Indonesian Trust Nursing Journal (ITNJ)
Volume 4, No. 2 – Mei 2026
ISSN Online: 2986-0164
ISSN Cetak: 2986-2116

KECERDASAN BUATAN DALAM PENDIDIKAN KEDOKTERAN: KAJIAN NARATIF TENTANG ASESMEN, UMPAN BALIK, DAN PENALARAN KLINIS

Erwin Handoko[1,*]
[1]Program Studi S1 Ilmu Keperawatan, Universitas Murni Teguh
[*]Koresponding: erwinhandoko@murniteguhuniversity.ac.id

Abstract

Artificial intelligence (AI) is increasingly shaping medical education, particularly in assessment, feedback, and clinical reasoning. This narrative review synthesizes recent evidence on the use of large language models, conversational agents, and AI-enhanced simulation in these three areas. The review highlights that AI can support educators by generating assessment materials, assisting rubric-based scoring, providing formative feedback, simulating patient encounters, and reviewing clinical reasoning documentation. These applications may help address persistent challenges in medical education, including limited faculty time, difficulty assessing higher-order reasoning, and the resource demands of high-quality simulation. However, the benefits of AI are accompanied by important risks. Concerns related to validity, bias, privacy, academic integrity, local relevance, and professional formation remain central, especially when tools developed in well-resourced and English-dominant contexts are applied in different educational and clinical settings. This review argues that AI should not be understood as a replacement for educational judgment, clinical supervision, or human relationships in medical training. Rather, AI is best positioned as a supervised educational tool that can extend the reach of good teaching when embedded in clear curricula, locally validated assessment systems, transparent governance, and faculty development programs. The central challenge for medical educators is therefore not whether AI should be used, but how it can be used responsibly while preserving the ethical, relational, and patient-centered foundations of medical education.

Keywords: Artificial Intelligence, Medical Education, Assessment, Feedback, Clinical Reasoning, Large Language Models

Abstrak

Kecerdasan buatan (*artificial intelligence*/AI) bukan lagi merupakan hal yang baru dalam pendidikan kedokteran. Teknologi ini sudah mulai mengubah cara pendidik merancang asesmen pembelajaran, memberikan umpan balik, dan mendukung proses penalaran klinis. Kajian naratif ini mensintesis bukti terkini mengenai penggunaan large language models, agen percakapan, dan simulasi berbasis AI pada ketiga area tersebut. Hasil kajian laporan terkini menunjukkan bahwa AI dapat membantu pendidik dalam menyusun bahan asesmen, mendukung asesmen berbasis rubrik, memberikan umpan balik formatif, mensimulasikan interaksi pasien, serta menelaah dokumentasi penalaran klinis. Berbagai penerapan tersebut berpotensi membantu mengatasi tantangan yang telah lama dihadapi pendidikan kedokteran, seperti keterbatasan waktu dosen, kesulitan menilai penalaran tingkat tinggi, dan besarnya sumber daya yang diperlukan untuk menyelenggarakan simulasi berkualitas. Namun, manfaat AI juga disertai sejumlah risiko penting. Isu validitas, bias, privasi, integritas akademik, relevansi lokal, dan pembentukan profesionalisme tetap perlu diperhatikan, terutama ketika

ISSN Online: 2986-0164
ISSN Cetak: 2986-2116

alat yang dikembangkan dalam konteks berbahasa Inggris dan bersumber daya tinggi digunakan pada lingkungan pendidikan dan klinis yang berbeda. Review ini menegaskan bahwa AI tidak seharusnya dipahami sebagai pengganti pertimbangan edukasional, supervisi klinis, atau relasi manusia dalam pendidikan kedokteran. Sebaliknya, AI lebih tepat ditempatkan sebagai alat pendidikan yang diawasi dan digunakan dalam kurikulum yang jelas, sistem asesmen yang tervalidasi secara lokal, tata kelola yang transparan, serta program pengembangan dosen. Tantangan utama bagi pendidik kedokteran bukan lagi apakah AI perlu digunakan, melainkan bagaimana AI dapat digunakan secara bertanggung jawab sambil tetap menjaga fondasi etis, relasional, dan berpusat pada pasien dalam pendidikan kedokteran.

Kata Kunci: Kecerdasan Buatan, Pendidikan Kedokteran, Asesmen, Umpan Balik, Penalaran Klinis, Large Language Models

LATAR BELAKANG DAN RUANG LINGKUP

Pendidikan kedokteran sejak lama bergantung pada pertimbangan profesional pendidik. Tanggung jawa dari pendidik di dunia kedokteran, tidak hanya menyampaikan materi, tetapi juga menciptakan pengalaman belajar yang bermakna, menilai kesiapan mahasiswa untuk memperoleh tanggung jawab yang lebih besar, serta memberikan umpan balik yang mendukung perkembangan kompetensi tanpa melemahkan motivasi belajar. Dalam ruang pendidikan yang sangat bergantung pada penilaian manusia, relasi pedagogis, dan sensitivitas terhadap tahap perkembangan peserta didik, kecerdasan buatan (*artificial intelligence*/AI), khususnya *large language models* (LLMs), kini hadir dengan kecepatan yang belum pernah terjadi sebelumnya.

Daya tarik AI dalam pendidikan kedokteran mudah dipahami. Berbagai permasalahan seperti waktu dosen yang terbatas, inkonsistensi penalaran klinis, dan sulitnya mempersiapkan simulasi klinis yang berkualitas tinggi dapat diatasi dengan relatif mudah melalui penggunaan AI. Scoping review BEME mengidentifikasi 278 publikasi mengenai AI dalam pendidikan kedokteran dan melaporkan penerapan AI pada berbagai bidang terkait, seperti seleksi mahasiswa, pembelajaran, asesmen, dan penalaran klinis, dengan sebagian besar penelitian berasal dari institusi di Amerika Utara dan Eropa (Gordon et al., 2024). Konsentrasi geografis ini penting untuk diperhatikan karena teknologi yang dikembangkan dalam konteks berbahasa Inggris, dengan sumber daya tinggi dan standar klinis tertentu, belum tentu menunjukkan kinerja yang sama dalam program pendidikan yang memiliki bahasa, pola penyakit, norma klinis, dan infrastruktur yang berbeda.

Kajian naratif ini difokuskan pada tiga penggunaan AI yang berhadapan langsung dengan peran pendidik kedokteran, yaitu perancangan dan penilaian asesmen, penyusunan umpan balik, serta pengembangan penalaran klinis. Kajian ini mengintegrasikan berbagai studi empiris, penelitian validasi, dan review utama untuk menjawab satu pertanyaan praktis: Bagaimana pendidik di bidang kesehatan dapat menggunakan AI tanpa kehilangan inti manusiawi, relasional, dan etis dalam pendidikan kedokteran? Pembahasan ini terutama menggunakan literatur yang diterbitkan antara tahun 2023 hingga 2026, yang diidentifikasi melalui

ISSN Online: 2986-0164
ISSN Cetak: 2986-2116

penelusuran terarah pada PubMed, Scopus, dan Google Scholar dengan kata kunci terkait AI, LLMs, ChatGPT, pendidikan kedokteran, penalaran klinis, asesmen, dan umpan balik. Sintesis disusun secara tematik dan menekankan implikasi bagi pendidik yang perlu mengambil keputusan saat ini, sering kali sebelum kebijakan dan bukti ilmiah berkembang secara memadai. Lucas et al., 2024) juga menunjukkan bahwa penggunaan LLMs dalam pendidikan kedokteran berkembang lebih cepat daripada kualitas bukti yang mendukung banyak penerapannya.

ASESMEN PEMBELAJARAN DALAM ERA AI GENERATIF

Asesmen dalam pendidikan kedokteran secara tradisional berupaya menyeimbangkan dua tujuan utama, yaitu autentisitas dan akuntabilitas. *Workplace-based assessment*, OSCE, dan *programmatic assessment* dirancang untuk menangkap performa klinis yang mendekati praktik nyata, sedangkan prosedur psikometrik digunakan untuk menjaga keadilan dan reliabilitas penilaian. Kehadiran AI generatif membuat keseimbangan ini menjadi lebih kompleks. Di satu sisi, AI melemahkan akuntabilitas banyak tugas tertulis yang tidak diawasi karena mahasiswa dapat menghasilkan jawaban yang tampak masuk akal terhadap berbagai prompt umum hanya dalam hitungan detik. Di sisi lain, AI berpotensi memperkuat asesmen dengan membantu menyusun kasus, meninjau dokumentasi klinis, dan memberikan umpan balik terstruktur dalam skala yang sulit dicapai bila hanya mengandalkan dosen (Masters et al., 2025).

AMEE Guide yang ditulis oleh Masters et al. (2025) relevan dalam konteks ini karena memandang asesmen berbasis AI bukan sebagai jalan pintas teknis, melainkan sebagai persoalan pedagogis, etis, dan tata kelola. Bukti bahwa LLMs dapat menunjukkan performa baik pada ujian konvensional semakin sulit diabaikan. Penelitian Kung et al. (2023) melaporkan bahwa ChatGPT mampu mencapai atau mendekati ambang kelulusan pada ketiga tahapan USMLE tanpa pelatihan khusus. Penelitian yang lebih baru pada Spanish Medical Intern Resident examination juga menemukan keadaan yang serupa. Luengo Vera et al. (2025) mendapati bahwa model berbasis Claude yang telah disesuaikan mampu memperoleh skor 195 dari 210 pada tahun 2025, meskipun model tersebut kurang andal pada soal yang membutuhkan pengetahuan epidemiologi lokal atau interpretasi gambar. Temuan ini tidak berarti bahwa ujian tertulis harus dihapus. Namun, temuan tersebut menunjukkan bahwa tugas asesmen yang terutama bertumpu pada hafalan atau penalaran rutin menjadi semakin rentan. Program pendidikan perlu memperkuat asesmen terawasi untuk keputusan berisiko tinggi, memperbarui bank soal secara lebih berkala, serta mengembangkan tugas yang menuntut penjelasan, kritik, pembelaan lisan, pertimbangan kontekstual, atau observasi langsung.

AI juga dapat membantu pendidik membangun asesmen yang lebih baik jika digunakan secara hati-hati. Hudon et al. (2025) menemukan bahwa *script concordance tests* yang dihasilkan AI layak dan dapat diterima untuk mengajarkan penalaran klinis dalam situasi ketidakpastian. Artsi et al. (2024) juga melaporkan bahwa LLMs dapat menghasilkan soal pilihan ganda medis dengan kualitas yang cukup baik. Studi oleh Goh et al. (2024) dan Jamieson et al. (2024) juga menunjukkan bahwa LLMs dapat menilai vignette klinis atau catatan pasca-OSCE dengan tingkat kesesuaian yang menjanjikan dibandingkan penilaian pakar.

ISSN Online: 2986-0164
ISSN Cetak: 2986-2116

Salah satu perkembangan yang paling menjanjikan adalah penggunaan LLMs untuk mengevaluasi dokumentasi klinis autentik. Schaye et al. (2025) memvalidasi asesmen berbasis LLM terhadap catatan admisi residen menggunakan rubrik Revised-IDEA, dengan lebih dari 1.000 catatan retrospektif dan satu set validasi prospektif. Temuan ini penting karena catatan klinis dapat memperlihatkan aspek penalaran diagnostik yang sulit diamati secara langsung, sementara supervisor yang sibuk tidak selalu memiliki waktu untuk membaca dan menilai setiap catatan secara mendalam.

Meskipun demikian, perlu ditekankan bahwa banyak studi belum menguji soal-soal yang dibuat oleh AI pada peserta didik nyata. Hal ini penting karena meskipun AI dapat menghasilkan draft awal soal, vignette, rubrik, atau pernyataan umpan balik, tetapi validitas tetap bergantung pada telaah manusia, kesesuaian dengan blueprint, dan data performa peserta didik. Dalam ranah tertentu, AI dapat mengambil peran penilai, tetapi pendidik tetap bertanggung jawab terhadap desain dan interpretasi. Jika digunakan secara formatif, telaah berbasis AI dapat membantu mahasiswa atau residen memperoleh umpan balik yang lebih sering mengenai cara mereka membingkai masalah, menyusun diagnosis banding, dan menghubungkan bukti dengan rencana tata laksana. Dalam fungsi ini, AI lebih tepat dipahami bukan sebagai penguji, tetapi sebagai pembaca berskala besar yang membantu pendidik menemukan pola yang layak didiskusikan.

PEMANFAATAN AI GENERATIF UNTUK PROSES UMPAN BALIK

Umpan balik sering dipandang sebagai inti pembelajaran klinis, tetapi dalam praktiknya tetap menjadi tantangan karena sumber daya yang terbatas. Dosen mungkin ingin memberikan komentar yang rinci, namun tuntutan layanan klinis, pekerjaan administratif, dan jumlah mahasiswa yang besar membatasi apa yang dapat dilakukan. AI dapat membantu mengatasi tantangan penyediaan umpan balik karena mampu menghasilkan respons yang cepat dan terstruktur dengan biaya marginal yang rendah. Bukti yang tersedia masih awal, tetapi mendukung penggunaan yang hati-hati. Dalam sebuah uji acak tersamar ganda, Brügge et al. (2024) et al. (2024) menemukan bahwa mahasiswa yang menerima umpan balik terstruktur berbasis AI setelah percakapan dengan pasien simulasi menunjukkan peningkatan yang lebih baik dibandingkan mahasiswa yang berlatih tanpa umpan balik tersebut, khususnya dalam membangun konteks dan menggali informasi. Ukuran sampel penelitian ini kecil, tetapi hasilnya mengindikasikan bahwa umpan balik berbasis AI dapat mendukung latihan yang dilakukan secara terarah dan berulang, dan disertai umpan balik untuk memperbaiki aspek kemampuan tertentu.

Kualitas umpan balik AI tidak selalu konsisten. Çiçek et al. (2025) et al. menemukan bahwa umpan balik yang dihasilkan ChatGPT memberikan hasil pembelajaran yang sebanding dengan umpan balik pakar pada sebagian pertanyaan penalaran klinis, tetapi umpan balik pakar tetap lebih baik pada kasus yang kompleks. Pola ini dapat dipahami secara pedagogis. Umpan balik AI biasanya cepat, terorganisasi, dan tampak komprehensif. Namun, kualitas seperti ini tidak selalu identik dengan ketepatan. AI dapat menghasilkan komentar yang terlalu umum, terlalu percaya diri, atau kurang peka terhadap tahap perkembangan peserta didik. Hal ini penting diperhatikan karena mahasiswa dapat menganggap umpan balik yang terdengar fasih sebagai

ISSN Online: 2986-0164
ISSN Cetak: 2986-2116

umpan balik yang akurat. Dalam pendidikan kedokteran, umpan balik yang baik tidak hanya perlu terdengar jelas, tetapi juga harus relevan, kontekstual, dan mendukung perkembangan kompetensi. Oleh karena itu, AI sebaiknya diposisikan sebagai lapisan awal dukungan formatif, bukan sebagai penilaian akhir.

Tiga prinsip desain dapat ditarik dari temuan tersebut. Pertama, umpan balik AI perlu ditambatkan pada rubrik yang jelas, seperti Revised-IDEA, instrumen *workplace-based assessment*, atau kerangka kompetensi. Tanpa rubrik, AI hanya menghasilkan komentar yang terdengar rapi, tetapi belum tentu mengarah pada kompetensi yang ingin dibentuk. Kedua, pemanfaatannya sebaiknya ditempatkan terlebih dahulu pada konteks formatif dan berisiko rendah sampai tersedia bukti validitas lokal. Dalam pendidikan kedokteran, semakin besar dampak suatu keputusan terhadap mahasiswa, semakin besar pula kebutuhan untuk memastikan bahwa alat yang digunakan benar-benar valid, adil, dan dapat dipertanggungjawabkan. Ketiga, pendidik perlu mengajarkan literasi umpan balik.

AI dapat menghasilkan banyak komentar, tetapi banyaknya komentar tidak otomatis membuat mahasiswa bertumbuh. Mahasiswa perlu belajar membedakan mana umpan balik yang relevan, mana yang perlu diverifikasi, dan mana yang harus didiskusikan dengan dosen. Dengan demikian, peran pendidik bergeser dari satu-satunya penulis umpan balik menjadi kurator yang membantu mahasiswa mengubah banyak komentar menjadi perbaikan yang bermakna. Peran ini tetap membutuhkan kehadiran, dukungan, dan pertimbangan profesional, terutama ketika umpan balik menyentuh identitas, kepercayaan diri, atau keselamatan pasien.

MANFAAT DAN RISIKO AI DALAM LATIHAN PENALARAN KLINIS

Penalaran klinis merupakan area di mana AI tampak sangat menjanjikan, tetapi juga paling berisiko. Mahasiswa membutuhkan paparan berulang terhadap kasus, kesempatan untuk mengambil keputusan, dan umpan balik terhadap cara berpikir mereka. AI dapat menyediakan kesempatan latihan dalam jumlah yang sulit dicapai oleh banyak program pendidikan. Agen percakapan dapat mensimulasikan wawancara pasien, menghasilkan diagnosis banding, menantang premature closure, dan meminta mahasiswa menjelaskan rencana klinis mereka. Cross et al. (2025) menemukan bahwa mahasiswa menilai ChatGPT sebagai pasien standar virtual yang fleksibel, terutama karena dapat digunakan untuk latihan berulang tanpa batas. Namun, mereka juga mencatat masalah seperti keluaran yang tidak terkurasi, ketergantungan pada keterampilan *prompting*, serta ketiadaan isyarat nonverbal dan temuan pemeriksaan fisik. Platform yang terkurasi menawarkan fidelitas yang lebih baik, sedangkan LLMs umum menawarkan fleksibilitas yang lebih besar.

Literatur yang lebih luas menunjukkan arah yang serupa. Roveta et al. (2025) menyimpulkan bahwa simulasi berbasis AI paling bermanfaat ketika dipadukan dengan umpan balik terstruktur dan debriefing. Aster et al. (2025) juga menekankan bahwa manfaat pendidikan lebih banyak ditentukan oleh desain pembelajaran di sekitar teknologi daripada oleh kebaruan alat itu sendiri. Hal ini sejalan dengan prinsip dasar pendidikan kedokteran. Simulasi bekerja ketika mahasiswa mempersiapkan diri, bertindak, menerima umpan balik, berefleksi, dan mencoba kembali. AI tidak menghapus siklus tersebut, tetapi dapat membuat

ISSN Online: 2986-0164
ISSN Cetak: 2986-2116

beberapa bagiannya lebih mudah diterapkan dalam skala luas.

Risikonya perlu dinyatakan secara jelas. LLMs dapat menghasilkan halusinasi, menyederhanakan ketidakpastian, melewatkan isyarat kontekstual, atau mereproduksi narasi pasien yang stereotip. Mahasiswa tahap awal mungkin belum mampu mengenali ketika pasien AI memberikan respons yang tidak realistis atau ketika penjelasan diagnostik tampak benar tetapi sebenarnya keliru.

Lebih penting lagi, penalaran klinis bukan hanya kemampuan kognitif. Penalaran klinis juga merupakan praktik sosial dan moral yang dibentuk melalui perjumpaan dengan pasien nyata, yang kisahnya sering tidak lengkap, emosional, dan tidak selalu dapat diklasifikasikan secara rapi. Oleh karena itu, simulasi AI sebaiknya mempersiapkan mahasiswa untuk pelayanan pasien, bukan menggantikannya. Urutan kurikulum yang paling aman adalah menempatkan latihan berbasis AI sebagai persiapan, pendukung, dan kemudian pembanding terhadap pengalaman klinis langsung yang disupervisi.

VALIDITAS, KESETARAAN, DAN BIAS

Pendidik tidak dapat menyerahkan persoalan validitas kepada vendor. Asesmen berbasis AI hanya dapat dianggap valid untuk tujuan, kelompok peserta didik, konteks, dan keputusan tertentu. Masters (2023) mengidentifikasi sejumlah isu yang perlu masuk dalam diskusi institusional, termasuk persetujuan, privasi, kepemilikan data, transparansi, tanggung jawab, bias, dan beneficence. Isu-isu ini bukan sekadar abstraksi etis. Semuanya menentukan apakah mahasiswa dapat mempercayai sistem dan apakah keputusan asesmen dapat dipertanggungjawabkan.

Dua masalah pengukuran sangat relevan. Pertama, construct underrepresentation. LLM yang menilai catatan klinis dapat mengevaluasi jejak tertulis dari penalaran, tetapi tidak dapat secara langsung mengamati proses berpikir, interaksi di samping tempat tidur pasien, atau dimensi emosional dan etis dalam pelayanan. Kedua, *construct-irrelevant variance*. Mahasiswa yang menulis dalam bahasa Inggris, menggunakan idiom klinis Barat yang familier bagi model, atau mendeskripsikan penyakit yang banyak terdapat dalam data pelatihan model dapat memperoleh keuntungan dibandingkan mahasiswa yang bekerja dalam bahasa atau konteks epidemiologi berbeda, seperti di Indonesia. Program di Asia Tenggara, Afrika, Amerika Latin, dan berbagai wilayah dengan sumber daya bervariasi tidak dapat mengasumsikan bahwa model yang divalidasi di tempat lain akan otomatis valid secara lokal.

Bias juga telah terdokumentasi dengan baik. Zack et al. (2024) menemukan bahwa GPT-4 dapat melanggengkan bias ras dan gender dalam pembuatan kasus pendidikan kedokteran, penalaran diagnostik, perencanaan terapi, dan penilaian subjektif terhadap pasien. Omiye et al. (2023) menunjukkan bahwa beberapa LLM komersial mereproduksi asumsi medis berbasis ras yang berbahaya. Dalam pendidikan, bias semacam ini membentuk apa yang berulang kali dilihat mahasiswa sebagai “normal”. Jika kasus yang dihasilkan AI terlalu sering menggambarkan kelompok tertentu sebagai kompleks, tidak patuh, atau berisiko, sementara kelompok lain kurang terwakili, mahasiswa dapat menyerap prototipe klinis yang menyimpang. Oleh karena itu, program pendidikan perlu mengaudit kasus yang dihasilkan AI untuk menilai keseimbangan

ISSN Online: 2986-0164
ISSN Cetak: 2986-2116

demografis, relevansi lokal, dan potensi asumsi yang merugikan sebelum digunakan dalam pembelajaran.

Privasi pasien merupakan permasalahan lain yang perlu dipertimbangkan. Ketika alat AI menganalisis catatan klinis pasien, data pribadi pasien dapat berpindah ke sistem pihak ketiga. Hal ini membutuhkan tata kelola yang jelas, de-identifikasi, persetujuan institusional, atau *hosting* lokal. Jika status hukum belum jelas, pendidik sebaiknya menggunakan materi sintetis atau data yang telah dideidentifikasi, bukan rekam klinis pasien secara langsung.

TANTANGAN DALAM ADOPSI

Hambatan adopsi AI bersifat praktis sekaligus etis. Biaya merupakan hambatan yang paling terlihat. Penggunaan model LLMs dalam volume besar dapat menjadi mahal, dan program dengan akses internet yang tidak stabil, mekanisme pembayaran institusional yang terbatas, atau aturan berbagi data yang ketat menghadapi kendala yang mungkin tidak terlihat di institusi dengan sumber daya tinggi. Wang et al. (2025) mencatat bahwa publikasi dan adopsi AI terkonsentrasi pada negara berpenghasilan tinggi, sebagian karena kendala biaya, privasi, dan infrastruktur. Kondisi ini penting bagi institusi yang harus membuat keputusan kurikuler tanpa anggaran AI atau tim teknis khusus.

Kesiapan dosen juga merupakan tantangn lain yang tidak kalah pentingnya. Banyak pendidik klinis belum memiliki cukup waktu untuk memahami cara kerja LLMs saat ini. Tanpa literasi dasar, mereka dapat menghindari AI sepenuhnya atau justru menggunakannya secara terlalu percaya diri. Keduanya tidak ideal bagi mahasiswa.

Integritas akademik membutuhkan kebijakan yang cermat. Alat yang sama yang membantu menyediakan umpan balik juga dapat menghasilkan tugas mahasiswa, sementara perangkat deteksi AI tidak dapat terlalu diandalkan. Liang et al. (2023) menunjukkan bahwa GPT detectors dapat salah mengklasifikasikan tulisan non-penutur asli bahasa Inggris sebagai tulisan yang dihasilkan AI. Hal ini membuat keputusan pelanggaran akademik berbasis detektor menjadi sangat berisiko. Institusi membutuhkan kebijakan spesifik berdasarkan jenis tugas, yang menjelaskan kapan AI boleh digunakan, wajib digunakan, atau tidak boleh digunakan. Mahasiswa juga perlu diminta mengungkapkan penggunaan AI dengan cara yang mendukung pembelajaran, bukan sekadar pengawasan.

Tata kelola dan keberlanjutan sering kali kurang diperhatikan. Jika AI berkontribusi pada keputusan berisiko tinggi, harus ada pihak yang bertanggung jawab terhadap kesalahan, proses banding, dan perubahan performa model. Alur kerja yang divalidasi dengan satu model dapat berubah ketika vendor memperbarui sistemnya. Proyek percontohan juga dapat berhenti ketika dosen penggeraknya tidak lagi terlibat. Penggunaan yang berkelanjutan membutuhkan *version control* untuk prompt, revalidasi berkala, dokumentasi telaah manusia, dan batas yang jelas antara dukungan formatif dan keputusan sumatif.

PILIHAN GRATIS DAN TERJANGKAU UNTUK KEBUTUHAN PENDIDIKAN

Kendala biaya tidak seharusnya menghentikan eksperimen yang dirancang secara hati-hati. Versi konsumen dari ChatGPT, Claude, Gemini, Copilot, dan Mistral dapat membantu pendidik membuat prototipe prompt kasus, menyusun draft soal, menghasilkan contoh umpan balik, dan mensimulasikan percakapan singkat dengan pasien. Alat-alat ini sebaiknya digunakan dengan materi sintetis atau data

ISSN Online: 2986-0164
ISSN Cetak: 2986-2116

yang telah dideidentifikasi. Model open-source seperti Llama, Mistral, Qwen, Gemma, dan DeepSeek membuka peluang tambahan untuk hosting lokal atau penggunaan institusional berbiaya rendah. Model *open-source* dapat meningkatkan kesetaraan dalam AI kesehatan dengan mengurangi ketergantungan pada sistem pembiayaan yang mahal (Wu et al., 2026). Bahkan, model open-source juga dapat mendukung pendidikan kedokteran tingkat lanjut, seperti radiologi (Ray, 2024).

Bagi banyak program, titik awal yang paling realistis bukanlah membangun platform AI khusus, tetapi merancang alur kerja kecil yang dapat dibalik atau dihentikan dengan mudah. Misalnya, dosen dapat menggunakan AI untuk menyusun draft kasus, memvalidasinya terhadap blueprint, mengujinya pada mahasiswa, lalu merevisinya berdasarkan data performa. Prinsip kuncinya adalah kedaulatan data dan reversibilitas. Informasi pasien tidak boleh dimasukkan ke dalam alat AI tanpa persetujuan yang sesuai, dan materi inti seperti rubrik, bank kasus, serta blueprint asesmen harus tetap disimpan dalam format yang tidak bergantung pada satu vendor atau model. Pendekatan ini memungkinkan pendidik belajar dari AI tanpa membangun ketergantungan pada satu produk tertentu.

PERUBAHAN PADA PERAN PENDIDIK

AI tidak membuat pendidik menjadi kurang penting. Sebaliknya, AI membuat bentuk keahlian pendidikan yang berbeda menjadi lebih terlihat. Pendidik semakin berperan sebagai arsitek asesmen yang merancang blueprint, mendefinisikan konstruk, dan menentukan bukti apa yang diperlukan untuk menilai kompetensi. Mereka juga menjadi supervisor umpan balik yang mengaudit komentar AI dan membantu mahasiswa menentukan langkah perbaikan berikutnya. Selain itu, pendidik menjadi penjaga pengalaman klinis autentik. Ketika AI membuat latihan simulasi lebih mudah disediakan, perjumpaan nyata dengan pasien, keluarga, tim kesehatan, ketidakpastian, dan tanggung jawab klinis menjadi semakin bernilai.

Pengembangan dosen tidak cukup hanya berupa demonstrasi alat yang menarik. Pendidik di bidang kesehatan memerlukan literasi AI yang cukup untuk menyusun *prompt*, mengenali halusinasi, mengevaluasi umpan balik secara kritis, dan menjelaskan keterbatasan AI kepada mahasiswa. Tolentino et al. (2024) menunjukkan bahwa pendidikan AI semakin dipandang sebagai persoalan desain kurikulum bagi mahasiswa, residen, dan dokter. Lebih lanjut, Ng et al. (2023) mengusulkan bahwa klinisi perlu dipersiapkan sebagai pengguna, penerjemah, atau pengembang AI, dengan kedokteran berbasis bukti sebagai fondasi. Kerangka ini berguna karena menempatkan kritik terhadap AI sebagai bagian dari penalaran profesional, bukan sebagai keterampilan teknis yang terpisah.

Komite kurikulum atau unit pendidikan kedokteran perlu memulai dengan mengidentifikasi capaian pembelajaran mana yang terancam oleh AI, mana yang dapat diperkuat oleh AI, dan mana yang kini memerlukan kompetensi penggunaan AI. Mahasiswa dapat diberi tugas berisiko rendah, seperti mengkritisi keluaran AI, membandingkannya dengan bukti, dan menjelaskan bagaimana mereka merevisinya. Sementara itu, keputusan berisiko tinggi mengenai kompetensi tetap perlu berbasis asesmen yang terawasi, teramati, atau dapat dipertahankan secara lisan sampai tersedia bukti validitas yang lebih kuat.

Penanggung jawab asesmen perlu menetapkan satu prinsip dasar: keluaran AI tidak boleh memengaruhi keputusan formal

ISSN Online: 2986-0164
ISSN Cetak: 2986-2116

sebelum ditambatkan pada rubrik yang jelas dan diuji dalam konteks lokal. AI mungkin bekerja baik di satu lingkungan, tetapi belum tentu menghasilkan penilaian yang sama adilnya pada kelompok bahasa, tingkat mahasiswa, topik klinis, atau gaya dokumentasi yang berbeda. Karena itu, bahkan studi validasi kecil tetap bernilai. Ia dapat menjadi semacam "uji kewarasan awal" sebelum institusi mempercayakan keputusan yang berdampak pada mahasiswa kepada sistem berbasis AI.

Unit pengembangan sumber daya manusia pada institusi pendidikan juga perlu bergerak lebih jauh dari sekadar mengenalkan alat. Yang dibutuhkan bukan hanya pelatihan tentang cara menggunakan AI, tetapi literasi praktis untuk memahami kapan AI membantu, kapan AI menyesatkan, dan bagaimana keluarannya harus dikritisi. Integrasi AI ke dalam alur kerja pembelajaran perlu dirancang agar tidak menambah beban administratif baru bagi dosen. Pada saat yang sama, kemampuan mengkritisi AI perlu diajarkan sebagai bagian dari kebiasaan penalaran klinis, karena calon dokter tidak hanya perlu menggunakan teknologi, tetapi juga harus mampu mempertanyakan kualitas, relevansi, dan konsekuensi dari informasi yang dihasilkan teknologi tersebut.

Untuk konteks dengan sumber daya yang bervariasi, jalur yang paling aman adalah memulai secara hati-hati, lokal, dan formatif. AI sebaiknya terlebih dahulu digunakan untuk mendukung latihan, umpan balik, dan refleksi, bukan langsung untuk keputusan sumatif yang berisiko tinggi. Adopsi yang lebih luas baru layak dilakukan ketika akurasi, keadilan, privasi, dan keberlanjutan telah dapat ditunjukkan. Dengan demikian, institusi tidak sekadar mengikuti arus teknologi, tetapi membangun kemampuan untuk menggunakannya secara bertanggung jawab.

HARAPAN KEDEPAN

Tahap berikutnya dari AI dalam pendidikan kedokteran kemungkinan tidak akan berlangsung secara merata. Model multimodal dapat membuat pasien virtual menjadi lebih realistis melalui integrasi teks, suara, gambar, dan video. Sistem yang menggunakan *retrieval-augmented generation* dan kurasi klinis juga berpotensi mengurangi halusinasi serta meningkatkan performa pada bidang-bidang spesifik. Pada saat yang sama, kerangka asesmen dalam pendidikan kedokteran akan terus berubah ketika kemampuan menggunakan AI secara kritis dan bertanggung jawab mulai dipandang sebagai bagian dari kompetensi profesional, bukan sekadar jalan pintas teknologi.

Namun, inti pendidikan kedokteran tetap berada pada pertimbangan manusia. AI dapat membantu mahasiswa berlatih lebih sering, menerima umpan balik lebih cepat, dan menelaah penalaran klinis secara lebih mendalam. Akan tetapi, AI tidak dapat menggantikan tanggung jawab pendidik dalam menjaga validitas asesmen, keadilan, integritas akademik, pembentukan identitas profesional, dan orientasi pembelajaran yang berpusat pada pasien. Karena itu, pendidik kedokteran perlu menentukan secara cermat apa yang dapat diotomatisasi, apa yang harus disupervisi, dan apa yang harus tetap dipertahankan sebagai pengalaman belajar yang bersifat manusiawi dan relasional.

Jika digunakan dengan tepat, AI dapat memperluas jangkauan pengajaran yang bermutu. Jika digunakan tanpa desain dan pengawasan yang memadai, AI justru dapat mempercepat asesmen yang lemah, memperluas umpan balik yang bias, dan mengurangi kedalaman pengalaman klinis. Dengan demikian, kualitas integrasi AI

ISSN Online: 2986-0164
ISSN Cetak: 2986-2116

dalam pendidikan kedokteran tidak hanya ditentukan oleh kecanggihan model yang digunakan, tetapi terutama oleh pendidik yang merancang, memvalidasi, mengawasi, dan mengajarkan penggunaannya.

REFERENSI

Artsi, Y., Sorin, V., Konen, E., Glicksberg, B. S., Nadkarni, G., & Klang, E. (2024). Large language models for generating medical examinations: systematic review. *BMC Medical Education*, *24*(1), 354. https://doi.org/10.1186/s12909-024-05239-y

Aster, A., Laupichler, M. C., Rockwell-Kollmann, T., Masala, G., Bala, E., & Raupach, T. (2025). ChatGPT and Other Large Language Models in Medical Education: Scoping Literature Review. *Medical Science Educator*, *35*(2), 1015–1030. https://doi.org/10.1007/s40670-024-02206-6

Brügge, E., Ricchizzi, S., Arenbeck, M., Keller, M. N., Schur, L., Stummer, W., Holling, M., Lu, M. H., & Darici, D. (2024). Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Medical Education*, *24*(1), 1391. https://doi.org/10.1186/s12909-024-06399-7

Çiçek, F. E., Ülker, M., Özer, M., & Kıyak, Y. S. (2025). ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. *Postgraduate Medical Journal*, *101*(1195), 458–463. https://doi.org/10.1093/postmj/qgae170

Cross, J., Kayalackakom, T., Robinson, R. E., Vaughans, A., Sebastian, R., Hood, R., Lewis, C., Devaraju, S., Honnavar, P., Naik, S., Joseph, J., Anand, N., Mohammed, A., Johnson, A., Cohen, E., Adeniji, T., Nnenna Nnaji, A., & George, J. E. (2025). Assessing ChatGPT's Capability as a New Age Standardized Patient: Qualitative Study. *JMIR Medical Education*, *11*, e63353. https://doi.org/10.2196/63353

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024). Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open*, *7*(10), e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969

Gordon, M., Daniel, M., Ajiboye, A., Uraiby, H., Xu, N. Y., Bartlett, R., Hanson, J., Haas, M., Spadafore, M., Grafton-Clarke, C., Gasiea, R. Y., Michie, C., Corral, J., Kwan, B., Dolmans, D., & Thammasitboon, S. (2024). A scoping review of artificial intelligence in medical education: BEME Guide No. 84. *Medical Teacher*, *46*(4), 446–470. https://doi.org/10.1080/0142159X.2024.2314198

Hudon, A., Phan, V., Charlin, B., & Wittmer, R. (2025). Teaching Clinical

ISSN Online: 2986-0164
ISSN Cetak: 2986-2116

Reasoning in Health Care Professions Learners Using AI-Generated Script Concordance Tests: Mixed Methods Formative Evaluation. *JMIR Formative Research*, *9*, e76618. https://doi.org/10.2196/76618

Jamieson, A. R., Holcomb, M. J., Dalton, T. O., Campbell, K. K., Vedovato, S., Shakur, A. H., Kang, S., Hein, D., Lawson, J., Danuser, G., & Scott, D. J. (2024). Rubrics to Prompts: Assessing Medical Student Post-Encounter Notes with AI. *NEJM AI*, *1*(12). https://doi.org/10.1056/AIcs2400631

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, *4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Lucas, H. C., Upperman, J. S., & Robinson, J. R. (2024). A systematic review of large language models and their implications in medical education. *Medical Education*, *58*(11), 1276–1285. https://doi.org/10.1111/medu.15402

Luengo Vera, C., Ferro Picon, I., del Val Nunez, M. T., Gomez Gandia, J. A., Borrego, A. G., Romero Tapia, J. A., & Gomez, A. T. (2025). Evaluating Large Language Models on the Spanish Medical Intern Resident (MIR) Examination 2024/2025: A Comparative Analysis of Clinical Reasoning and Knowledge Application. *ArXiv*. https://doi.org/10.48550/arXiv.2503.00025

Masters, K., MacNeil, H., Benjamin, J., Carver, T., Nemethy, K., Valanci-Aroesty, S., Taylor, D. C. M., Thoma, B., & Thesen, T. (2025). Artificial Intelligence in Health Professions Education assessment: AMEE Guide No. 178. *Medical Teacher*, *47*(9), 1410–1424. https://doi.org/10.1080/0142159X.2024.2445037

Ng, F. Y. C., Thirunavukarasu, A. J., Cheng, H., Tan, T. F., Gutierrez, L., Lan, Y., Ong, J. C. L., Chong, Y. S., Ngiam, K. Y., Ho, D., Wong, T. Y., Kwek, K., Doshi-Velez, F., Lucey, C., Coffman, T., & Ting, D. S. W. (2023). Artificial intelligence education: An evidence-based medicine approach for consumers, translators, and developers. *Cell Reports Medicine*, *4*(10), 101230. https://doi.org/10.1016/j.xcrm.2023.101230

Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. *Npj Digital Medicine*, *6*(1), 195.

ISSN Online: 2986-0164
ISSN Cetak: 2986-2116

https://doi.org/10.1038/s41746-023-00939-z

Ray, P. P. (2024). Opening doors for open-source large language models in radiology education. *Radiologia Brasileira*, *57*, e20240037. https://doi.org/10.1590/0100-3984.2024.0037

Roveta, A., Castello, L. M., Massarino, C., Francese, A., Ugo, F., & Maconi, A. (2025). Artificial Intelligence in Medical Education: A Narrative Review on Implementation, Evaluation, and Methodological Challenges. *AI*, *6*(9), 227. https://doi.org/10.3390/ai6090227

Schaye, V., DiTullio, D., Guzman, B. V., Vennemeyer, S., Shih, H., Reinstein, I., Weber, D. E., Goodman, A., Wu, D. T. Y., Sartori, D. J., Santen, S. A., Gruppen, L., Aphinyanaphongs, Y., & Burk-Rafel, J. (2025). Large Language Model-Based Assessment of Clinical Reasoning Documentation in the Electronic Health Record Across Two Institutions: Development and Validation Study. *Journal of Medical Internet Research*, *27*, e67967. https://doi.org/10.2196/67967

Tolentino, R., Baradaran, A., Gore, G., Pluye, P., & Abbasgholizadeh-Rahimi, S. (2024). Curriculum Frameworks and Educational Programs in AI for Medical Students, Residents, and Practicing Physicians: Scoping Review. *JMIR Medical Education*, *10*, e54793. https://doi.org/10.2196/54793

Wang, H., Shan, W., Liu, R., & Wang, Z. (2025). Can large language models serve as digital assistants for medical undergraduates? A bibliometric mapping and scoping analysis of the medical-education literature. *Digital Health*, *11*. https://doi.org/10.1177/20552076251390280

Wu, S., Zou, W., Tu, J., Wang, C., Jin, C., Liao, J., Tang, K.-C., Liu, N., & Hao, C. (2026). Open-Source Large Language Models and AI Health Equity: A Health Service Triangle Model Perspective. *Journal of Medical Internet Research*, *28*, e86769. https://doi.org/10.2196/86769

Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., Abdulnour, R.-E. E., Butte, A. J., & Alsentzer, E. (2024). Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, *6*(1), e12–e22. https://doi.org/10.1016/S2589-7500(23)00225-X