



ORIGINAL ARTICLE

Artificial Intelligence in Medical Education: A Narrative Review of Current Applications and Future Perspectives

Paweł Klimas⁵, Piotr Wysokiński¹, Sebastian Szmeja¹, Julia Masalska², Kacper Robert Toczydłowski², Natalia Mazek², Paweł Adam Król², Martyna Michalik³, Maria Mamczak⁴, Stanisław Ferdynand Strzępek⁶

¹Śniadeckiego Voivodeship Hospital in Białystok, ul. M.C. Skłodowskiej 26, 15-278 Białystok, Poland

²Clinical Hospital of the Ministry of Internal Affairs and Administration with the Warmia-Mazury Oncology Centre in Olsztyn

³National Medical Institute of the Ministry of Interior and Administration: Warsaw, Mazovia, PL

⁴Prague Hospital of the Transfiguration of Our Lord: Aleja Solidarności 67 03-401, Warsaw

⁵Św. Wincentego a Paulo Hospital, 1 Wójta Radtkego St., 81-348 Gdynia, Poland

⁶Medical University of Białystok: Białystok, Podlasie, PL

Corresponding author: Paweł Klimas

e-mail: klimas.pawel42@gmail.com, <https://orcid.org/0009-0002-8803-0476>

ARTICLE INFORMATION



Received: 24.07.2026



Revised: 15.08.2026



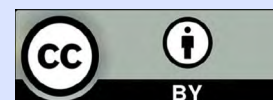
Accepted: 31.08.2026



Article type: Original Article

LICENSE

This work is licensed under Creative Common Attribution License 4.0 International License CC BY 4.0.



Klimas, P., Wysokiński, P., Szmeja, S., Masalska, J., Toczydłowski, K. R., Mazek, N., Król, P. A., Michalik, M., Mamczak, M., Strzępek, S. F. 2026. Artificial Intelligence in Medical Education: A Narrative Review of Current Applications and Future Perspectives. Journal of Education, Health and Sport, 95(1): 73695. DOI: [10.12775/JEHS.2026.95.73695](https://doi.org/10.12775/JEHS.2026.95.73695)

ABSTRAKT

Introduction and purpose. Artificial intelligence has gained growing attention in medical education, especially since ChatGPT appeared in late 2022. Large language models proved capable of reaching or exceeding passing thresholds on USMLE-style questions or official practice materials, which sparked both enthusiasm and concern among educators. This review aims to present how AI tools are currently used across the medical education continuum, what problems they create and where the field may be heading. **Brief description of the state of knowledge.** Machine learning and natural language processing are already being used in preclinical teaching, especially in anatomy and pharmacology, through adaptive quizzes, intelligent tutoring, and question generation. In clinical training, AI-powered virtual patients allow students to practise history-taking and diagnostic reasoning in a safe setting, while AI-enhanced surgical simulators give objective, real-time feedback on technical skills. Large language models are also being tested as item writers for medical examinations, with early studies showing acceptable question quality after expert review. At the same time, serious problems remain: AI systems can produce confident but wrong medical information, student overreliance may weaken independent reasoning, academic integrity is harder to enforce and algorithmic bias may disadvantage certain groups.

Summary. AI offers real opportunities to personalise learning, broaden access, and improve assessment in medical education. Realising these benefits without harm will require clear institutional policies, faculty development, rigorous outcome research, and a commitment to preserving the human side of medicine.

KEYWORDS

artificial intelligence; medical education; machine learning; large language models; clinical reasoning; educational technology

1. Introduction and Purpose

Artificial intelligence is no longer a distant promise in medicine. Over the past five years it has moved from research laboratories into diagnostic imaging suites, electronic health records and more recently to the classroom^{1,2}. The speed of this shift caught many medical educators off guard. When OpenAI released ChatGPT in November 2022, within weeks researchers showed that the system could pass the United States Medical Licensing Examination (USMLE) at or near the passing threshold of roughly 60 percent^{3,4}. The next-generation model, GPT-4, cleared that bar by an even wider margin⁵, and Google's Med-PaLM 2 approached expert-level accuracy on medical question-answering benchmarks^{6,7}. These results made front-page news and forced a reckoning: if a machine can pass the same exam as a graduating medical student, what does that mean for how we teach and test?

The question matters because medical education already faces considerable pressure. The volume of biomedical knowledge doubles faster than ever, yet the length of medical training has hardly changed. Curricula must somehow cover more content while also developing higher-order skills like clinical reasoning, communication, teamwork, professionalism, that cannot be reduced to memorisable facts. Traditional tools such as didactic lectures and multiple-choice examinations have clear limits here⁸. AI-based tools hold the potential to fill some of these gaps: they can tailor content to each learner, give instant feedback at any hour of the day, simulate patient encounters, and take over routine assessment tasks so that faculty time is freed for mentoring and bedside teaching^{9,10}.

At the same time, good AI integration does not happen on its own. These systems sometimes make things up. The technical term for this is "hallucination." They can say false things with the same confidence as true ones¹¹. Students who lack the background to spot errors may absorb false information and students who grow dependent on AI-generated answers may never develop the kind of independent thinking they need at the bedside. Academic dishonesty is harder to detect when the text a student submits is original yet machine-written. Algorithms trained on biased data can silently reproduce the very health disparities that medical education claims to combat¹².

Several recent reviews have examined the use of large language models in assessment¹³, AI in surgical simulation¹⁴, ethical frameworks for AI in health professions education¹⁵. What is still missing is a narrative review that ties these threads together across the full span of medical training, from preclinical science to

residency, while paying equal attention to opportunities and risks. That is the gap this paper tries to fill.

The purpose of this review is threefold: first, to present how AI technologies are currently applied in preclinical education, clinical training and assessment; second, to examine the main challenges and ethical concerns that accompany adoption; and third, to outline directions for future research, policy, and curriculum design. The review draws on peer-reviewed literature in PubMed and Scopus, with emphasis on publications from 2020 onward, supplemented by guidance documents from the Association of American Medical Colleges¹⁶, the American Medical Association¹⁷, and the World Health Organization^{18,19}.

2. Materials and Methods

2.1. Study Design

This study was conducted as a narrative review of the literature on artificial intelligence applications in medical education. A narrative review design was chosen because the field is broad, rapidly evolving, and methodologically heterogeneous, conditions under which a rigid systematic approach would either force the exclusion of much relevant work or produce misleading summary statistics. The review was guided by the Scale for the Assessment of Narrative Review Articles (SANRA) checklist to ensure methodological transparency, even though no formal systematic review protocol was registered²⁰.

2.2. Search Strategy

A literature search was performed between November and December 2025 across three electronic databases: PubMed/MEDLINE, Scopus, and Google Scholar. The search combined terms related to artificial intelligence with terms related to medical education using the Boolean AND operator. The primary search strings included: („artificial intelligence" OR „machine learning" OR „deep learning" OR „large language model" OR „ChatGPT" OR „GPT-4" OR „natural language processing") AND („medical education" OR „medical students" OR „clinical training" OR „undergraduate medical education" OR „residency training"). Additional targeted searches were conducted for specific subtopics using the following term combinations: „AI" AND „anatomy education"; „AI" AND „surgical simulation"; „AI" AND „virtual patients"; „ChatGPT" AND „MCQ generation"; „LLM" AND „USMLE"; „AI" AND „clinical reasoning"; „AI" AND „academic

integrity” AND „medical education”; „AI bias” AND „healthcare”.

In accordance with the editorial requirements of the Journal of Education, Health and Sport, additional searches were performed in the databases of three affiliated journals: Journal of Education, Health and Sport (<https://apcz.umk.pl/JEHS/search/search>), Quality in Sport (<https://apcz.umk.pl/QS/search/search>), and Pedagogy and Psychology of Sport (<https://apcz.umk.pl/PPS/search/search>). The keywords used in these searches included „artificial intelligence,” „medical education,” „e-learning,” „digital education,” „simulation,” and „technology education.”

2.3. Eligibility Criteria

Articles were considered eligible if they met all of the following inclusion criteria: (a) published in a peer-reviewed journal or as an official organisational report; (b) written in English; (c) published between January 2019 and December 2025; (d) focused on the application, evaluation, or ethical implications of artificial intelligence technologies in medical education at any level (undergraduate, postgraduate, or continuing professional development). The lower date boundary of 2019 was chosen because earlier publications largely predate the current generation of AI tools, though seminal works on foundational concepts were included regardless of date when they provided essential context.

Studies were excluded if they: (a) were published only as conference abstracts, editorials, or letters to the editor without original data or substantive analysis; (b) dealt exclusively with AI in clinical practice without any educational component; (c) were non-peer-reviewed preprints, with the exception of one widely cited Microsoft Research report on GPT-4 capabilities⁵, which was included because of its direct relevance and extensive citation in subsequent peer-reviewed literature; (d) were published in languages other than English; (e) were duplicates across databases.

2.4. Study Selection and Data Extraction

Titles and abstracts of retrieved records were screened for relevance to the research questions. Full texts of potentially eligible articles were then read and assessed against the inclusion and exclusion criteria described above. Reference lists of included articles and recent review papers were hand-searched to identify additional relevant publications not captured by the electronic database searches (backward citation tracking). Given the narrative nature of the review, no formal risk-of-bias assessment or quantitative synthesis

was performed. Instead, findings were grouped thematically into the following domains: AI technologies relevant to medical education, preclinical applications, clinical applications, assessment, personalised learning, LLM performance on medical examinations, challenges, and ethical concerns. The final reference list comprises 42 sources, including 37 peer-reviewed journal articles, one preprint, and four organisational guidance documents or reports.

2.5. Limitations of the Search Methodology

Several methodological limitations should be noted. First, because this is a narrative rather than a systematic review, the literature search was not exhaustive, and the selection of sources inevitably reflects the authors’ judgement regarding relevance and quality. Second, the restriction to English-language publications may have excluded relevant work published in other languages. Third, given the very rapid pace of publication in this field, particularly since the release of ChatGPT in late 2022, some studies published after the search period may not be captured. Fourth, no formal quality appraisal tool was applied to individual studies, which means the evidence presented ranges from randomised controlled trials to descriptive reports and expert commentaries. Readers should weigh the cited evidence accordingly.

3. Description of the State of Knowledge

3.1. AI Technologies Relevant to Medical Education

For the clinician-educator who did not train in computer science, a brief orientation may be useful. Machine learning (ML) is a broad category of algorithms that learn patterns from data instead of following hand-coded rules. Deep learning, a subset of ML, uses layered neural networks to handle complex inputs such as medical images or free text. Natural language processing (NLP) deals specifically with human language by understanding it, generating it, and translating between languages¹.

Large language models (LLMs) sit at the intersection of deep learning and NLP. Systems like chatGPT (OpenAI), Claude (Anthropic), Gemini (Google), and LLaMA (Meta) are trained on huge amounts of text and can write clearly and naturally about almost any topic. In the medical domain, Google developed Med-PaLM and Med-PaLM 2 by fine-tuning a general-purpose model on clinical literature. The second

version reached 86.5 percent accuracy on the MedQA benchmark, close to expert physician performance^{6,7}. These models can explain disease mechanisms, suggest differential diagnoses, create clinical case studies, and provide feedback on student answers at scale and without fatigue^{21,22}.

A key distinction is between discriminative and generative AI. Discriminative models classify inputs, for instance, labelling a chest X-ray as normal or abnormal. Generative models create new content: text, images, and synthetic patient records. It is the generative capacity of modern LLMs that makes them so relevant to education, because they can produce teaching materials, exam items and feedback. But that same capacity is what makes hallucinations dangerous: a model that can invent a convincing clinical explanation can just as easily invent a wrong one¹¹.

3.2. AI in Preclinical Medical Education

Preclinical courses like anatomy, physiology, biochemistry and pharmacology form the knowledge base that clinical reasoning rests on. AI is entering each of these disciplines in slightly different ways.

In anatomy, three-dimensional visualisation platforms now incorporate AI that adapts content to student performance²³. ChatGPT has been tested as a conversational anatomy tutor and can answer many standard questions correctly, though its error rate remains high enough that unsupervised use is not advisable²⁴. A recent pilot study found that ChatGPT can create case-based anatomy multiple-choice questions with acceptable quality, as long as the prompts are well designed and the questions are checked by an expert²⁵.

In pharmacology, AI-driven adaptive platforms use algorithms to schedule reviews of previously studied material at intervals calculated to maximise retention and a more individualised version of the spaced-repetition approach already popular with medical students. These systems track response time, confidence ratings, and topic relationships to fine-tune study plans⁹. In preclinical education, available studies and reviews most often highlight anatomy teaching and AI-supported adaptive tutoring or feedback systems, although the evidence remains heterogeneous and insufficient to establish clear superiority over applications in other foundational disciplines.

AI-powered virtual patients add another layer. Unlike static case-based learning, these systems use NLP to hold an actual conversation with the student. The student asks questions, and the virtual patient answers in plain language, creating a more lifelike bridge between basic science and clinical practice^{26,27}.

3.3. AI in Clinical Medical Education

Clinical training is where the stakes are highest and where AI may bring the greatest change. A key goal of clerkships and residency is to teach clinical reasoning: gathering information, considering possible diagnoses and adjusting your thinking as new facts come in. This is among the hardest skills to teach and among the hardest to assess.

Recent studies suggest that GPT-4 can generate plausible differential diagnoses in selected test settings, with performance that in some analyses approached, and in some dimensions exceeded, physician comparators^{31,32}. A randomized study found that doc-

Table 1. Overview of AI applications in medical education

Domain	AI Application	Example Tools	Key Evidence	Reference
Anatomy	Intelligent tutoring, 3D visualization, MCQ generation	ChatGPT, AI-enhanced anatomy platforms	AI-generated MCQs show acceptable psychometrics after expert review	Totlis et al., 2023; Kyriak et al., 2025a
Virtual patients	Conversational AI simulating patient encounters	AI training VR platform	AI virtual patients enable realistic diagnostic practice	Combs et al., 2019; Mergen et al., 2023
Surgical simulation	Real-time performance feedback, gesture recognition	AI-integrated surgical simulators	AI tutoring outperformed remote expert instruction in simulated skill acquisition	Fazlollahi et al., 2022; Park et al., 2022
MCQ generation	Automated item writing using LLMs	ChatGPT, GPT-4	AI items comparable to human-written items when expert-reviewed	Cheung et al., 2023; Cogkun et al., 2025
Personalised learning	Adaptive quizzes, learning analytics, AI tutors	AI-enhanced LMS, chatbots	24/7 tutoring; predictive analytics for at-risk students	Lee et al., 2021; Boscardin et al., 2024

tors changed the way they thought about diagnosis when they used an LLM, though the effect on final accuracy was mixed³³. These findings suggest that AI could help teach clinical reasoning by presenting cases, encouraging learners to generate hypotheses, and giving structured feedback. However, the evidence is still too limited to draw firm conclusions.

Standardised patients remain a cornerstone of clinical education, but they are expensive, need training and can portray only a limited range of conditions. AI-driven virtual patients powered by NLP can simulate a much wider clinical repertoire and are available around the clock. The medical training platform, for example, places students in a virtual reality hospital where they examine an AI-driven patient, order tests, and make treatment decisions²⁷. Every student action is logged, which opens the door to detailed learning analytics³⁴.

In surgical education, AI is already proving its value. Machine-learning classifiers can distinguish novice from expert performance on simulated procedures by analysing instrument movements and tissue interactions³⁵. A randomized trial found that AI tutoring was as effective as expert human teaching for simulated surgical skills in medical students. This could be especially useful where experienced surgical teachers are hard to find²⁸. Computer-vision algorithms embedded in simulators can give real-time feedback on

hand position, force, and procedural efficiency, making practice more structured than the traditional “see one, do one” model¹⁴.

3.5. Performance of LLMs on Medical Examinations

The first benchmark studies were published in early 2023. Kung and colleagues found that ChatGPT (GPT-3.5) scored around 60% on all three USMLE steps, which is roughly the passing level, even without any medical fine-tuning³. Gilson and colleagues confirmed these results and discussed what they meant for knowledge assessment⁴. Shortly after, GPT-4 cleared the bar by a wider margin⁵, and a study directly comparing chatbot answers with medical student responses on free-response clinical reasoning questions found that blinded evaluators rated the AI’s answers higher³⁷.

In specialised medical AI, Google’s Med-PaLM reached 67.6 percent on MedQA6. Its successor, Med-PaLM 2, reached up to 86.5 percent on MedQA, but this result should be interpreted as benchmark-specific progress rather than broad expert-level clinical performance⁷. These numbers raise an obvious question: if AI can pass the same tests we use to certify doctors, are those tests measuring the right things? Some argue the results should push us toward asses-

Table 2. Performance of large language models on medical examinations.

Model	Examination	Year	Score / Result	vs. Humans	Reference
ChatGPT (GPT-3.5)	USMLE (Steps 1, 2CK, 3)	2023	~60% accuracy, at or near passing threshold	Near passing level for medical graduates	Kung et al., 2023
GPT-4	USMLE (all steps)	2023	Exceeded passing threshold by ~20 points	Exceeded average student performance on select items	Nori et al., 2023
Med-PaLM	MedQA (USMLE-style)	2023	67.6% accuracy	Below expert physician level	Singhal et al., 2023
Med-PaLM 2	MedQA (USMLE-style)	2025	86.5% accuracy	High benchmark performance on MedQA; does not establish equivalence to expert clinical reasoning in real-world care	Singhal et al., 2025
GPT-4	Clinical reasoning cases	2023	Higher-rated than student responses	Outperformed medical students	Strong et al., 2023

sing skills AI cannot easily replicate: physical examination, empathic communication, moral judgement, teamwork^{8,15}. Others point out that LLM performance on MCQs may reflect pattern matching more than genuine understanding³¹. What everyone agrees on is that LLMs have serious blind spots. They can hallucinate and sometimes make up references, drug interactions or explanations of disease mechanisms. Their performance also varies across specialties and question types; they often struggle with questions that require image interpretation or the integration of laboratory data^{11,33}.

3.6. Challenges, Risks, and Ethical Concerns

The risks of AI in medical education are no longer theoretical. They are already being seen in classrooms and exam settings. The most immediate concern is academic integrity. Large language models generate new text rather than copy existing material, so traditional plagiarism checkers often fail to detect their use. Although AI-detection tools are available, their performance has been disappointing: false positives can wrongly accuse honest students, while false negatives can allow cheating to go unnoticed³⁶. The AMEE Guide No. 158 on ethical AI use in health professions education recommends that institutions draft explicit

policies on what is and is not acceptable and redesign assignments to test processes rather than products¹⁵.

Hallucinations are a quieter but potentially more dangerous problem. When an LLM states a wrong drug dose or fabricates a clinical guideline with perfect fluency, a student without enough background to notice may internalise the error. The ability to critically assess AI-generated output, often called “AI literacy,” is therefore not just a technical extra, but an essential clinical skill^{38,39}.

Bias in AI systems is highly relevant. A widely cited study in health care showed that an algorithm used for population health management underestimated the needs of Black patients compared with White patients with similar illness burden¹². In an educational context, biased AI could generate case scenarios that overrepresent certain demographics or provide subtly different feedback depending on a student’s writing style. Inclusive design and regular fairness audits are essential³⁹.

Overreliance is a concern that experienced educators voice repeatedly. If students reach for an AI answer every time they face uncertainty, they miss out on the productive struggle that builds clinical expertise. The ability to sit with diagnostic ambiguity and work through a problem step by step is exactly what distinguishes a competent clinician from someone

Table 3. Challenges and proposed solutions for AI integration in medical education

Challenge	Description	Proposed Solution	Reference
Academic integrity	LLMs generate original text, making conventional plagiarism screening insufficient; detector output should not be used as a stand-alone adjudication tool	Redesign assessments; draft clear institutional AI policies; require disclosure rules proportional to task type	Masters et al., 2023; Weber-Wulff et al., 2023
Hallucinations	AI produces confident but wrong medical information	Teach critical appraisal of AI output; integrate AI literacy	Lee et al., 2023; Mondal, 2025
Algorithmic bias	Biased training data may perpetuate health disparities	Diverse training data; regular fairness audits	Obermeyer et al., 2019
Overreliance	Students may skip independent reasoning	Use AI as a supplement; keep unaided assessment formats	Boscardin et al., 2024
Data privacy	AI platforms may exploit student data	Comply with data-protection law; informed consent	WHO, 2021
Regulatory gaps	No binding cross-institutional guidelines exist	Develop policies guided by AAMC, AMA, WHO frameworks	AAMC, 2024; AMA, 2024

who can only follow an algorithm⁸.

Data privacy should not be overlooked. AI platforms that process student learning data may store, analyse or even monetise that information. Compliance with data-protection regulations and genuine informed consent are baseline requirements^{15,18}.

Finally, the regulatory picture is still fragmented. The AAMC has published principles for AI use in medical education¹⁶, the AMA has issued broader guidance on augmented intelligence¹⁷, and the WHO has addressed AI governance for health in general^{18,19}. Binding, cross-institutional standards remain limited, and published literature points to ongoing curricular and pedagogical adaptation needs in AI-related education^{40,41}.

3.7. Future Perspectives

Several developments are likely to shape the next few years. Multimodal AI systems can process text, images, audio, and video together. This means they could act as tutors who discuss a chest X-ray, interpret an ECG, and talk through a clinical case in one session, which is closer to how clinicians actually work^{7,19}.

AI-enhanced Objective Structured Clinical Examinations (OSCEs) are another real possibility. NLP could automate parts of the scoring, virtual standardized patients could populate practice stations and video analysis could assess non-verbal communication. All of this could make OSCEs cheaper, more consistent, and easier to scale^{10,8}.

In continuing professional development, AI could deliver personalized learning modules based on a physician's actual practice patterns and knowledge gaps, turning continuing education from a checkbox exercise into genuinely targeted learning⁴². Perhaps the most important shift will be conceptual. Rather than treating AI as a rival or a gimmick, the idea of the "AI-augmented physician" is gaining traction: a doctor who knows how to use AI, when to trust it, and when to override it. Training this kind of physician means weaving AI literacy into the curriculum from day one, not bolting it on as an elective^{40,16}.

Finally, the evidence base needs to catch up with the hype. Most of what we know about AI in medical education comes from pilot studies, cross-sectional surveys, and descriptive reports. Randomised controlled trials, longitudinal follow-up, and qualitative research exploring the lived experience of students and teachers with AI are all badly needed^{10,28}.

4. Summary

This review has outlined the main ways AI is being used in medical education, including preclinical teaching, clinical training, assessment, and personalised learning, as well as the challenges that come with it. Several key points stand out.

First, large language models now have enough medical knowledge to challenge the usefulness of traditional exam formats. If assessment is meant to measure skills that truly matter in patient care, educators will need to reconsider what those skills are in a world where factual recall can be delegated to a machine^{3,7}.

Second, AI shows real promise in specific domains: generating examination items, providing surgical training feedback, and simulating patient encounters. In each case, however, the quality of the output depends on expert human oversight. AI is a tool, not a replacement for clinical faculty^{28,30}.

Third, the risks of AI, including hallucinations, bias, overreliance, threats to academic integrity, and privacy concerns, are not minor issues. They require active institutional responses, such as clear policies on acceptable AI use, faculty training, fairness audits, and curriculum changes that help students learn how to use and critically assess AI systems^{15,16}.

Fourth, the evidence base is thin. Much of the literature consists of small pilots and opinion pieces. Well-designed prospective studies are needed before AI tools can be recommended with confidence for routine educational use¹⁰.

In short, AI is neither a perfect solution nor something that should simply be feared. When used well, AI may improve selected aspects of personalisation and efficiency in medical education, but the magnitude and generalisability of these benefits remain uncertain. When used poorly, it can weaken the very skills and values it is meant to support. Moving forward will require careful testing, honest evaluation, and a strong commitment to keeping patients, not algorithms, at the heart of medical training.

References

1. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. *JMIR Med Educ*. 2019;5(1):e13930. doi: <https://doi.org/10.2196/13930>

2. Masters K. Artificial intelligence in medical education. *Med Teach*. 2019;41(9):976-980. doi: <https://doi.org/10.1080/0142159X.2019.1595557>
3. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. doi: <https://doi.org/10.1371/journal.pdig.0000198>
4. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? *JMIR Med Educ*. 2023;9:e45312. doi: <https://doi.org/10.2196/45312>
5. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv*. Published March 24, 2023. doi: <https://doi.org/10.48550/arXiv.2303.13375>
6. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi: <https://doi.org/10.1038/s41586-023-06291-2>
7. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. *Nat Med*. 2025;31(3):943-950. doi: <https://doi.org/10.1038/s41591-024-03423-7>
8. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. *Acad Med*. 2024;99(1):22-27. doi: <https://doi.org/10.1097/ACM.0000000000005439>
9. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. *Acad Med*. 2021;96(11S):S62-S70. doi: <https://doi.org/10.1097/ACM.0000000000004291>
10. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. *Med Teach*. 2024;46(4):446-470. doi: <https://doi.org/10.1080/0142159X.2024.2314198>
11. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388(13):1233-1239. doi: <https://doi.org/10.1056/NEJMs2214184>
12. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi: <https://doi.org/10.1126/science.aax2342>
13. Kiyak YS, Emekli E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. *Postgrad Med J*. 2024;100(1189):858-865. doi: <https://doi.org/10.1093/postmj/qae065>
14. Park JJ, Tiefenbach J, Demetriades AK. The role of artificial intelligence in surgical simulation. *Front Med Technol*. 2022;4:1076755. doi: <https://doi.org/10.3389/fmedt.2022.1076755>
15. Masters K. Ethical use of artificial intelligence in health professions education: AMEE Guide No. 158. *Med Teach*. 2023;45(6):574-584. doi: <https://doi.org/10.1080/0142159X.2023.2186203>
16. Association of American Medical Colleges. Principles for the Responsible Use of Artificial Intelligence in and for Medical Education. Published 2024.
17. American Medical Association. AMA Principles for Augmented Intelligence Development, Deployment, and Use. Published 2024.
18. World Health Organization. Ethics and Governance of Artificial Intelligence for Health. Published 2021.
19. World Health Organization. Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models. Published 2024.
20. Baethge C, Goldbeck-Wood S, Mertens S. SANRA-a scale for the quality assessment of narrative review articles. *Res Integr Peer Rev*. 2019;4:5. doi: <https://doi.org/10.1186/s41073-019-0064-8>
21. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: applications and implications. *JMIR Med Educ*. 2023;9:e50945. doi: <https://doi.org/10.2196/50945>
22. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med Educ*. 2023;9:e48291. doi: <https://doi.org/10.2196/48291>
23. Abdellatif H, Al Mushaiqri M, Albalushi H, Al-Zaabi AA, Roychoudhury S, Das S. Teaching, learning and assessing anatomy with artificial intelligence: the road to a better future. *Int J Environ Res Public Health*. 2022;19(21):14209. doi: <https://doi.org/10.3390/ijerph192114209>
24. Totlis T, Otountzidis N, Papadopoulos S, Piagkou M, Natsis K. The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. *Surg Radiol Anat*. 2023;45(10):1321-1329. doi: <https://doi.org/10.1007/s00276-023-03229-1>
25. Kiyak YS, Soylu A, Coskun O, Budakoglu Il, Peker TV. Can ChatGPT generate acceptable case-based multiple-choice questions for medical school anatomy exams? *Clin Anat*. 2025;38(4):505-510. doi: <https://doi.org/10.1002/ca.24271>
26. Combs CD. Emerging roles of virtual patients in the age of AI. *AMA J Ethics*. 2019;21(2):E153-E159. doi: <https://doi.org/10.1001/amajethics.2019.153>
27. Mergen M, Junga A, Risse B, Valkov D, Graf N, Marschall B. Immersive training of clinical decision making with AI driven virtual patients-a new VR platform called medical tr.AI.ning. *GMS J Med*

- Educ. 2023;40(2):Doc18. doi: <https://doi.org/10.3205/zma001600>
28. Fazlollahi AM, Bakhaidar M, Alsayegh A, et al. Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students: a randomized clinical trial. *JAMA Netw Open*. 2022;5(2):e2149008. doi: <https://doi.org/10.1001/jamanetworkopen.2021.49008>
 29. Cheung BHH, Lau GKK, Wong GTC, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions-a multinational prospective study. *PLoS One*. 2023;18(8):e0290691. doi: <https://doi.org/10.1371/journal.pone.0290691>
 30. Coskun O, Kiyak YS, Budakoglu IL. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: a randomized controlled experiment. *Med Teach*. 2025;47(2):268-274. doi: <https://doi.org/10.1080/0142159X.2024.2327477>
 31. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA*. 2023;330(1):78-80. doi: <https://doi.org/10.1001/jama.2023.8288>
 32. Cabral S, Restrepo D, Kanjee Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA Intern Med*. 2024;184(5):581-583. doi: <https://doi.org/10.1001/jamainternmed.2024.0295>
 33. Goh E, Gallo R, Hom J, et al. Influence of a large language model on diagnostic reasoning: a randomized clinical vignette study. *JAMA Netw Open*. 2024;7(10):e2440969. doi: <https://doi.org/10.1001/jamanetworkopen.2024.40969>
 34. Hamilton A, Molzahn A, McLemore K. The evolution from standardized to virtual patients in medical education. *Cureus*. 2024;16(10):e71224. doi: <https://doi.org/10.7759/cureus.71224>
 35. Winkler-Schwartz A, Bissonnette V, Mirchi N, et al. Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation. *J Surg Educ*. 2019;76(6):1681-1690. doi: <https://doi.org/10.1016/j.jsurg.2019.05.015>
 36. Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, et al. Testing of detection tools for AI-generated text. *Int J Educ Integr*. 2023;19:26. doi: <https://doi.org/10.1007/s40979-023-00146-z>
 37. Strong E, DiGiammarino A, Weng Y, Kumar A, Hosamani P, Hom J, Chen JH. Chatbot vs medical student performance on free-response clinical reasoning examinations. *JAMA Intern Med*. 2023;183(9):1028-1030. doi: <https://doi.org/10.1001/jamainternmed.2023.2909>
 38. Mondal H. Ethical engagement with artificial intelligence in medical education. *Adv Physiol Educ*. 2025;49(1):163-165. doi: <https://doi.org/10.1152/advan.00188.2024>
 39. Katznelson G, Gerke S. The need for health AI ethics in medical school education. *Adv Health Sci Educ Theory Pract*. 2021;26(4):1447-1458. doi: <https://doi.org/10.1007/s10459-021-10040-3>
 40. Grunhut J, Wyatt ATM, Marques O. Educating future physicians in artificial intelligence (AI): an integrative review and proposed changes. *J Med Educ Curric Dev*. 2021;8:23821205211036836. doi: <https://doi.org/10.1177/23821205211036836>
 41. Rudnik A, Jendro O, Mastěj A, et al. Artificial intelligence is revolutionizing everyday medical practice. *J Educ Health Sport*. 2023;28(1):155-165. doi: <https://doi.org/10.12775/JEHS.2023.28.01.012>
 42. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Med Educ*. 2024;58(11):1276-1285. doi: <https://doi.org/10.1111/medu.15402>

Author's Contribution:**Conceptualization:** PK, MMi**Methodology:** PK, SFS**Check:** SSz, PAK, KRT**Investigation:** PW, NM, KRT**Data curation:** SFS, PAK, MMA, JM**Writing - rough preparation:** SSz, NM, MMi**Writing - review and editing:** PK, PW, JM**Visualization:** PK, MMA, KRT**Project administration:** SFS, PW, JM**Receiving funding:** not applicable

All authors have read and agreed with the published version of the manuscript.

Funding: The article did not receive any funding.**Institutional Review and Board Statement:** Not applicable.**Informed Consent Statement:** Not applicable.**Data Availability Statement:** Not applicable.**Conflict of Interest Statement:** Not applicable.