) dan *unsupervised* (tidak terawasi/LLM). Pendekatan MT ganda yang dikombinasikan dengan ANN (MMT_ANNs) mencapai kinerja tertinggi di berbagai negara dan butir soal, dengan rata-rata kesepakatan eksak (*exact agreement*) berkisar antara 92,5% hingga 95% dan nilai Cohen's kappa antara 0,83 hingga 0,87. Metode MMT_ANNs secara strategis memilih antara Google Translate atau ChatGPT untuk bahasa dan respons individu yang berbeda, sehingga menunjukkan efektivitas bagi bahasa-bahasa dengan sumber daya terbatas (*low-resource languages*). Yang patut dicatat, pendekatan ini berhasil menerjemahkan lebih dari 99% respons dengan rata-rata hanya 0,81% yang tidak diterjemahkan serta mencapai selisih rata-rata terstandarisasi sebesar 0, yang menunjukkan bahwa skor dari penilai manusia dan sistem otomatis hampir identik.

Analisis terhadap ketidaksesuaian penilaian mengungkapkan bahwa perbedaan tersebut terutama berasal dari tiga faktor: ketidakkonsistenan penilaian manusia baik antarnegara maupun di dalam negara yang sama, kesulitan penerjemahan pada respons dengan banyak kesalahan ejaan, serta tantangan terkait bahasa dengan sumber daya terbatas (*low-resource languages*). Studi ini mencatat pola-pola penting: di satu negara, penilai manusia menerapkan kriteria penilaian secara lebih ketat dibandingkan negara lain menganggap salah respons yang menggunakan terminologi lama atau penjelasan kata kunci secara tidak langsung sementara sistem otomatis secara konsisten mengklasifikasikannya sebagai jawaban benar berdasarkan kesesuaian konten dengan rubrik.

Demikian pula, respons dengan kesalahan ejaan yang sama dalam bahasa yang digunakan di berbagai negara dinilai secara berbeda di negara-negara tersebut; penilai di satu negara menganggapnya salah, sedangkan penilai di negara lain menerimanya, sementara sistem otomatis memberikan penilaian yang konsisten dalam setiap kasus. Temuan ini menyoroti tantangan mendasar di bidang ini: ketika penilaian manusia menunjukkan variasi antarnegara, penetapan tolok ukur validitas yang tepat menjadi hal yang kompleks. Yang tak

kalah penting, hasil ini mengindikasikan bahwa penilaian otomatis dapat berfungsi sebagai langkah pengendalian mutu (QC) yang efektif dan hemat biaya dalam studi penilaian skala besar internasional (ILSA), yang berpotensi mengurangi ketergantungan pada penilai manusia tambahan sekaligus menjaga konsistensi penilaian di berbagai konteks linguistik yang beragam.

Untuk butir soal respons konstruksi (CR) berbasis gambar, von Davier dan rekan-rekan dan Tyackdan rekan-rekan, menunjukkan potensi besar *convolutional neural networks* (CNN) dalam pemberian skor pada butir soal grafis TIMSS. Para penulis menunjukkan bahwa CNN mengungguli *feed-forward neural networks* secara signifikan pada butir soal grafis TIMSS 2019 yang mengharuskan mahasiswa menggambar pada kisi digital; CNN mencapai akurasi klasifikasi hingga 97,53% dan, dalam beberapa kasus, berhasil mengidentifikasi dengan benar respons yang sebelumnya salah diklasifikasikan oleh penilai manusia. Dengan memperluas pembelajaran ini ke delapan butir soal respons grafis, Tyack dan rekan rekan melaporkan tingkat akurasi di atas 98% pada berbagai tugas matematika dan tingkat kompleksitas yang beragam. Yang penting, analisis kesalahan berbasis model mengungkapkan bahwa hingga 2% respons dinilai secara keliru atau tidak konsisten oleh manusia; hal ini menggambarkan keterbatasan penilaian manusia sebagai standar acuan sekaligus nilai sistem penilaian otomatis (AS) untuk pengendalian mutu (QC).

Secara bersama-sama, studi-studi ini membuka jalan bagi penerapan prosedur QC otomatis selama proses penilaian butir soal CR TIMSS 2023 guna meningkatkan konsistensi, transparansi, dan reliabilitas penilaian dalam skala global. Secara paralel, upaya saat ini sedang memperluas inovasi tersebut ke jenis respons yang lebih kompleks, melampaui sekadar butir soal CR jawaban singkat dan butir soal grafis. Sebuah alur kerja penilaian yang sepenuhnya tanpa pengawasan (*fully unsupervised*) telah dikembangkan untuk menilai ulang butir soal CR PIRLS 2021 tanpa memerlukan data pelatihan yang diberi label oleh manusia.

Pendekatan ini memanfaatkan *Large Language Models* (LLM), penerjemahan mesin (MT), teknik rekayasa *prompt*, dan metrik kemiripan semantik untuk menetapkan serta memvalidasi skor. Perkembangan ini menandai pergeseran signifikan menuju solusi penilaian berbasis AI yang dapat diskalakan, mendukung multibahasa, dan beroperasi tanpa pengawasan dalam studi penilaian skala besar internasional (ILSA); hal ini sekaligus menantang paradigma penilaian tradisional yang berpusat pada manusia dan memperluas cakupan AS untuk mencakup respons mahasiswa yang lebih bernuansa. Kemajuan teknis ini juga menuntut pertimbangan cermat mengenai validitas inferensi yang ditarik dari skor hasil AI. Meskipun kesesuaian dengan penilaian manusia sering kali dianggap sebagai indikator utama kinerja model, proses penilaian oleh manusia dan sistem otomatis memiliki perbedaan yang mendasar.

Kita berasumsi bahwa penilai manusia yang terlatih dengan baik mengerahkan upaya terbaik mereka untuk memahami konten dan membuat penilaian yang berdasar, sementara sistem AI mengidentifikasi pola sistematis yang berkorelasi dengan indikator konstruk. Ketika tugas penilaian dirancang untuk penilaian oleh manusia, perbedaan ini dapat menciptakan kesenjangan inferensial jika fitur-fitur yang ditangkap oleh AI tidak selaras dengan bukti yang

relevan dengan konstruk. Perspektif ECD membantu memitigasi risiko ini dengan memastikan koherensi antara model konstruk, model tugas, dan model bukti, sehingga memastikan bahwa penilaian otomatis menyasar fitur-fitur yang bermakna secara teoretis. Kerangka kerja validitas tradisional perlu ditinjau kembali berdasarkan lima sumber bukti: konten, proses respons, struktur internal, generalisasi, dan validitas konsekuensial.

Selain pendekatan-pendekatan tersebut, kerangka kerja validasi berbasis argumen telah mendapatkan perhatian luas dalam pembelajaran penilaian otomatis, dengan menyediakan kriteria operasional yang menghubungkan sistem penilaian otomatis (AS) dengan definisi konstruk serta memandu standar evaluasi, termasuk aspek keterkaitan dengan skor manusia, keadilan, hubungan dengan variabel eksternal, reliabilitas lintas tugas dan bentuk tes, serta dampak penggunaannya. Berbagai kerangka kerja ini menekankan perlunya bukti validitas yang komprehensif, alih-alih hanya mengandalkan statistik kesepakatan antara manusia dan mesin. Untuk memenuhi persyaratan validitas ini, kami mengusulkan penerapan operasional penilaian berbasis LLM untuk ILSA yang mengikuti alur kerja terintegrasi dengan pengendalian mutu (QC). Pra-pemrosesan berfungsi sebagai lapisan perlindungan pertama, memastikan kepatuhan etis melalui deteksi dan penyembunyian informasi identitas pribadi (PII) lintas bahasa sebelum respons masuk ke sistem berbasis cloud.

Pada saat yang sama, pemeriksaan validasi bahasa memastikan kualitas data dan menjawab kekhawatiran validitas konsekuensial terkait privasi Mahasiswa serta penanganan data yang tepat. Tahap penyusunan *prompt* penilaian secara langsung menerapkan prinsip-prinsip ECD: *prompt* mencakup pokok soal (*item stem*), panduan penilaian lengkap, contoh respons pada berbagai tingkat skor, dan instruksi penilaian yang eksplisit, sehingga proses penilaian AI secara konseptual serupa dengan cara penilai manusia dilatih untuk menerapkan kriteria penilaian. Selama penilaian berlangsung, jejak audit yang lengkap mendokumentasikan semua interaksi API dan keluaran model, guna memenuhi persyaratan transparansi yang penting bagi bukti validitas terkait konten. Pemeriksaan kualitas pasca-pemrosesan menangani masalah keadilan dan struktur internal dengan meninjau kualitas terjemahan untuk mendeteksi halusinasi yang dapat menimbulkan varians yang tidak relevan dengan konstruk. Selain itu, pemeriksaan konsistensi memastikan respons yang secara leksikal serupa mendapatkan skor yang sama, sehingga mengurangi variasi penilaian yang arbitrer.

Terakhir, analisis reliabilitas lintas negara dan lintas bahasa memberikan bukti untuk generalisasi, dengan penandaan sistematis terhadap respons yang menunjukkan ketidakkonsistenan antarnegara atau antarbahasa, yang memicu pemantauan ahli ketika penilaian otomatis mungkin menangkap fitur-fitur yang tidak dimaksudkan. Proses bertahap ini menghasilkan dokumentasi dan langkah pengendalian mutu yang diperlukan untuk mendukung argumen validitas yang dapat dipertanggungjawabkan bagi penilaian berbasis AI dalam ILSA. Inovasi-inovasi ini menawarkan manfaat substansial bagi program penilaian berskala besar. Pertama, inovasi ini mengurangi beban operasional penilaian oleh manusia, yang sering kali melibatkan perekrutan dan pelatihan ratusan penilai di berbagai negara peserta. Pelatihan penilaian mencakup sekitar 25% dari biaya pertemuan internasional yang diperlukan untuk melaksanakan penilaian internasional, di luar biaya besar untuk

mempekerjakan penilai manusia di setiap negara peserta. Kedua, inovasi ini memungkinkan waktu penyelesaian yang lebih cepat dibandingkan dengan kerangka kerja penilaian yang sepenuhnya dilakukan oleh manusia. Ketiga, inovasi ini membantu memastikan komparabilitas penilaian, sebuah syarat penting untuk inferensi lintas negara yang valid dalam ILSA. Singkatnya, penilaian otomatis dalam ILSA bukan lagi tujuan yang jauh di masa depan, melainkan solusi praktis yang semakin terbukti validitasnya. Dengan mengintegrasikan MT, ANN, LLM, dan kerangka kerja QC yang selaras dengan domain, ILSA dapat beralih menuju sistem yang lebih skalabel, adil, dan kaya data, sembari tetap mempertahankan prinsip-prinsip pengukuran inti.

Pembuatan Butir Soal Otomatis

Pengembangan butir soal merupakan salah satu aspek yang paling memakan biaya, waktu, dan tenaga dalam pengembangan asesmen karena sangat bergantung pada pakar materi pelajaran (*subject-matter experts* atau SME). Para SME menyusun draf setiap butir soal, yang kemudian diikuti oleh serangkaian pemantauan dan revisi berulang untuk memastikan kesesuaian dengan standar kualitas yang telah ditetapkan. Namun, pendekatan manual ini menimbulkan ketidakkonsistenan akibat perbedaan keahlian dan interpretasi antar-SME, serta kurang memiliki skalabilitas dan efisiensi operasional, sehingga membatasi responsivitas terhadap kebutuhan asesmen yang terus berkembang.

Untuk mengatasi keterbatasan tersebut, pembuatan butir soal otomatis atau *Automated Item Generation* (AIG) muncul sebagai alternatif yang menjanjikan. AIG, yang pada awalnya didefinisikan sebagai penggunaan teori kognitif dan psikometrik yang didukung oleh teknologi komputer untuk menghasilkan butir soal tes secara sistematis, menawarkan solusi berskala besar yang mengurangi beban kerja manusia sembari tetap menjaga keselarasan konstruk. Implementasi awal AIG terutama mengandalkan templat butir soal, yaitu format butir soal terstruktur yang telah ditetapkan sebelumnya oleh SME untuk menghasilkan pertanyaan. Meskipun efektif untuk menghasilkan sejumlah besar butir soal, pendekatan ini terbatas pada jenis butir soal yang definisinya sempit dan kurang fleksibel untuk mendukung beragam format butir soal. Selain itu, tahap awal pembuatan templat tersebut masih memerlukan masukan yang signifikan dari SME, yang pada akhirnya membatasi skalabilitas pendekatan ini dalam lingkungan operasional. Strategi alternatif memanfaatkan teknik NLP tradisional seperti penandaan kelas kata (*part-of-speech tagging*), pemodelan topik, dan ekstraksi frasa nomina untuk menghasilkan butir soal secara otomatis.

Inovasi-inovasi ini mengatasi beberapa kendala AIG berbasis templat, namun strukturnya yang berbasis aturan sering kali gagal menghasilkan butir soal dengan kekayaan semantik atau keselarasan yang memadai dengan target kognitif. Terobosan yang lebih signifikan muncul dengan adanya model jaringan saraf tiruan (*neural network*), khususnya jaringan saraf berulang (*recurrent neural network*) berbasis LSTM, yang memungkinkan sistem untuk belajar secara langsung dari korpus teks dan mensimulasikan pembuatan butir soal layaknya manusia. Kemunculan model bahasa berbasis *transformer* menandai lompatan besar dalam kemampuan AIG. Model-model seperti *Bidirectional Encoder Representations from Transformers*, *Text-to-Text Transformer*, dan *Generative Pretrained Transformer* sering kali

mencapai kinerja setara manusia dalam berbagai tugas kebahasaan. Sebagai contoh, GPT-2 digunakan untuk menyusun materi ujian lisensi medis, membuat pertanyaan kepribadian , dan menghasilkan pilihan jawaban pengecoh (*distractor*); namun, model ini memerlukan proses pra-pelatihan dan penyesuaian halus (*fine-tuning*) yang ekstensif pada kumpulan data besar yang spesifik untuk suatu domain agar dapat beradaptasi secara efektif dengan kebutuhan AIG. Seiring dengan pengembangan GPT-3, pembelajaran AIG memasuki fase baru. GPT-3 memperkenalkan kemampuan pembelajaran *few-shot* dan *zero-shot*, yang memungkinkan pengguna menghasilkan keluaran yang koheren dan sesuai konteks dengan masukan minimal melalui teknik *prompt engineering*. Kemajuan ini memungkinkan pergeseran dari ketergantungan pada kumpulan data pelatihan besar yang telah diberi label (*labeled*) menuju pembuatan butir soal yang fleksibel dan dapat ditingkatkan skalanya (*scalable*) menggunakan instruksi bahasa alami.

Studi-studi awal menggunakan GPT-3 untuk menyusun teks pemahaman bacaan baik yang bersifat ekspositori maupun naratif serta membuat butir soal pemahaman dan pilihan pengecoh untuk pertanyaan pilihan ganda pada *Duolingo English Test*, dengan menerapkan pendekatan "AI yang melibatkan manusia" (*human-in-the-loop AI*) dalam alur kerja mereka untuk mengevaluasi keadilan dan potensi bias pada teks serta butir soal yang dihasilkan secara otomatis tersebut. Bezirhan dan von Davier melakukan salah satu studi empiris pertama yang secara sistematis mengevaluasi penggunaan LLM untuk menghasilkan materi penilaian membaca. Mereka memperkenalkan kerangka kerja rekayasa *prompt* terstruktur yang memandu GPT-3 untuk menghasilkan teks bacaan dengan tingkat keterbacaan (skor Lexile) yang terkontrol, yang sesuai untuk penilaian membaca dan dimodelkan berdasarkan teks PIRLS yang sudah ada. Mereka juga menerapkan pendekatan *human-in-the-loop* (melibatkan manusia dalam proses) ke dalam kerangka kerja pembuatan cerita tersebut untuk memungkinkan penyuntingan akhir yang ringan serta memitigasi potensi bias dan ketidakakuratan faktual. Setelah pemeriksaan tersebut, panel penilai mengevaluasi setiap teks buatan AI bersamaan dengan teks asli PIRLS yang setara menggunakan skala Likert empat poin.

Dalam seluruh perbandingan, 92%–97% penilai menganggap teks buatan AI tersebut memadai bagi pembaca kelas empat. Temuan ini memberikan bukti yang menjanjikan terkait validitas tampakan (*face validity*) dan kelayakan konten; namun, penilaian kelayakan saja tidak menunjukkan validitas konstruk secara utuh. Hal ini perlu dilengkapi dengan uji coba lapangan bersama mahasiswa dan validasi psikometrik di luar penilaian kelayakan oleh ahli, serupa dengan proses validasi untuk materi yang ditulis manusia.

Selain itu, tingkat keterlibatan (*engagement*) untuk teks buatan AI setara dengan—atau dalam beberapa kasus melampaui—teks asli, sementara skor koherensi tetap sebanding. Yang menarik, para penilai tidak menganggap teks buatan AI mengandung lebih banyak materi yang mengalihkan perhatian dibandingkan teks buatan manusia; bahkan, dalam satu perbandingan karya sastra, mereka menilai teks buatan AI mengandung lebih sedikit konten yang mengalihkan perhatian. Temuan ini mengindikasikan bahwa, jika dipadukan dengan desain *prompt* yang efektif dan pengawasan manusia yang terarah, LLM dapat menghasilkan teks bacaan utuh yang sesuai secara psikolinguistik dan menyaingi kualitas materi buatan manusia,

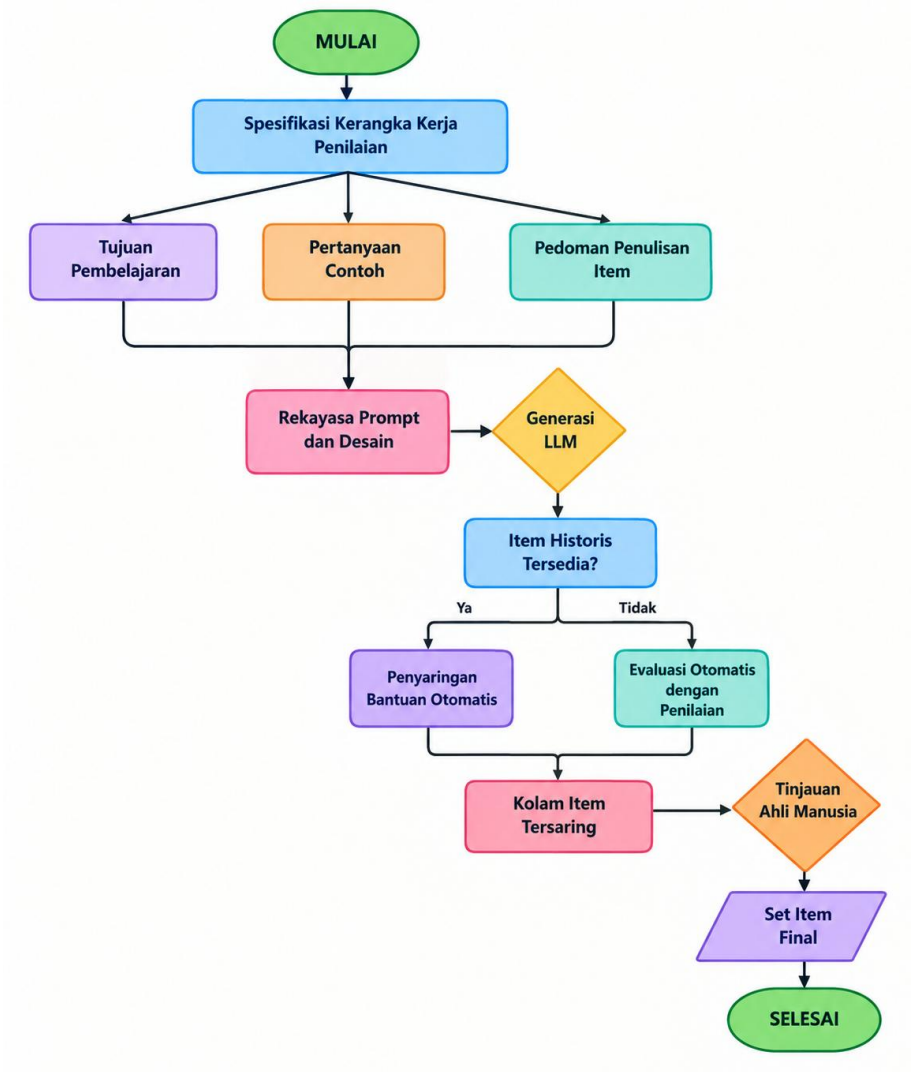
sehingga menawarkan alternatif yang dapat diskalakan dan valid untuk pengembangan konten dalam penilaian pendidikan. Studi-studi selanjutnya memperluas paradigma ini ke ranah lain dengan menggunakan GPT-3.5 dan GPT-4. Sebagai contoh, LLM telah digunakan untuk menghasilkan konten dalam bidang pemahaman bacaan, ilmu kedokteran, pemrograman tingkat sarjana, dan matematika analisis psikometrik menunjukkan bahwa 40%–70% butir soal yang dihasilkan AI memenuhi tingkat kesulitan dan ambang batas daya pembeda yang konvensional. Memahami alasan mengapa 30%–60% butir soal sisanya gagal memenuhi ambang batas tersebut baik karena masalah sistematis dalam kerangka kerja rekayasa *prompt*, keterbatasan inheren model, maupun kebutuhan akan mekanisme penyaringan yang lebih canggih—merupakan arah penting untuk meningkatkan kualitas dan efisiensi AIG (*AI-generated content/items*).

Perkembangan ini mencerminkan pergeseran yang lebih luas dari pendekatan berbasis templat menuju alur kerja yang fleksibel dan digerakkan oleh *prompt*. Rekayasa *prompt* telah muncul sebagai komponen krusial dalam perancangan kerangka kerja AIG yang efektif dan bertanggung jawab. Meskipun secara teknis kini dimungkinkan untuk memberikan instruksi (*prompt*) kepada *chatbot* (seperti ChatGPT) guna menghasilkan butir soal, melakukannya tanpa adanya kerangka panduan (*scaffolding*) sering kali menghasilkan konten yang kualitasnya setara dengan buatan penulis soal pemula yakni berkualitas rendah, tidak selaras dengan target kognitif, dan tidak konsisten dengan kerangka kerja penilaian yang telah ditetapkan. Oleh karena itu, pengembangan alur kerja yang menjamin keselarasan konstruk dan kerangka kerja dalam AIG menjadi hal yang sangat penting.

Dalam lingkungan pendidikan, butir soal penilaian harus mematuhi cetak biru ketat yang menetapkan tingkat kelas, domain konten, dan proses kognitif. Sebagai contoh, dalam penilaian membaca PIRLS, pertanyaan harus bergantung pada teks, tidak tumpang tindih dalam hal inferensi, bersifat independen antarbutir soal, serta dipetakan pada batasan proses pemahaman yang tidak dapat dipenuhi hanya dengan *prompt* umum. Dengan demikian, rekayasa *prompt* harus menerjemahkan spesifikasi desain yang kompleks ini menjadi instruksi bahasa alami yang presisi. Para peneliti telah mengeksplorasi berbagai strategi untuk meningkatkan fidelitas *prompt*, seperti menyertakan tujuan pembelajaran yang terstruktur (misalnya, taksonomi Bloom), menentukan format keluaran, dan menggunakan contoh *few-shot* untuk memandu pembuatan respons.

Selain itu, penggunaan *prompt* berbasis peran di mana model diinstruksikan untuk bertindak sebagai ahli materi pelajaran (SME) atau penulis butir soal telah terbukti meningkatkan relevansi maupun kualitas. Sebuah studi terkini oleh Bezirhan dan rekan-rekan memperkenalkan kerangka kerja terstruktur untuk menghasilkan butir soal pemahaman bacaan yang selaras dengan asesmen PIRLS menggunakan LLM. Pembelajaran ini mengembangkan riset rekayasa *prompt* sebelumnya dengan menyematkan pedoman penulisan butir soal PIRLS secara langsung ke dalam *prompt* serta mengoperasionalkan cetak biru asesmen sebagai input bahasa alami untuk pembuatan butir soal berbasis GPT. Proses tersebut mengikuti alur kerja AIG (*AI-Generated*) bertahap, yang dimulai dengan memasukkan teks stimulus, teks referensi, dan contoh butir soal.

Templat *prompt* dirancang untuk mencerminkan prinsip-prinsip utama PIRLS seperti target pemahaman yang spesifik terhadap proses dan independensi butir soal serta digunakan untuk menghasilkan butir soal kandidat awal melalui GPT-4. Butir-butir soal tersebut kemudian melalui tahap penyaringan menggunakan serangkaian metrik kemiripan semantik dan leksikal (BLEU, ROUGE-L, METEOR, dan kemiripan kosinus melalui *embedding* BERT) guna mengidentifikasi butir soal yang paling merepresentasikan isi teks dan proses pemahaman yang dituju. Himpunan bagian yang tersisa kemudian ditinjau oleh pakar manusia; dalam tahap ini, butir soal disunting atau disisihkan berdasarkan kesesuaiannya dengan konstruk PIRLS, proses kognitif, dan kelayakan linguistik. Terakhir, butir soal yang tersisa dievaluasi secara psikometris menggunakan responden mahasiswa dan dibandingkan dengan butir soal buatan manusia, yang menunjukkan kinerja sebanding dalam hal indeks tingkat kesulitan dan daya pembeda. Berdasarkan pekerjaan kami sebelumnya, kami mengusulkan alur kerja AIG yang digeneralisasi (lihat Gambar 14.1) yang mengintegrasikan komponen-komponen kunci pengembangan item: penyelarasan dengan kerangka penilaian, rekayasa *prompt*, pembuatan konten otomatis, dan jaminan kualitas melalui pemantauan otomatis dan manusia.



Gambar 14.1 Alur kerja AIG umum

Alur kerja dimulai dengan spesifikasi kerangka penilaian target, termasuk tingkat kelas, domain konten, proses kognitif, dan batasan format item, yang berfungsi sebagai dasar untuk desain prompt. Prompt disusun dengan cermat untuk mencerminkan elemen-elemen cetak biru ini dan dapat mencakup item contoh, teks stimulus, atau tujuan pembelajaran yang ditargetkan. Setelah diselesaikan, prompt diteruskan ke LLM yang menghasilkan kumpulan item kandidat yang selaras dengan kerangka kerja.

Dalam konteks di mana item historis tersedia, item yang dihasilkan dapat difilter menggunakan metrik kesamaan leksikal dan semantik untuk mengidentifikasi item yang paling sesuai dengan gaya, kompleksitas, dan konten item yang telah divalidasi sebelumnya. Langkah penyaringan otomatis ini membantu mempersempit kumpulan item menjadi kandidat berkualitas tinggi sebelum pemantauan manusia. Namun, dalam lingkungan tanpa akses ke item-item lama, peneliti dapat mengadopsi strategi evaluasi otomatis alternatif yang diambil dari riset terbaru. Ini termasuk menggunakan model peninjau berbasis LLM (misalnya, varian GPT yang disesuaikan dengan instruksi) untuk menilai kualitas item berdasarkan dimensi seperti kejelasan, keselarasan, dan tuntutan kognitif, atau mendorong model untuk menilai sendiri dan merevisi outputnya.

Meskipun teknik otomatis ini belum menjadi pengganti validasi manusia, teknik ini menawarkan cara yang terukur untuk menghilangkan output berkualitas rendah dan mengurangi beban peninjau. Item yang tersisa kemudian menjalani peninjauan dengan melibatkan manusia, di mana pakar materi pelajaran (SME) menilai kesetiaan terhadap konstruk, keakuratan faktual, kesesuaian linguistik, dan keadilan. Setelah dikurasi, item dapat diuji di lapangan dan dikalibrasi secara psikometrik. Alur kerja AIG yang digeneralisasi ini juga dapat disertai dengan indikator pemantauan konkret yang mendukung penggunaan operasional dalam penilaian skala besar. Metrik yang berguna meliputi proporsi item yang dihasilkan AI yang dipertahankan setelah pemantauan ahli materi pelajaran (SME); sejauh mana item yang diterima memenuhi batasan cetak biru pada konten, proses kognitif, dan format; persentase item yang diujicobakan yang mencapai statistik kesesuaian item yang dapat diterima; dan analisis fungsi item diferensial (DIF) komparatif untuk item yang dihasilkan AI versus item yang dibuat oleh manusia.

Pelaporan indikator-indikator ini akan memberikan dasar yang lebih jelas untuk mengevaluasi alur kerja AIG, membantu memahami tidak hanya apakah AI dapat menghasilkan item tetapi juga dalam kondisi apa item tersebut secara psikometrik dan substantif sebanding dengan kumpulan item yang sudah mapan. Dengan menggabungkan cetak biru penilaian ke dalam proses generatif, alur kerja ini memastikan bahwa keluaran LLM secara pedagogis tepat dan secara psikometrik dapat dipertahankan. Struktur modularnya memungkinkan adaptabilitas di berbagai domain dan jenis model, sementara penggunaan gabungan validasi otomatis dan manusia memastikan skalabilitas dan kontrol kualitas (QC). Dengan demikian, alur kerja ini mewakili pendekatan yang kompatibel ke depan untuk AIG yang dapat berkembang seiring dengan kemajuan di bidang pemrosesan bahasa dan psikometri.

Verifikasi dan Penyelarasan Butir Soal

Verifikasi butir soal, penyelarasan, dan pemetaan konten telah lama menjadi komponen dasar dalam kerangka kerja validasi asesmen. Secara historis, prosedur penyelarasan mengharuskan para ahli materi pelajaran (SME) untuk mengevaluasi secara manual kesesuaian antara butir soal tes, materi pembelajaran, dan standar pembelajaran. Penyelarasan sering kali dilakukan melalui lokakarya beberapa hari atau pemantauan panel, dengan menggunakan teknik pemetaan kurikulum dan penilaian kualitatif untuk memastikan validitas konten. Meskipun efektif, pendekatan-pendekatan ini pada dasarnya memakan banyak tenaga dan memiliki keterbatasan dalam hal skalabilitas.

Kemajuan terkini dalam bidang NLP dan LLM telah menghadirkan pendekatan yang lebih sistematis dan dapat ditingkatkan skalanya untuk penyelarasan konten melatih model berbasis *transformer* untuk menyelaraskan butir soal tes dengan standar pendidikan; mereka mencapai kinerja prediksi yang moderat karena ketergantungan pada *embedding* tingkat kata yang dirata-ratakan dan kurangnya nuansa kontekstual. Sebaliknya, Khan dan rekan-rekan memperkenalkan sistem *Catalog* yang memanfaatkan *embedding* kalimat berbasis *transformer* untuk mencocokkan teks bacaan sains dengan *Next-Generation Science Standards*.

Mengembangkan hal tersebut, pihak lain mengusulkan kerangka kerja yang dibantu oleh NLP (bukan semata-mata berbasis NLP) yang mengintegrasikan *embedding* kalimat dan kemiripan semantik untuk mendukung pengambilan keputusan oleh SME. Butterfuss dan Doran mengusulkan metode yang menggunakan kemiripan kosinus (*cosine similarity*) dari *embedding* teks yang dihasilkan *transformer* untuk melakukan pemeringkatan awal terhadap potensi kecocokan antar-standar, sehingga secara signifikan mengurangi beban kognitif bagi SME. Kerangka kerja mereka menekankan bahwa *machine learning* (ML) sebaiknya mengatur dan memprioritaskan kandidat penyelarasan, namun keputusan akhir harus tetap berada di tangan peninjau ahli. Demikian pula, Camilli memperluas peran *embedding* dengan menerapkan teknik regresi untuk menyelidiki penyelarasan antara *Common Core State Standards* dan spesifikasi butir soal NAEP. Selanjutnya, Camilli dan Suter menerapkan teknik pengelompokan (*clustering*), termasuk *k-means*, untuk mengelompokkan standar dan butir soal berdasarkan kemiripan semantik mereka.

Temuan mereka memperkuat kegunaan kerangka kerja berbasis *embedding*, tidak hanya untuk tugas penyelarasan tetapi juga untuk analisis validitas konstruk dan koherensi di seluruh konten yang bersifat hierarkis. Meskipun kesamaan semantik berbasis *embedding* memberikan bukti awal yang berharga mengenai keselarasan konstruk, hal ini hanyalah salah satu komponen dari kerangka kerja validasi yang lebih luas yang harus dilengkapi dengan bukti psikometrik empiris. Walaupun selama ini fokus keselarasan tertuju pada standar dan kerangka kerja kurikulum, upaya memastikan bahwa butir soal yang dihasilkan oleh AI memenuhi ekspektasi psikometrik menghadirkan serangkaian tantangan baru. Dalam beberapa tahun terakhir, upaya pembelajaran yang signifikan telah beralih pada prediksi *a priori* mengenai karakteristik butir soal sebagai proksi bagi keselarasan psikometrik.

Pembelajaran awal di bidang ini menggunakan fitur-fitur permukaan seperti frekuensi kata, keterbacaan, dan panjang kalimat untuk memprediksi tingkat kesulitan. Seiring dengan perkembangan NLP, para peneliti mulai menggunakan *embedding* yang peka terhadap konteks untuk mengekstrak fitur semantik, sehingga memungkinkan prediksi tingkat kesulitan berbasis teks yang lebih akurat. Model berbasis Transformer telah memajukan ranah pembelajaran ini lebih jauh. Sebagai contoh, Yeung dan rekan-rekan menggunakan BERT untuk menghasilkan pengecoh (*distractor*) dengan tingkat kesulitan yang terkendali melalui penghitungan kemiripan semantik dalam konteks soal pilihan ganda. Penerapan yang lebih baru telah menggunakan model-model ini baik melalui *fine-tuning* maupun pendekatan *zero-shot* untuk memprediksi tingkat kesulitan butir soal dan waktu respons dengan akurasi yang agak lebih tinggi dibandingkan metode tradisional. Tran dan rekan-rekan memperluas arah pembelajaran ini dengan menunjukkan bahwa fitur manual, penalaran *chain-of-thought*, dan LLM yang telah melalui *fine-tuning* secara bersama-sama menjelaskan sekitar 60% varians dalam parameter kesulitan IRT pada butir soal matematika. Temuan mereka menyoroti potensi ekstraksi fitur berbasis AI untuk mengidentifikasi butir soal yang tingkat kesulitan empirisnya kemungkinan berada di luar rentang target.

Pendekatan-pendekatan ini sangat relevan dalam pembuatan butir soal berbantuan AI (AIG), di mana kualitas psikometrik harus diantisipasi dan dipantau selama proses pembuatan, bukan hanya setelah validasi empiris. Kebutuhan ini sebagian muncul akibat skala dan kecepatan AI dalam menghasilkan butir soal. Hasil produksi dalam volume besar ini meningkatkan risiko terciptanya butir soal yang tidak selaras dengan domain konten, tuntutan kognitif, dan distribusi tingkat kesulitan yang ditargetkan. Tanpa pemeriksaan psikometrik pada tahap awal, terdapat kemungkinan lebih besar untuk menghasilkan butir soal yang tidak dapat digunakan, yang akan memakan biaya tinggi untuk diidentifikasi dan disaring setelah proses pembuatan selesai (*post hoc*). Oleh karena itu, mengintegrasikan pemantauan psikometrik ke dalam alur kerja pembuatan butir soal menjadi suatu kebutuhan praktis. Meskipun butir soal yang dibuat manusia juga melalui tahap peninjauan dan validasi, volume yang dapat dikelola dan laju produksi yang lebih lambat biasanya memungkinkan proses-proses tersebut dilakukan setelah pembuatan tanpa menghabiskan sumber daya secara berlebihan.

Terlepas dari kemajuan signifikan dalam penilaian tingkat nasional maupun negara bagian, pembelajaran mengenai keselarasan dalam studi penilaian skala besar internasional (ILSA) masih relatif kurang dieksplorasi. TIMSS secara sistematis menjaga keselarasan konten melalui proses pengembangan butir soal yang ketat dan pemantauan oleh para ahli. Berlandaskan kemajuan tersebut serta prinsip-prinsip desain berbasis bukti, Bezirhan dan von Davier mengusulkan model klasifikasi berbasis LLM untuk memastikan keselarasan butir soal dengan kerangka kerja penilaian TIMSS. Studi mereka mengevaluasi kemampuan GPT-4o dalam mengklasifikasikan butir soal matematika TIMSS ke dalam domain konten dan kognitif serta menentukan tingkat kesulitannya.

Dengan menggunakan pendekatan rekayasa *prompt* terstruktur yang berlandaskan kerangka kerja TIMSS, model tersebut menunjukkan akurasi tinggi untuk klasifikasi konten (~95%) serta kinerja moderat untuk domain kognitif dan tingkat kesulitan. Hasil ini

mengindikasikan bahwa LLM bila dipandu oleh kerangka kerja terstruktur dan *prompting* yang dinamis dapat berperan sebagai mitra peninjau (*co-reviewer*) yang andal dalam alur kerja pengembangan penilaian, sehingga mengurangi beban kerja ahli materi pelajaran (SME) sekaligus menjaga validitas konstruk. Perkembangan pembelajaran mengenai verifikasi dan penyelarasan butir soal telah beralih dari model penilaian yang sepenuhnya bergantung pada pakar menuju kerangka kerja hibrida yang memadukan skalabilitas NLP dengan nuansa keahlian manusia. *Text embedding* dan model berbasis *transformer* kini menjadi alat yang ampuh untuk mengurangi beban kerja SME serta mendukung penyelarasan antara butir soal, standar, dan spesifikasi. Konsensus yang muncul saat ini menyatakan bahwa alur kerja hibrida yang memanfaatkan LLM atau model *embedding* untuk penyaringan cepat, diikuti dengan pemantauan terarah oleh SME atau ahli psikometri menawarkan keseimbangan terbaik antara efisiensi dan validitas untuk proses verifikasi dan penyelarasan butir soal berskala besar.

Penyusunan Tes Otomatis

Penyusunan tes secara konvensional masih didominasi oleh pakar materi manusia yang memilih butir soal secara manual, melakukan berbagai siklus peninjauan, serta menyeimbangkan cakupan materi dan target statistik secara manual. Proses ini melibatkan peninjauan yang memakan banyak waktu, di mana para pakar secara berulang memilih dan merevisi butir soal agar memenuhi spesifikasi cetak biru (*blueprint*) atau kerangka kerja tes. Namun, metode semacam itu semakin tidak praktis mengingat skala dan kompleksitas asesmen modern. Penyusunan perangkat tes secara manual tidak hanya mahal dan tidak efisien, tetapi juga rentan terhadap ketidakkonsistenan dan kurangnya representasi konstruk, terutama saat berupaya memenuhi target distribusi materi dan kognitif yang sangat rinci sebagaimana diwajibkan oleh kerangka kerja internasional seperti TIMSS atau PIRLS.

Kepatuhan terhadap cetak biru sangatlah krusial dalam asesmen berskala besar karena hal ini menjadi landasan bagi validitas interpretasi skor serta keterbandingan skor lintas negara atau secara longitudinal. Kerangka kerja biasanya menetapkan proporsi butir soal yang rinci di berbagai domain materi, proses kognitif, dan jenis butir soal. Sebagai contoh, kerangka kerja matematika TIMSS 2023 mensyaratkan bahwa 65% butir soal untuk kelas delapan harus selaras dengan domain kognitif penerapan (*applying*) atau penalaran (*reasoning*).

Ketika tes menyimpang dari spesifikasi tersebut baik karena kelebihan representasi pada area materi atau kognitif tertentu maupun kekurangan representasi pada area lainnya inferensi mengenai cakupan konstruk berdasarkan skor tes dapat terganggu. Pelanggaran terhadap cetak biru terbukti dapat menghasilkan model penyusunan tes yang tidak dapat dilaksanakan (*infeasible*), terutama jika bank soal tidak memiliki cakupan yang memadai atas domain yang ditargetkan. Selain itu, kegagalan dalam memenuhi batasan cetak biru memiliki implikasi lanjutan terhadap analisis psikometrik, pelaporan, dan interpretasi kebijakan, khususnya dalam asesmen tren yang mengandalkan representasi materi yang konsisten di berbagai siklus pelaksanaan.

Untuk mengatasi tantangan tersebut, pendekatan penyusunan tes otomatis (*automated test assembly* atau ATA) muncul sebagai solusi yang layak. Kerangka kerja model linear dari van der Linden merumuskan ATA sebagai masalah pemrograman linear bilangan

bulat campuran (*mixed-integer linear programming* atau MILP); dalam pendekatan ini, algoritma menggunakan *solver* modern seperti lp_solve untuk menghasilkan perangkat tes yang mendekati optimal secara andal, dengan tetap mematuhi berbagai batasan psikometrik dan cetak biru. Akan tetapi, ketika bank soal berukuran besar dan kompleks, penggunaan formulasi MILP standar sering kali menghasilkan solusi yang tidak dapat dilaksanakan atau tidak optimal. Integrasi algoritma metaheuristik seperti algoritma genetika dan *cuckoo search*—telah meningkatkan kemampuan ATA dengan menyediakan solusi tangguh untuk masalah optimasi multi-kriteria yang menuntut keseimbangan cermat antara target informasi tes, spesifikasi tumpang tindih butir soal, dan parameter pengendalian paparan soal.

Secara bersamaan, pendekatan berbasis AI mentransformasi proses penyusunan tes dengan memanfaatkan *machine learning* (ML), *reinforcement learning*, dan *natural language processing* (NLP) untuk meningkatkan aspek konstruksi bentuk tes maupun penyajian tes secara adaptif. Sebagai contoh, *neural computerized adaptive testing* merumuskan pemilihan butir soal sebagai masalah *reinforcement learning* dan mempelajari algoritma pemilihan secara langsung dari data. Pendekatan lain seperti *Bandit CAT*, menggunakan *contextual bandits* dan *Thompson sampling* untuk mengakomodasi ketidakpastian serta mengendalikan paparan butir soal secara adaptif selama penyajian tes, sebagaimana diterapkan pada *Duolingo English Test*. Sistem-sistem ini terbukti mampu mengurangi durasi tes tanpa mengorbankan presisi pengukuran, sehingga memberikan peningkatan efisiensi yang signifikan.

Metode penyusunan tes yang diperkuat AI juga memungkinkan pelaksanaan asesmen yang lebih adaptif dan terarah pada kelompok tertentu, sehingga lebih akurat dalam mencerminkan kompetensi individu peserta didik. Kemajuan terkini dalam *explainable AI* (XAI) semakin menyempurnakan sistem ini dengan menambahkan lapisan interpretabilitas ke dalam proses tes adaptif dan penyusunan tes, yang memungkinkan *Large Language Models* (LLM) menghasilkan penjelasan rasional yang dapat dipahami manusia terkait pemilihan butir soal. Pendekatan-pendekatan ini menjanjikan bagi bidang asesmen pendidikan, dengan potensi menghasilkan sistem yang tidak hanya efisien dan memiliki landasan psikometris yang kuat, tetapi juga transparan, bersifat diagnostik, dan berpusat pada peserta didik.

14.3 DAMPAK AI TERHADAP PRAKTIK PENILAIAN

AI telah mengubah lanskap penilaian pendidikan secara mendasar dengan mengurangi upaya manual yang diperlukan dalam penskoran dan pengembangan penilaian. Tugas-tugas yang sebelumnya membutuhkan banyak tenaga manusia kini dapat diperkuat atau diotomatisasi sebagian menggunakan LLM dan teknik NLP. Sistem penskoran otomatis, misalnya, telah menunjukkan tingkat kesesuaian yang tinggi dengan penilai manusia, sehingga secara signifikan mengurangi waktu penyelesaian sembari tetap menjaga reliabilitas. Demikian pula, sistem AIG (pembuatan butir soal otomatis) berbasis AI memiliki potensi untuk mempercepat pembuatan butir soal yang selaras dengan kerangka penilaian, sehingga memperkaya bank soal dengan lebih cepat dan mencakup cakupan domain yang lebih luas.

Selain peningkatan efisiensi, algoritma-algoritma ini menawarkan solusi yang dapat ditingkatkan skalanya (*scalable*). Peningkatan skalabilitas ini sangat berharga bagi ILSA (penilaian berskala besar internasional), yang harus menyajikan penilaian yang konsisten dan valid di berbagai konteks linguistik, budaya, dan pendidikan. Sistem perakitan tes dan penskoran yang diperkuat AI memungkinkan penilaian ini ditingkatkan skalanya tanpa mengorbankan kualitas psikometrik. Selain itu, seiring dengan semakin matangnya model AI yang berkembang dengan kecepatan luar biasa hingga bisa terlewatkan dalam sekejap mata model-model ini menawarkan transparansi dan kemampuan audit yang lebih baik; fitur-fitur yang sangat penting dalam studi perbandingan internasional di mana reliabilitas, keadilan, dan kemampuan untuk dipertanggungjawabkan (*defensibility*) menjadi hal yang utama.

Semakin meluasnya penggunaan penilaian berbasis komputer memungkinkan pengumpulan data proses yang kaya seperti waktu respons, frekuensi peninjauan kembali butir soal, penekanan tombol (*keystrokes*), aliran klik (*clickstreams*), dan pola navigasi yang merekam bagaimana peserta tes berinteraksi dengan tugas-tugas yang diberikan. Data yang terstruktur secara temporal ini memberikan wawasan berharga mengenai perilaku yang menyimpang, strategi pengerjaan tes, dan keterlibatan peserta. Namun, kompleksitas data tersebut menuntut metode yang canggih untuk menggali maknanya. Kemajuan terkini dalam ML dan AI telah memungkinkan para peneliti untuk menganalisis data proses yang bersifat sekuensial melalui teknik-teknik seperti *autoencoder*, *contextual embedding*, RNN, dan model Markov. Pendekatan-pendekatan ini memungkinkan deteksi anomali, identifikasi strategi kognitif laten, dan perbandingan antar-kelompok dalam proses pemecahan masalah.

Bidang kajian ini sejalan dengan psikometrika komputasional, yang mengintegrasikan model pembelajaran mesin (ML) dan kecerdasan buatan (AI) dengan teori psikometrika untuk menangani data berskala besar dan multimodal yang berasal dari pembelajaran dan penilaian digital. Pemanfaatan data multimodal semacam itu membuka peluang baru untuk menafsirkan proses pembelajaran mahasiswa, khususnya melalui penggunaan butir soal terbuka (*open-ended*) dan butir soal berbasis teknologi.

Pada tingkat yang lebih luas, wawasan yang diperoleh dari sistem penilaian berbasis AI dapat menjadi landasan bagi pengambilan keputusan instruksional dan kebijakan pendidikan. Dengan mengagregasi dan menganalisis data penilaian lintas sekolah, wilayah, atau populasi, AI dapat membantu mengidentifikasi kesenjangan pencapaian yang bersifat sistemik, memantau efektivitas kurikulum, serta mendukung jalur pembelajaran yang lebih personal dan berkeadilan. Model ML semakin banyak diterapkan pada data ILSA untuk memprediksi kinerja Mahasiswa dan mengungkap faktor-faktor yang berkontribusi terhadap keberhasilan akademik. Sebagai contoh, *explainable deep learning* dan model *gradient-boosting* yang dilatih menggunakan data TIMSS menunjukkan bahwa variabel seperti penggunaan strategi metakognitif, durasi pembelajaran, dan indikator iklim sekolah merupakan faktor pendorong utama pencapaian matematika. Pada saat yang sama, penerapan model jaringan saraf (*neural network*) dan *ensemble* pada data sains PISA serta penggunaan pengklasifikasi yang peka terhadap kurikulum pada data sains TIMSS menunjukkan bahwa ML mampu memprediksi kinerja secara akurat. Jika dikombinasikan dengan metode XAI, model-model ini dapat

menghasilkan wawasan yang dapat ditafsirkan seperti peringkat kepentingan fitur yang membantu pembuat kebijakan dan pendidik dalam memahami mengapa kelompok mahasiswa tertentu mungkin tertinggal. Kemampuan semacam itu memberikan wawasan yang dapat ditindaklanjuti bagi para pemangku kepentingan, pendidik, administrator, dan pembuat kebijakan dalam upaya meningkatkan hasil pendidikan.

14.4 INTEGRASI AI KE DALAM PENILAIAN PENDIDIKAN

Integrasi AI ke dalam penilaian pendidikan membawa serta tantangan etis. Seiring dengan semakin besarnya pengaruh sistem AI terhadap cara penilaian dikembangkan, dinilai, dan ditafsirkan, bidang ini harus menghadapi berbagai isu terkait bias algoritmik, transparansi, privasi data, bias otomatisasi, dan berkurangnya pengawasan manusia. Salah satu kesenjangan mendasar dalam bidang ini adalah tidak adanya pedoman formal mengenai penggunaan AI dalam pengukuran pendidikan. *Standards for Educational and Psychological Testing* yang menjadi acuan utama bagi validitas, reliabilitas, dan keadilan dalam penilaian saat ini belum memuat panduan khusus terkait AI. Untuk mengatasi kesenjangan ini, Burstein dan rekan-rekannya memperkenalkan *Responsible AI Standards* (RAIS), sebuah kerangka kerja institusional pertama yang menguraikan tanggung jawab etis dan teknis dalam penerapan AI pada penilaian pendidikan. Dokumen RAIS menyoroti prinsip-prinsip utama seperti transparansi, inklusivitas, dan pemantauan berkelanjutan, serta menjadi landasan penting bagi revisi *Standards* di masa mendatang.

Kontribusi penting lainnya terhadap wacana etika AI dalam penilaian pendidikan adalah laporan dari Johnson dan rekan-rekan, yang menjabarkan lima prinsip utama AI yang bertanggung jawab dalam pengukuran dan pembelajaran: keadilan dan mitigasi bias, privasi dan keamanan, transparansi dan akuntabilitas, dampak pendidikan dan integritas, serta perbaikan berkelanjutan. Kelompok Minat Khusus (*Special Interest Group*) NCME mengenai AI dalam Pengukuran dan Pendidikan telah menerbitkan sebuah kertas putih (*white paper*) yang menyajikan perspektif komprehensif dan menyeluruh mengenai isu-isu tersebut.

Upaya kolaboratif ini mempertemukan berbagai peneliti untuk mengeksplorasi implikasi etis dari perangkat berbasis AI, menangani tantangan seperti bias otomatisasi dan keberlanjutan lingkungan, serta mengusulkan kerangka kerja yang dapat diterapkan untuk integrasi AI yang bertanggung jawab dalam pengukuran pendidikan. Transparansi merupakan salah satu kekhawatiran yang paling sering muncul terkait sistem penilaian berbasis AI. Model berbasis *transformer* sering kali beroperasi sebagai "kotak hitam" (*black box*), sehingga sulit untuk melacak dasar pemikiran di balik keluaran (*output*) tertentu. Tanpa proses pengambilan keputusan yang dapat ditafsirkan, validasi keluaran AI berdasarkan standar psikometrik yang mapan menjadi sulit dilakukan.

Kurangnya kemampuan interpretasi ini dapat melemahkan dasar pertanggungjawaban hasil tes serta membatasi kemampuan pemangku kepentingan untuk mempertanyakan atau memverifikasi keputusan penilaian. Untuk memitigasi hal ini, para peneliti menganjurkan penggunaan metode XAI dan sistem hibrida, di mana hasil keluaran AI ditinjau atau dilengkapi oleh pakar manusia. Meskipun transparansi dan bias merupakan masalah yang diakui secara

luas, penerapan langkah-langkah pengamanan secara operasional memerlukan metodologi yang konkret. Beberapa perangkat psikometrik yang sudah mapan dapat diadaptasi ke dalam alur kerja berbasis AI. Sebagai contoh, analisis DIF dapat diterapkan untuk membandingkan skor yang dihasilkan manusia dan AI di berbagai subkelompok demografis guna mendeteksi ketimpangan sistematis dalam perilaku model.

Dasbor bias, yang memantau keluaran model secara berkelanjutan di berbagai subkelompok, jenis butir soal, dan bahasa, menyediakan mekanisme yang dapat diskalakan untuk mendeteksi pola pergeseran (*drift*) atau varians yang tidak relevan dengan konstruk. Deteksi anomali pada tingkat butir soal dan respons, serta desain yang melibatkan manusia dalam prosesnya (*human-in-the-loop*), dapat memberikan pengawasan terhadap kasus-kasus khusus (*edge cases*), respons baru, atau ketidakkonsistenan yang terdeteksi. Bias algoritmik tetap menjadi perhatian kritis dalam sistem AI. Karena LLM dilatih menggunakan data historis yang mencerminkan struktur dan kesenjangan sosial, sistem ini dapat secara tidak sengaja mereproduksi bias yang mengakar seperti penggunaan bahasa yang bias gender, kerangka kurikulum bernuansa kolonial, atau kesenjangan akses digital dalam hasil keluarannya. Sebagai contoh, LLM mungkin menyerap asosiasi stereotipikal atau gagal mengenali ungkapan khas dialek dan budaya tertentu, sehingga merugikan mahasiswa dari kelompok bahasa atau budaya yang kurang terwakili.

Bias dalam model AI dapat muncul dari berbagai sumber: data pelatihan yang tidak seimbang, representasi subpopulasi yang tidak setara, atau korelasi tak disengaja yang dipelajari selama proses pelatihan. Dalam penilaian pendidikan, bias semacam itu dapat menimbulkan kerugian sistematis bagi kelompok bahasa atau budaya yang kurang terwakili, serta mengancam keadilan dalam interpretasi skor tes. Untuk memitigasi risiko ini, evaluasi keadilan yang kokoh sangatlah penting. Hal lain yang sama mengkhawatirkannya adalah fenomena ilusi otomatisasi sebuah bias kognitif di mana para pemangku kepentingan terlalu mengandalkan hasil keluaran mesin karena kecepatan, presisi, dan persepsi objektivitasnya, bahkan ketika terdapat ketidakpastian atau kesalahan. Sebagaimana ditekankan di sepanjang bab ini, sistem AI paling efektif jika dirancang untuk memperkuat, bukan menggantikan, keahlian manusia setidaknya pada tahap saat ini. Ketergantungan tanpa sikap kritis terhadap model penilaian otomatis atau klasifikasi butir soal, terutama tanpa adanya pengawasan manusia, meningkatkan risiko diterimanya hasil keluaran yang cacat tanpa pemeriksaan yang memadai. Hal ini menegaskan pentingnya mempertahankan kerangka kerja yang melibatkan manusia (*human-in-the-loop*) guna memastikan keterpahaman hasil, kepekaan terhadap konteks, dan keselarasan dengan konstruk yang dituju.

Kekhawatiran mengenai privasi dan keamanan data juga sama krusialnya dalam penerapan AI pada penilaian pendidikan. Sistem AI, khususnya yang dijalankan pada platform berbasis *cloud*, dapat secara tidak sengaja mengekspos informasi sensitif mahasiswa. Untuk memitigasi risiko privasi, institusi sebaiknya memprioritaskan penggunaan model yang diterapkan secara lokal jika memungkinkan, guna meminimalkan transmisi data ke pihak luar. Selain itu, penerapan protokol API yang menjaga privasi seperti saluran komunikasi terenkripsi, akses berbasis token, dan pencatatan penggunaan yang ketat dapat membantu memastikan

bahwa data mahasiswa tetap terlindungi di sepanjang alur proses penilaian. Memasukkan langkah-langkah perlindungan ini ke dalam desain sistem sangatlah penting untuk menjaga kepercayaan serta mematuhi standar hukum dan etika terkait penggunaan data pendidikan.

Dampak lingkungan dari AI generatif telah banyak diperdebatkan, khususnya terkait kebutuhan energi untuk melatih dan mengoperasikan LLM. Namun, penting untuk menempatkan kekhawatiran ini dalam konteks penilaian pendidikan. Pada praktiknya, sebagian besar aplikasi AI di bidang ini lebih mengandalkan proses inferensi daripada pelatihan, dan konsumsi dayanya tergolong rendah. Sebagai contoh, pemberian skor pada jawaban terbuka menggunakan model generatif yang di-hosting secara lokal menghabiskan energi sekitar 2 watt-jam per jawaban, yang setara dengan biaya energi kurang dari Rp 10. Jika diperhitungkan untuk pelaksanaan penilaian secara menyeluruh (misalnya, 135 butir soal, 800 jawaban per butir soal, di 60 negara), dampaknya setara dengan jejak karbon dari 0,0147 penerbangan lintas Atlantik.

Sebaliknya, penilaian tradisional oleh manusia dalam studi penilaian berskala besar internasional (ILSA) memerlukan pelatihan ekstensif; sesi pelatihan ini melibatkan aspek logistik seperti sekitar 280 penerbangan internasional, 700 malam menginap di hotel, dan 1.500 hari kerja penilaian di berbagai pusat nasional, yang mencerminkan beban lingkungan yang cukup besar. Meskipun keberlanjutan harus tetap menjadi pertimbangan dalam penerapan AI, bukti menunjukkan bahwa banyak kasus penggunaan di bidang Pendidikan terutama yang menggantikan alur kerja beremisi tinggi dapat menjadi bagian dari solusi, bukan justru menjadi masalah.

14.5 MASA DEPAN ASESMEN BERBASIS AI

Seiring berkembangnya bidang ini, beberapa area utama memerlukan pembelajaran dan pengembangan lebih lanjut. Salah satu area penting adalah bidang yang sedang berkembang pesat, yaitu model multimodal berskala besar (*large multimodal models*) yang mengintegrasikan input teks, audio, video, dan gambar. Sejauh ini, hanya sedikit aplikasi AI dalam pembelajaran pendidikan yang berfokus pada model multimodal. Model-model ini menjanjikan representasi kompetensi pembelajar yang lebih kaya, khususnya dalam ranah seperti eksperimen ilmiah, kemahiran berbahasa lisan, dan pemecahan masalah secara kreatif. Namun, kemajuan semacam itu akan menuntut adanya argumen validitas baru, protokol anotasi multimodal, dan model psikometrik yang mampu menangani data yang.

Seiring berkembangnya kebutuhan pendidikan, asesmen akan semakin diarahkan untuk mengukur kompetensi baru seperti kolaborasi, kreativitas, literasi digital, dan literasi AI. Kompetensi-kompetensi ini merupakan konstruk yang sulit diukur menggunakan format tes tradisional. AI berpotensi memainkan peran vital dalam mengoperasionalkan dan mengevaluasi kompetensi tersebut, terutama melalui simulasi dinamis dan analisis kinerja secara *real-time*. Akan tetapi, hal ini menuntut perhatian cermat terhadap representasi konstruk, kesesuaian budaya, dan kesetaraan bahasa dalam desain berbasis AI, khususnya dalam lingkungan multibahasa seperti ILSA (*International Large-Scale Assessments*).

Penting untuk dicatat bahwa ILSA memerlukan sistem AI yang dirancang sejak awal untuk konteks multibahasa dan lintas budaya. Hal ini mencakup pengembangan model yang responsif secara budaya, pengembangan LLM (*Large Language Models*) pendidikan multibahasa, serta penciptaan kerangka kerja internasional untuk kinerja AI yang adil di berbagai kelompok bahasa dan budaya. Terakhir, arah masa depan yang krusial melibatkan pengembangan pedoman AI yang baku dan spesifik untuk ranah pengukuran pendidikan. Meskipun upaya seperti *Duolingo Responsible AI Standards* dan *NCME's Special Interest Group on AI in Measurement and Education* menunjukkan potensi yang menjanjikan, diperlukan diskusi di tingkat komunitas mengenai perluasan standar validitas dan reliabilitas yang spesifik untuk AI. Hao dan rekan-rekan menganjurkan pendekatan berbasis dimensi waktu (*temporal stance*) yang mengakui adanya pergeseran skala (*scale drift*) dan mewajibkan validasi ulang sistem AI secara berkala, terutama seiring dengan perubahan distribusi data, model, dan konteks sosial. Untuk mendukung upaya-upaya tersebut, kolaborasi lintas disiplin ilmu yang melibatkan bidang psikometri, ilmu komputer, sains data, pendidikan, dan etika akan menjadi hal yang sangat penting. Pembelajaran di masa mendatang harus memprioritaskan XAI, studi dampak longitudinal, serta infrastruktur untuk berbagi tolok ukur (*benchmark*) dan protokol evaluasi lintas program dan konteks.

14.6 AI YANG BERTANGGUNG JAWAB DALAM PENILAIAN PENDIDIKAN

AI telah menghadirkan kemampuan transformatif dalam penilaian pendidikan berskala besar, memungkinkan tingkat skalabilitas, efisiensi, dan personalisasi yang belum pernah terjadi sebelumnya. Melalui penerapannya dalam penilaian otomatis, pembuatan butir soal, penyusunan tes, dan verifikasi butir soal, AI memperluas berbagai kemungkinan dalam praktik psikometrik. Namun, kemajuan ini harus disikapi dengan kehati-hatian yang terukur. Risiko adaptasi berlebihan termasuk ketergantungan tanpa sikap kritis terhadap hasil otomatis, validasi sistem AI yang tidak memadai, serta hilangnya keahlian manusia sama signifikannya dengan peluang yang ada. Kemajuan ini menuntut komitmen baru terhadap keadilan, transparansi, dan validitas konstruk. Sebagaimana telah diilustrasikan dalam bab ini, penerapan AI yang bertanggung jawab memerlukan lebih dari sekadar kecanggihan teknis; hal ini menuntut pendekatan yang berlandaskan prinsip dan terukur, serta didasarkan pada teori psikometrik. Penerapan AI dalam pendidikan tidak dimaksudkan untuk menggantikan keahlian manusia, melainkan untuk memperkuat dan mendukungnya, guna menciptakan sistem di mana penilaian manusia dan kecerdasan mesin berkembang bersama untuk menjunjung tinggi nilai-nilai dasar pengukuran pendidikan. Ke depannya, bidang ini harus menghadapi tantangan ganda berupa inovasi dan tata kelola. Dengan berinvestasi pada validasi, perlindungan etis, dan standar komunitas, kita dapat memastikan bahwa penilaian berbasis AI tidak hanya efisien dan dapat ditingkatkan skalanya, tetapi juga adil, inklusif, dan bermakna secara edukatif.

DAFTAR PUSTAKA

Akbar, A. F., Akbar, J. S., & Musa'adah, S. (2026). Analysis Of The Influence Of The Use Of Artificial Intelligence Assistants On The Digital Literacy Of Generation Z History Education Students. *Keraton: Journal Of History Education And Culture*, *8*(1), 111-117.

Ananda, R., Clarisa, A. L., Nadhifah, S., & Ahyar, M. S. (2026). Integrasi Artificial Intelligence (Ai) Dalam Pembelajaran Ips Berbasis Metakognisi. *Pendas: Jurnal Ilmiah Pendidikan Dasar*, *11*(02), 259-266.

Ardiyanti, R. D. F. (2026). Optimalisasi Ai Dalam Pendidikan Karakter Islami Untuk Meningkatkan Literasi Akhlak Siswa. *Al-Ilmiya: Jurnal Pendidikan Islam*, *1*(6), 206-213.

Aulia, D., Putri, N. D., Aini, H., & Meisofa, Y. (2026). Integrasi Teknologi Artificial Intelligence (Ai) Dalam Perencanaan Pembelajaran Di Era Kurikulum Merdeka. *Jurnal Penelitian Ilmiah Multidisipliner*, *3*(01).

Cholid, N., & Husna, M. A. (2026). Pendidikan Islam Transformatif Dalam Menghadapi Disrupsi Digital, Kecerdasan Buatan, Dan Krisis Moral: Rekonstruksi Paradigma Pendidikan Islam Abad Ke-21. *Ihsan: Jurnal Pendidikan Islam*, *4*(4), 633-653.

Dzulfiqar, A. S., Putri, D. D. V., Masruro, A., Shofiyyullaht, D. S., Ramadhani, A., Zahra, D. A. A., & Syafi'i, I. (2026). Profil Dan Tantangan Guru Di Era Artificial Intelligence: Relevansi Pemikiran Ki Hajar Dewantara. *Jurnal Kepemimpinan Dan Pengurusan Sekolah*, *11*(4), 2463-2477.

Fadillah, S. A., Misdiyana, D., & Pramono, A. (2026). Transformasi Pendidikan Berbasis Teknologi Sebagai Implementasi Sdgs 4 Di Era Revolusi Industri 4.0. *Dewantara: Jurnal Pendidikan Sosial Humaniora*, *5*(2), 315-329.

Fa'izah, A. N., Ramadhani, A. P., Amelia, E., Nur'aeni, F., Nurmala, I., Ruhiyat, S., ... & Mashudi, E. A. (2026). Tinjauan Literatur Seputar Revolusi Ai Dalam Pendidikan Anak Usia Dini. *Child Kingdom: Jurnal Pendidikan Anak Usia Dini*, *4*(1), 73-93.

Fajriyah, S., Irfan, M., & Fadhil, A. (2026). Dominasi Teknologi Di Era Ai Dan Pergeseran Otoritas Guru Dalam Pendidikan. *Advances In Education Journal*, *3*(1), 61-71.

Fikri, F., & Maskur, S. (2026). Pendidikan Islam Di Era Artificial Intelligence (Ai): Peluang, Tantangan, Dan Transformasi Pembelajaran. *Didacta: Journal Of Educational Studies*, *2*(1), 119-137.

Fitriana, N. (2026). Pengaruh Teknologi Ai Terhadap Transformasi Pendidikan Di Era Revolusi Industri 4.0. *Maliki Interdisciplinary Journal*, *4*(2), 384-393.

Hasan, Z., Lb, K. A., Rexy, A., & Victorius, D. A. (2026). Lunturnya Moralitas Dalam Dunia Pendidikan Di Era Globalisasi Dan Kecerdasan Buatan (Ai). *Jurnal Ilmiah Research Student*, *3*(1), 77-84.

Ilmudinulloh, R., Ridwan, I., Sumaizar, S. E., Winata, K. A., & Suardika, I. K. (2026). *Memahami Perubahan Dalam Dunia Pendidikan: Evolusi Atau Revolusi*. Pt. Literatus Digitus Indonesia.

Khan, S. (2026). Brave New Words: Bagaimana Ai Merevolusi Pendidikan (Dan Mengapa Itu Diperlukan).

Khola, N. J., & Febriansyah, A. R. (2026). Transformasi Budaya Pendidikan Agama Islam Dalam Dinamika Digital Konten Sosial Media & Artificial Intelligence (Ai). *Qolamuna: Keislaman, Pendidikan, Literasi Dan Humaniora*, *3*(1), 291-300.

Kurniawan, D., & Rojabi, M. A. (2026). *Stimuler Ai: Revolusi Pelatihan Bahasa Berbasis Ai*. Afdan Rojabi Publisher.

Lutfi, M., Manzah, A. A., Fiandini, S. P., Karima, Q., & Restyan Vanny Varera, S. A. P. (2026). *Revolusi Pendidikan Di Indonesia*. Pustaka Garuda.

Mulya, S. (2026). Artificial Intelligenci Dalam Islam: Implikasi Bagi Literasi Dan Pendidikan Di Era Digital. *Qolamuna: Keislaman, Pendidikan, Literasi Dan Humaniora*, *3*(1), 119-126.

Munawwarah, Z. (2026). Modernisasi Manajemen Pendidikan Islam Melalui Artificial Intelligence (Ai). *Al-Zayn: Jurnal Ilmu Sosial & Hukum*, *4*(2), 3741-3757.

Nugraha, D., Sutia, I., Faiz, I. N., & Mahyani, A. (2026). Transformasi Kurikulum Pendidikan Islam Berbasis Ai Dan Ecotheology Dalam Menghadapi Tantangan Krisis Lingkungan Global. *Takuana: Jurnal Pendidikan, Sains, Dan Humaniora*, *5*(2), 2071-2083.

Nurita, N. F. P., Salsabila, K. N., & Abdun Nasir, M. F. (2026). Analisis Efektivitas Penggunaan Generative Ai Dalam Mengingkatkan Kemandirian Belajar Mahasiswa. *Rekognisi*, *3*(1), 30-43.

Nurul, N. H. (2026). Revolusi Belajar: Pemanfaatan Ai Sebagai Alat Bantu Dalam Pembelajaran Di Kalangan Mahasiswa/I Pgmi Stiq Amuntai. *Al-Majid: Jurnal Ilmiah Manajemen Pendidikan Dan Inovasi Pembelajaran*, *2*(1), 1-16.

Oktavia, I. W., Septian, F. I., Farasalsabila, F., Su, Y. R., Rochmahwati, P., Cahyo, A. N., ... & Anasti, H. P. (2026). *Artificial Intelligence Dan Masa Depan Pendidikan*. Cv. Edu Akademi.

Putri, A., Sari, H. P., Hafiz, M. S., & Zahra, Q. A. (2026). Artificial Intelligence (Ai) Dalam Pendidikan Islam: Peluang, Tantangan, Dan Etika Keislaman. *Al-Abshor: Jurnal Pendidikan Agama Islam*, *3*(2), 239-251.

Qusairi, Q., & Zaini, F. (2026). Transformasi Peran Guru Di Era Artificial Intelligence: Dari Sumber Pengetahuan Menuju Fasilitator Pembelajaran. *Abuya: Jurnal Pendidikan Dasar*, *4*(1), 79-100.

Rahmah, N. S. A., & Salamah, S. (2026). Transformasi Kurikulum Pai Era Digital Melalui Kecerdasan Buatan (Ai), Ekopedagogi, Pendidikan Seksualitas, Dan Moderasi Beragama Dalam Bingkai Kearifan Lokal. *Advances In Education Journal*, *2*(6), 242-253.

Ramadhani, I., Kusumawati, D., & Waziroh, I. (2026). Penggunaan Artificial Intelligence (Ai) Dalam Pengembangan Konsep Pendidikan Islam: Tinjauan Filosofis Dan Praktik. *Edukasi: Jurnal Mahasiswa Kependidikan Islam*, *5*(2), 50-65.

Rosadi, A. B., Fauziyah, A. N., Ihsan, F. N., & Amalia, N. N. (2026). Implementasi Media Digital Dan Artificial Intelligence Dalam Pembelajaran Pendidikan Agama Islam. *Jurnal Manajemen Dan Pendidikan Agama Islam*, *4*(3), 196-217.

Rozie, F., & Pratikno, A. S. (2026). *Revolusi Media Pembelajaran Berbasis Ai (Membangun Ekosistem Pendidikan Digital)*. Pt Bukuloka Literasi Bangsa.

Rusdi, R., Subaidi, S., & Akmal, S. (2026). Rekonstruksi Kurikulum Pesantren Di Era Artificial Intelligence: Integrasi Literasi Digital, Etika Ai, Dan Spiritualitas Islam. *Gahwa*, *4*(2), 20-42.

Salsabila, O. N. (2026). *Mengenal Scale Ai: Pilar Utama Infrastruktur Data Dalam Revolusi Kecerdasan Buatan*. Afdan Rojabi Publisher.

Setiawan, A., Putri, D. A., Wahyu, S., Huda, M. M., Rahmad, I. N., Kurniawan, A., ... & Sutrisno, A. D. (2026). *Edtech 5.0 (Revolusi Media Dan Teknologi Pembelajaran Masa Depan)*. Ganesha Kreasi Semesta.

Sholihah, N. M., Yuwono, W., Rosidin, M. P. I., Friani, D. A., Mm, M. P., Marheni Rayung Puspaningrum, S. P., & Husnayain, M. F. (2026). *Pendidikan Di Era Ai: Belajar, Mengajar, Dan Hidup Di Tahun Digital*. Inspirasi Modern Publishing.

Sides, A., & Firdausih, F. (2026). Pendidikan Islam Di Era Ai: Antara Trasformasi, Tantangan Dan Harapan.

Simanjuntak, H. (2026, May). Pemanfaatan Artificial Intelligence (Ai) Sebagai Media Ajar Dalam Pembelajaran Pendidikan Agama Kristen (Pak). In *Thecred: Proceedings Of The Theology Conference And Religious Education Dialogue* (Vol. 1, No. 1, Pp. 46-52).

Sucipto, S., & Saputra, H. (2026). Transformasi Pembelajaran Pendidikan Agama Islam Melalui Artificial Intelligence Di Era Society 5.0. *Al-Jinan: Jurnal Studi Islam*, *1*(01), 100-111.

Tambunan, G., Tamba, N. D., Nova, A., & Panjahitan, S. R. (2026). Analisis Dampak Penggunaan Kecerdasan Buatan (Ai) Generatif Dalam Pendidikan Terhadap Motivasti Belajar Dan Integritas Akademik. *Ahsani Taqwim: Jurnal Pendidikan Dan Keguruan*, *3*(2), 333-348.

Tumangger, R. A. P., & Pohan, S. (2026). Artificial Intelligence Dan Pendidikan Agama Islam: Upaya Membangun Generasi Berakhlak Di Era Digital. *Hidayah: Cendekia Pendidikan Islam Dan Hukum Syariah*, *3*(2), 62-72.

Wardani, E., Indah, P. D. P., Ramadhani, V. M., Dana, G. W. P., & Wirautami, N. L. P. (2026). *Pendidikan Dan Tantangan Revolusi Industri 5.0*. Star Digital Publishing.

Wibowo, A. (2026). Tantangan Pendidikan Bahasa Di Era Ai (Artificial Intelligence). *Penerbit Yayasan Prima Agus Teknik*.

Willdahlia, A. G., Putra, D. M. D. U., Judijanto, L., Suandana, N. P. W., & Pramitha, G. D. (2026). *Artificial Intelligence Dalam Pendidikan Tinggi: Membangun Kesiapan Kerja Mahasiswa*. Pt. Sonpedia Publishing Indonesia.

Yahya, D. A., Muthayyibah, N. A., Syahida, G. D., & Supriyatna, S. M. (2026). Perkembangan Media Pembelajaran: Dari Media Konvensional Hingga Ai Generatif: Tinjauan Pedagogis Dan Manajemen Pendidikan. *Jurnal Manajemen Dan Pendidikan Agama Islam*, *4*(4), 333-349.

REVOLUSI PENDIDIKAN di zaman AI

Dr. Joseph Teguh Santoso, S.Kom, M.Kom.

BIODATA PENULIS



Dr. Joseph Teguh Santoso, M.Kom memiliki Jabatan Akademik Lektor Kepala dan praktisi industri yang berpengalaman. Saat ini menjabat sebagai Rektor Universitas Sains dan Teknologi Komputer (Universitas STEKOM), salah satu universitas terkemuka di Jawa Tengah, Indonesia. Dengan pengalaman lebih dari 13 tahun di dunia bisnis dan praktisi industri di China, beliau membawa perspektif global dan inovasi yang signifikan ke dalam dunia akademis.Sebagai seorang entrepreneur, penulis adalah pencipta TopLoker.com, sebuah platform inovatif yang merevolusi cara mencari dan menawarkan pekerjaan. TopLoker.com adalah portal lowongan bursa kerja terbesar di Indonesia, khusus untuk pendidikan SMA/SMK sederajat. TopLoker.com telah mendapatkan penghargaan sebagai juara 1 Startup4industry 2022 oleh Kementerian Perindustrian Republik Indonesia. Kontribusi Dr. Joseph dalam menyediakan akses pekerjaan yang luas bagi lulusan SMA/SMK telah membantu banyak individu menemukan peluang kerja yang sesuai dengan keahlian mereka. Selain itu, Dr. Joseph Teguh Santoso, M.Kom adalah pendiri dari dua organisasi yaitu (1) organisasi guru/pendidik PTIC (Perkumpulan Teacherpreneur Indonesia Cerdas) yang bertujuan untuk meningkatkan kualitas pendidikan dan kesejahteraan guru/pendidik dengan wawasan entrepreneurship, serta (2) organisasi industri PERKIVI (Perkumpulan Komunitas Industri dan Vokasi Indonesia) yang berfokus pada pengembangan link and match antara industri dan dunia pendidikan. Sebagai Rektor, Dr. Joseph Teguh Santoso, M.Kom memiliki kepemimpinan yang berorientasi pada hasil, dan berkomitmen untuk mendorong kemajuan Universitas Sains dan Teknologi Komputer (Universitas STEKOM). Saat ini Universitas STEKOM telah mengalami transformasi positif dalam peningkatan kualitas pendidikan, perluasan fasilitas, serta penguatan kemitraan Perguruan Tinggi Nasional dan Internasional. Beliau memprioritaskan pengembangan sumber daya manusia dan penelitian, serta memastikan bahwa universitas berada di garis depan dalam inovasi dan teknologi untuk mencapai tujuan akhir, yaitu lulusan yang mampu bekerja dan sukses setelah lulus. Dr. Joseph Teguh Santoso, M.Kom sering diundang sebagai pembicara di berbagai konferensi nasional maupun internasional dan telah menerima berbagai penghargaan atas dedikasinya dalam bidang pendidikan, industri, dan kewirausahaan.

YAYASAN PRIMA AGUS TEKNIK

PENERBIT :
YAYASAN PRIMA AGUS TEKNIK
Jl. Majapahit No. 605 Semarang
Telp. (024) 6723456. Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id