

Artificial Intelligence in Medical Education: Applications, Reported Effects, and Evidence Gaps

AI-Assisted Research Pipeline

Abstract

Background: Artificial intelligence (AI), especially large language models (LLMs), has entered medical education quickly, but the evidence on its educational effects is unclear. **Purpose:** This report examines how AI is being applied in medical education and what effects on learners, educators, and curriculum outcomes have been reported. **Method:** A structured literature search produced a corpus of reviews, position papers, benchmark studies, and one student survey. Synthesis relied on abstract-level information. Only a few sources are specific to medicine or health professions education. **Findings:** The strongest direct evidence shows that ChatGPT performs at or near the passing threshold on USMLE examinations. Evidence on what learners gain from using such tools is largely absent. Assessment integrity, educator readiness, AI literacy, hallucination, explainability, and governance are discussed mainly conceptually or in general higher education. No retrieved source addresses AI-enhanced simulation or virtual patients in medical training. **Implications:** Conclusions about effects on learners, educators, and curricula remain provisional. Priorities include controlled educational studies in medical populations, targeted work on simulation and feedback, and empirical study of ethics and policy.

Keywords: artificial intelligence, medical education, large language models, ChatGPT, assessment, AI literacy, ethics

1 Introduction

Artificial intelligence now spans clinical practice, research, and education. In healthcare, AI is described as transforming the practice of medicine (Bajwa et al., 2021; Alowais et al., 2023). In education, LLM-based chatbots such as ChatGPT have prompted both enthusiasm and concern (Kasneci et al., 2023; Tlili et al., 2023). Medical education is a demanding setting for these tools. Curricula depend on large bodies of knowledge, on the development of clinical

reasoning, and on high-stakes assessment. The end goal is safe patient care, so errors in training can have consequences beyond the classroom.

This report addresses the following research question: *How is artificial intelligence being applied in medical education, and what are its reported effects on learners, educators, and curriculum outcomes?* Five subtopics structured the search: AI-based assessment and feedback, LLMs in medical curricula, AI-enhanced simulation and virtual patients, educator readiness and AI literacy, and ethical and governance considerations.

1.1 Scope and Nature of the Evidence

The retrieved sources are mostly general higher-education or ChatGPT-in-education papers. Only a small subset is medical or health-specific: Kung et al. (2023), Gilson et al. (2023), Sallam (2023), Alowais et al. (2023), Bajwa et al. (2021), Tjoa and Guan (2020), and Kelly et al. (2019). Of these, only Kung et al. (2023) and Gilson et al. (2023) directly test AI in a medical-education context, and Sallam (2023) reviews healthcare education. The others concern clinical AI or explainability. Several bibliography entries carried subtopic labels that did not match their content, so each source is treated here according to what it actually addresses. The synthesis rests on abstract-level information, and claims are stated at the level the abstracts support. Where a statement extends beyond what a source reports, it is marked as inference.

The sections that follow cover five themes. Table 1 summarizes the type and strength of evidence for each.

Table 1: Evidence Base by Theme

Theme	Dominant evidence type	Medical-specific evidence
LLM performance and curricular use	Benchmark testing; reviews; position papers	Present (exam performance; one healthcare-education review)
Assessment, feedback, and integrity	Conceptual and discussion papers; one survey of methods	None retrieved
Simulation and virtual patients	General AI-in-education reviews	None retrieved
Readiness and AI literacy	Systematic and exploratory reviews; one survey	None retrieved
Reliability, ethics, and governance	Surveys; opinion and position pieces	Partial (explainability, clinical deployment)

Note. Classification is based on abstract-level information. Sources for each theme are cited in the corresponding section.

2 Large Language Models in Medical Curricula

2.1 Performance on Licensing Examinations

The most direct evidence concerns what LLMs can do on medical examination items. Kung et al. (2023) evaluated ChatGPT on the three United States Medical Licensing Examination (USMLE) steps (Step 1, Step 2CK, and Step 3). The model performed at or near the passing threshold on all three without specialized training or reinforcement, and its explanations showed a high level of concordance and insight. A second evaluation covered USMLE Step 1 and Step 2 questions and framed its results as implications for medical education and knowledge assessment (Gilson et al., 2023). The abstract available for that study was truncated, so detailed accuracy figures are not reported here.

These studies show that a general-purpose LLM can handle knowledge-recall and exam-style tasks at a level comparable to the passing standard. They do not show that learners who use such tools learn more, reason better, or perform better clinically. Capability evidence is therefore ahead of learning-outcome evidence. In addition, the results reflect early model versions and may not hold for current systems (Kung et al., 2023; Gilson et al., 2023).

2.2 Reviews of Healthcare Education Applications

A systematic review of ChatGPT in healthcare education, research, and practice describes its utility as promising but conditional on proactively addressing valid concerns (Sallam, 2023). This framing of promise tempered by concern recurs across the broader literature (Kasneci et al., 2023; Tlili et al., 2023; Dwivedi et al., 2023).

2.3 Benefits Claimed in General Education

General education sources add a set of claimed benefits. They include personalized, real-time support (Farrokhnia et al., 2023), increased engagement, collaboration, and accessibility (Cotton et al., 2023), and wider teaching and learning benefits (Baidoo-Anu and Ansah, 2023). Other work presents such benefits as opportunities alongside challenges (Kasneci et al., 2023; Tlili et al., 2023). A rapid review covering only the first three months after ChatGPT’s release (Lo, 2023) shows how early much of this literature is. None of these sources is specific to medicine. Their relevance to medical curricula is by analogy, and these claims are best read as perspectives, not demonstrated effects in medical learners.

3 Assessment, Feedback, and Academic Integrity

Strong LLM performance on examinations raises the question of whether traditional assessment formats remain valid. This debate is framed for higher education in general, not for medicine (Rudolph et al., 2023). Academic honesty and plagiarism are identified as the main assessment-side risks, alongside the engagement benefits noted above (Cotton et al., 2023). The sources agree that assessment design must adapt (Rudolph et al., 2023; Cotton et al., 2023).

One tension runs through this literature. The capability that makes LLMs useful as tutors, producing fluent and plausible answers on demand, is the same capability that threatens assessment integrity (Farrokhnia et al., 2023; Cotton et al., 2023). The benefit and the threat therefore cannot be cleanly separated.

Educational data mining and learning analytics offer an established, pre-LLM toolkit for learner assessment and feedback (Romero and Ventura, 2020). No retrieved source applies this toolkit to medical learners. More notably, no retrieved source tests AI-generated feedback or automated grading of clinical reasoning with medical students or residents, although this was a targeted subtopic of the research plan. Whether AI-based feedback in medicine is valid, reliable, and educationally useful remains unanswered by this corpus.

4 AI-Enhanced Simulation and Virtual Patients: An Evidence Gap

Simulation and virtual patients are widely regarded as promising applications of AI in clinical training, yet this subtopic is effectively unsupported by the retrieved sources. General reviews of AI in education describe applications in administration, instruction, and learning (Chen et al., 2020), and trends and challenges across 2010 to 2020 (Zhai et al., 2021). A further review notes that expectations often rest on misunderstanding of what current AI can do (Holmes and Tuomi, 2022). None of these addresses medical simulation or virtual patients in its abstract.

Medical AI overviews describe AI transforming clinical practice (Bajwa et al., 2021; Alowais et al., 2023). They concern practice, not training, so any link to simulation-based teaching would be inference. This section therefore reports a gap, not a finding. The absence of evidence may reflect either a genuine literature gap or a failure of the search queries. Targeted follow-up searches (for example, combining AI with “virtual patient” and “surgical simulation”) are needed to distinguish these explanations.

5 Learner and Educator Readiness and AI Literacy

A systematic review of AI in higher education asks “where are the educators?” and points to limited educator involvement and unclear pedagogical use of AI (Zawacki-Richter et al., 2019). A later review of 138 articles published between 2016 and 2022 maps the state of the field (Crompton and Burke, 2023). These reviews concern higher education broadly, not medical faculty.

Student attitudes appear receptive. A survey of 399 undergraduate and postgraduate students in Hong Kong found generally positive attitudes toward generative AI, alongside perceived challenges (Chan and Hu, 2023). The sample is non-medical and drawn from one region, so generalization to medical students, residents, or other regions is not supported.

AI literacy is being defined as a core competency (Ng et al., 2021), and prompt engineering and critical thinking have been proposed as complementary skills (Walter, 2024). The sources agree that literacy and educator capacity are prerequisites for effective use (Zawacki-Richter et al., 2019; Ng et al., 2021; Walter, 2024). Student enthusiasm (Chan and Hu, 2023) sits alongside the gap in educator preparedness (Zawacki-Richter et al., 2019). This mismatch could create risk if students adopt tools faster than faculty can guide them. However, this is an inference, and no medical-specific data test it.

6 Reliability, Ethics, and Governance

Reliability is a central concern for any educational use of LLMs. LLMs are prone to hallucination, generating plausible but nonfactual content (Huang et al., 2024). In medical training this matters because learners may lack the expertise to detect errors. That implication is an inference and not a finding of the cited survey.

Explainability is highlighted as important in high-accountability fields such as medicine (Tjoa and Guan, 2020). AI research in healthcare has produced few successful clinical deployments, and the challenges to impact are well documented (Kelly et al., 2019). This suggests, by extension, that similar adoption barriers could affect educational uses, although that extension is not tested in the source.

Generative AI also raises ethical and legal challenges with both positive and negative impacts (Dwivedi et al., 2023), and the healthcare-education review likewise lists valid concerns (Sallam, 2023). The sources agree that the risks are real. The AI-in-medicine sources are predominantly cautionary, whereas the education sources are more optimistic about benefits (Kelly et al., 2019; Tjoa and Guan, 2020; Kasneci et al., 2023; Baidoo-Anu and Ansah, 2023). This difference may reflect the safety-critical nature of medicine, though the sources

do not test the explanation. No empirical studies of bias, fairness, or institutional policy in medical training were retrieved. Evidence here consists of surveys and opinion or position pieces.

7 Discussion and Evidence Gaps

7.1 Synthesis

Taken together, the corpus supports a narrow factual claim and a broad set of expectations. The factual claim is that LLMs can reach approximately passing performance on medical licensing examination items (Kung et al., 2023; Gilson et al., 2023). The expectations concern personalization, engagement, and efficiency, together with risks to integrity, reliability, and equity (Sallam, 2023; Kasneci et al., 2023; Farrokhnia et al., 2023; Dwivedi et al., 2023). The recurring “promise and concern” framing suggests the field is still in an early, anticipatory phase.

Timing reinforces this reading. Much of the LLM literature clusters in 2023, soon after ChatGPT’s release (Lo, 2023), while earlier AI-in-education reviews predate generative AI (Zawacki-Richter et al., 2019; Chen et al., 2020; Zhai et al., 2021; Romero and Ventura, 2020). Overall evidence quality is low to moderate. Most sources are narrative or rapid reviews, position papers, opinion pieces, or a single survey. No randomized or controlled educational intervention in medical learners appears in the corpus.

7.2 Effects on Learners, Educators, and Curricula

On *learners*, the corpus offers no measurement of knowledge retention, clinical reasoning, or skill acquisition after AI-assisted instruction in medicine. Student attitudes are documented only in a non-medical sample (Chan and Hu, 2023). On *educators*, evidence is limited to general higher education and conceptual work on literacy (Zawacki-Richter et al., 2019; Ng et al., 2021; Walter, 2024). On *curriculum outcomes*, the sources discuss implications for assessment design and AI literacy (Rudolph et al., 2023; Cotton et al., 2023; Ng et al., 2021), but report no measured curricular outcomes.

7.3 Identified Gaps

- **Empirical gap.** No retrieved study measures learner outcomes from AI-assisted instruction in medical education.

- **Subtopic gap.** AI-driven simulation, virtual patients, and surgical training are not covered by any medically relevant source.
- **Population gap.** Medical students, residents, and faculty are largely absent from the readiness and literacy evidence. The one student survey comes from a single region (Chan and Hu, 2023).
- **Methodological gap.** The evidence leans on benchmark testing and reviews. Qualitative, mixed-methods, and longitudinal designs are missing.
- **Assessment gap.** There is no evidence on validated AI-based feedback or grading of clinical reasoning.
- **Ethics and governance gap.** Bias and fairness in training data and institutional policy are not empirically examined in medical training.
- **Temporal gap.** Exam-performance results reflect early model versions and may not hold for current systems (Kung et al., 2023; Gilson et al., 2023).

7.4 Limitations of This Report

This report has several limitations. First, the synthesis draws on abstract-level information only, so nuance in full texts may have been missed. Second, the corpus is small and contains few medical-specific sources, and some entries were originally tagged with subtopics that did not match their content. Third, the simulation subtopic may be underrepresented because of the search queries and not because the literature is absent. Fourth, several sources are opinion or position papers (Dwivedi et al., 2023; Kasneci et al., 2023; Farrokhnia et al., 2023) and are treated as perspectives, not empirical findings. Fifth, the report was not reviewed by a human expert.

8 Conclusion

The retrieved literature shows that AI, in the form of LLMs, is technically capable on medical examination tasks and is widely discussed as a tool for medical education (Kung et al., 2023; Gilson et al., 2023; Sallam, 2023). The same literature offers little direct evidence on its effects on learners, educators, or curriculum outcomes. Claims about benefits are mostly extrapolated from general higher education or stated conceptually, and claims about risks, including integrity threats, hallucination, and unresolved governance questions, are likewise

only partly tested in medical settings (Cotton et al., 2023; Huang et al., 2024; Tjoa and Guan, 2020; Kelly et al., 2019).

Three recommendations follow. First, medical educators and institutions should treat current claims about educational benefit as provisional, and should plan assessment design, faculty development, and AI literacy training with that uncertainty in mind (Rudolph et al., 2023; Ng et al., 2021). Second, researchers should prioritize controlled, longitudinal, and mixed-methods studies in medical populations, including studies of AI-generated feedback, simulation, and virtual patients. Third, supplementary targeted searches and empirical work on bias, fairness, and policy are needed before firm conclusions about effects on learners, educators, and curricula can be drawn.

AI Disclosure

This report was produced with AI-assisted research tools. The research pipeline included AI-powered literature search, source verification, evidence synthesis, and report drafting. All cited sources were retrieved from OpenAlex and their bibliographic metadata was checked automatically. The report was not reviewed by a human expert; readers should verify key claims against the original sources.

References

- Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A. O., Alrashed, M. A., Saleh, K. B., Badreldin, H. A., Yami, M. S. A., Harbi, S. A. A., and Albekairy, A. M. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education*, 23(1):689.
- Baidoo-Anu, D. and Ansah, L. O. (2023). Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, 7(1):52–62.
- Bajwa, J., Munir, U., Nori, A., and Williams, B. (2021). Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthcare Journal*, 8(2):e188–e194.
- Chan, C. K. Y. and Hu, W. (2023). Students’ voices on generative ai: perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20(1).

- Chen, L., Chen, P., and Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, 8:75264–75278.
- Cotton, D., Cotton, P. A., and Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in Education and Teaching International*, 61(2):228–239.
- Crompton, H. and Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, 20(1).
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A. A., Raghavan, V., Ahuja, M. K., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., Carter, L. D., Chowdhury, S., Crick, T., Cunningham, S. W., Davies, G. H., Davison, R. M., Dé, R., Dennehy, D., Duan, Y., Dubey, R., Dwivedi, R., Edwards, J. S., Flavián, C., Gault, R. D. C., Grover, V., Hu, M., Janssen, M., Jones, P. K., Junglas, I. A., Khorana, S., Kraus, S., Larsen, K. R. T., Latreille, P., Laumer, S., Malik, T., Mardani, A., Mariani, M., Mithas, S., Mogaji, E., Nord, J. H., O’Connor, S., Okumus, F., Pagani, M., Pandey, N., Papagiannidis, S., Pappas, I. O., Pathak, N., Pries-Heje, J., Raman, R., Rana, N. P., Rehm, S., Navarrete, S. R., Richter, A., Rowe, F., Sarker, S., Stahl, B. C., Tiwari, M. K., van der Aalst, W. M. P., Venkatesh, V., Viglia, G., Wade, M. R., Walton, P., Wirtz, J., and Wright, R. T. (2023). Opinion paper: “so what if chatgpt wrote it?” multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71(71):102642.
- Farrokhnia, M., Banihashem, S. K., Noroozi, O., and Wals, A. E. J. (2023). A swot analysis of chatgpt: Implications for educational practice and research. *Innovations in Education and Teaching International*, 61(3):460–474.
- Gilson, A., Safranek, C. W., Huang, T. Y., Socrates, V., Chi, L., Taylor, R. A., and Chartash, D. (2023). How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, 9:e45312.
- Holmes, W. and Tuomi, I. (2022). State of the art and practice in <scp>ai</scp> in education. *European Journal of Education*, 57(4):542–570.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2024). A survey on hallucination in large language models:

- Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M. J., Weller, J., Kühn, J., and Kasneci, G. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.
- Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. S., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):195.
- Kung, T. H., Cheatham, M., ChatGPT, Medenilla, A., Sillos, C., Leon, L. D., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., and Tseng, V. (2023). Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198.
- Lo, C. K. (2023). What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, 13(4):410.
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., and Qiao, M. S. (2021). Conceptualizing ai literacy: An exploratory review. *Computers and Education Artificial Intelligence*, 2:100041.
- Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, 10(3).
- Rudolph, J., Tan, S., and Tan, S. (2023). Chatgpt: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning Teaching*, 6(1).
- Sallam, M. (2023). Chatgpt utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, 11(6):887.
- Tjoa, E. and Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813.

- Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., and Agyemang, B. (2023). What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education. *Smart Learning Environments*, 10(1).
- Walter, Y. (2024). Embracing the future of artificial intelligence in the classroom: the relevance of ai literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, 21(1).
- Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1).
- Zhai, X., Chu, X., Chai, C. S., Jong, M. S., Starčić, A. I., Spector, M., Liu, J., Jing, Y., and Li, Y. (2021). A review of artificial intelligence (ai) in education from 2010 to 2020. *Complexity*, 2021(1).