

Technologies, opportunities, challenges, and future directions for integrating generative artificial intelligence into medical education: a narrative review

Junseok Kang¹, Jihyun Ahn^{2*}

¹Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

²Department of Internal Medicine, Korea Medical Institute, Seoul, Korea

Generative artificial intelligence (GenAI), including large language models such as GPT-4 and image-generation tools like DALL-E, is rapidly transforming the landscape of medical education. These technologies present promising opportunities for advancing personalized learning, clinical simulation, assessment, curriculum development, and academic writing. Medical schools have begun incorporating GenAI tools to support students' self-directed study, design virtual patient encounters, automate formative feedback, and streamline content creation. Preliminary evidence suggests improvements in engagement, efficiency, and scalability. However, GenAI integration also introduces substantial challenges. Key concerns include hallucinated or inaccurate content, bias and inequity in artificial intelligence (AI)-generated materials, ethical issues related to plagiarism and authorship, risks to academic integrity, and the potential erosion of empathy and humanistic values in training. Furthermore, most institutions currently lack formal policies, structured training, and clear guidelines for responsible GenAI use. To realize the full potential of GenAI in medical education, educators must adopt a balanced approach that prioritizes accuracy, equity, transparency, and human oversight. Faculty development, AI literacy among learners, ethical frameworks, and investment in infrastructure are essential for sustainable adoption. As the role of AI in medicine expands, medical education must evolve in parallel to prepare future physicians who are not only skilled users of advanced technologies but also compassionate, reflective practitioners.

Keywords: Artificial intelligence; Curriculum; Formative feedback; Medical education; Natural language processing

Introduction

Background

Generative artificial intelligence (GenAI) refers to a class of machine learning technologies capable of producing novel content—including text, images, code, audio, and video—by learning patterns from large datasets rather than simply retrieving existing information [1]. Unlike traditional artificial intelligence (AI) systems designed primarily for classification or prediction, GenAI can synthesize entirely new outputs. This has significant implications for fields that depend on knowledge generation and communication, such as medicine and medical education [2]. Generally, GenAI denotes AI systems capable of creating new outputs across modalities (text, image, audio, code, video). In contrast,

large language models (LLMs) represent a specific subclass of GenAI focused exclusively on language, trained on vast text corpora to generate and interpret human language. While all LLMs are GenAI, not all GenAI systems are LLMs—a distinction that will be maintained consistently throughout this manuscript (Fig. 1).

The emergence of advanced LLMs such as OpenAI's GPT-4, Google's Med-PaLM 2, and open-source platforms like LLaMA has accelerated the integration of GenAI into academic environments [3]. These models can answer complex biomedical questions, simulate patient interactions, and generate coherent, high-fidelity text with accuracy that in some cases approaches human performance [4]. Notably, ChatGPT has demonstrated passing-level performance across all 3 United States Medical Licensing

*Corresponding email: jiyunahnmd@gmail.com

Received: August 25, 2025 Revised: September 11, 2025 Accepted: September 11, 2025

© 2025 Ewha Womans University College of Medicine and Ewha Medical Research Institute

© This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



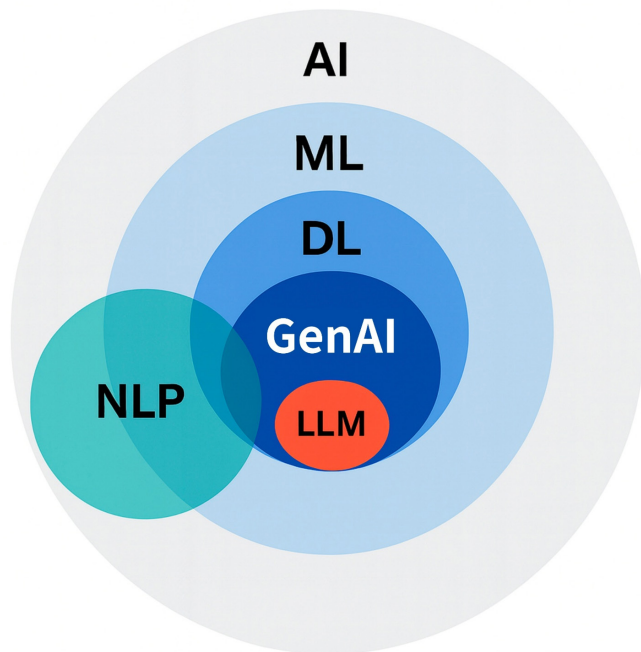


Fig. 1. Hierarchical relationship among artificial intelligence (AI), machine learning (ML), deep learning (DL), and generative artificial intelligence (GenAI). Large language models (LLMs) are depicted as a subset of GenAI, while natural language processing (NLP) overlaps multiple layers, reflecting its broad role in both traditional and GenAI applications. Figure generated by the authors using ChatGPT.

Examination (USMLE) Step exams without specialized medical training [5]. Comparable results have also been reported for licensing examinations in other countries, including Thailand [6]. In one study of the Thai National Licensing Examination, GPT-4 achieved 88.9% accuracy, substantially exceeding the national average and underscoring its potential as an equitable and multilingual preparatory tool [6]. This capability carries profound implications. If general-purpose LLMs can perform at or above the level of human trainees on standardized assessments, fundamental assumptions about how future physicians are taught, evaluated, and credentialed must be reconsidered [3]. The capacity of GenAI to serve as a real-time tutor, content generator, and even clinical reasoning assistant represents both a disruptive opportunity and a mandate for reform in medical pedagogy.

Since the public release of ChatGPT in late 2022, interest in GenAI tools among educators and learners has risen sharply [2]. Early adopters have applied these tools for exam preparation, note summarization, content development, and virtual patient scenario generation [1]. Institutions such as Harvard Medical School (HMS) have proactively incorporated GenAI into their curricula, introducing AI-focused courses and funding innovation projects

that integrate GenAI into teaching and assessment [2]. Beginning in fall 2023, HMS required all incoming students on the Health Sciences and Technology track to complete a one-month introductory course on AI in healthcare, making it the first medical school to mandate comprehensive AI training at the outset of medical education [7]. This course explores contemporary applications of AI in medicine, critically examines limitations in clinical decision-making, and prepares students for a healthcare environment demanding “good data skills, AI skills, and machine-learning skills” [8]. Additionally, the Department of Biomedical Informatics has launched a PhD (Doctor of Philosophy) track in AI in medicine to cultivate future leaders in healthcare AI technologies [9].

Despite these advances, enthusiasm is tempered by caution. Educators and scholars have expressed concerns regarding the accuracy of AI-generated content, particularly the risk of “hallucinations”—plausible yet incorrect outputs—as well as the ethical consequences of excessive reliance on automated systems [10]. Broader issues such as academic dishonesty, data privacy, algorithmic bias, and unequal access to GenAI tools further underscore the need for a measured, policy-guided approach to adoption [11]. Recent systematic reviews have mapped both the opportunities and challenges of LLMs in medical education while highlighting unresolved research gaps [12]. Such reviews identify opportunities including personalized learning plans, clinical simulation, and writing support, but also emphasize risks related to academic integrity, misinformation, and ethics—areas that demand careful regulation [12].

Objectives

The aim of this narrative review is to synthesize current knowledge on the integration of GenAI into medical education, emphasizing both the opportunities and challenges associated with its use. We provide a structured analysis of GenAI technologies relevant to medical education, examine their pedagogical applications, discuss ethical and logistical barriers to implementation, and propose recommendations for future research and institutional policy development.

Through this review, we aim to provide a balanced resource that enables medical educators, learners, and administrators to navigate this transformative shift responsibly, effectively, and ethically.

Ethics statement

This is a literature-based study; therefore, neither approval by an institutional review board nor informed consent was required.

Overview of GenAI technologies relevant to medical education

GenAI encompasses a wide range of models and modalities that can generate new, human-like outputs. In medical education, the most relevant GenAI technologies include LLMs, image-generation systems, multimodal models, and simulation tools that integrate natural language with visual or audio data [2] (Fig. 2, Table 1).

LLMs, one of the most prominent subtypes of GenAI, include GPT-4 (OpenAI), Med-PaLM 2 (Google), and open-source variants such as LLaMA and BLOOM. These models can generate fluent, contextually appropriate text responses based on user prompts [1]. Trained on massive corpora of web texts, scientific publications, and human conversations, they are capable of engaging in Q&A, summarizing content, simulating dialogues, and generating exam questions or clinical cases.

Importantly, general-purpose LLMs—without domain-specific fine-tuning—have achieved near or above-passing scores on high-stakes medical licensing exams, including USMLE Step 1, Step 2

GenAI technologies in medical education

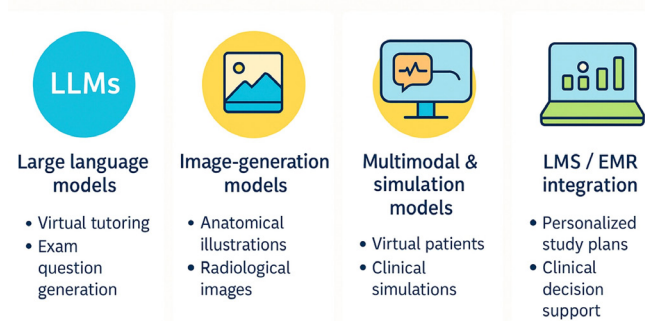


Fig. 2. Overview of generative artificial intelligence (GenAI) technologies in medical education. The infographic highlights 4 domains: large language models (LLMs) for tutoring and exam preparation; image-generation models for medical illustrations and dataset augmentation; multimodal and simulation models for immersive training; and integration with learning management systems (LMS) and electronic medical records (EMRs) for personalized support and clinical documentation. Figure generated by the authors using ChatGPT.

Table 1. Comparison of GenAI technologies in medical education

Technology type	Examples	Medical education applications	Key advantages	Limitations
LLMs	• GPT-4 (OpenAI) • Med-PaLM 2 (Google) • LLaMA • BLOOM	• Virtual tutoring and Q&A • Clinical reasoning support • Exam preparation • Content summarization • SOAP note generation	• 24/7 availability • Personalized responses • Passing scores on USMLE • Natural conversation • Scalable deployment	• Hallucinations • Knowledge cutoff dates • Lack of real-time data • Variable accuracy • No clinical judgment
Image-generation models	• DALL-E 3 (OpenAI) • Midjourney • Stable Diffusion • GANs	• Anatomical illustrations • Radiological findings • Dermatological cases • Histopathology slides • Clinical syndrome visualization	• Privacy protection • Unlimited variations • Rare case generation • No patient consent needed • Customizable features	• Potential inaccuracies • May generate unrealistic features • Limited medical specificity • Requires validation
Multimodal & simulation models	• GPT-4 with vision • AI-powered OSCEs • Virtual patient platforms • Integrated simulation environments	• Complete clinical encounters • Combined text/image/audio • Virtual standardized patients • Adaptive difficulty scenarios • Real-time feedback systems	• Immersive experiences • Multiple input types • Comprehensive training • Safe practice environment • Repeatable scenarios	• Technical complexity • High computational needs • Integration challenges • Limited emotional nuance
LMS/EMR integration	• AI-enhanced LMS platforms • EMR-integrated assistants • Performance analytics tools	• Personalized study plans • Documentation assistance • Performance tracking • Clinical decision support • Automated feedback	• Seamless workflow • Data-driven insights • Real-time assistance • Curriculum alignment • Progress monitoring	• Privacy concerns • System compatibility • Requires infrastructure • Policy barriers

GenAI, generative artificial intelligence; LLMs, large language models; SOAP, subjective, objective, assessment, and plan; USMLE, United States Medical Licensing Examination; GANs, generative adversarial networks; AI, artificial intelligence; OSCEs, objective structured clinical examinations; LMS, learning management systems; EMR, electronic medical records.

CK, and Step 3 [3,13], as well as specialty board examinations such as interventional radiology [14] and urology [15]. For example, GPT-4o correctly answered 67% of questions on the simulated European Board of Interventional Radiology exam, performing at a level comparable to trainees and even generating valid exam items for practice purposes [14]. By contrast, in a US urology knowledge assessment, GPT-4 outperformed GPT-3.5 (44% vs. 31%) but still fell below the 60% passing threshold, underscoring its role as a supplementary rather than standalone tool [15]. The ability of these models to approximate medical reasoning and generate coherent differential diagnoses positions them as valuable assistants for knowledge reinforcement and exam preparation. LLMs are also being adopted as conversational tutors in medical education. Students can interact with these systems to clarify complex concepts, explore reasoning pathways, or simulate patient interviews. In some studies, learners have reported using LLMs to create summaries, flashcards, and practice quizzes tailored to their weak areas [2].

Image-generation models

In addition to text, GenAI includes tools capable of producing high-quality images from textual prompts. Systems such as DALL-E 3 (OpenAI), Midjourney, and Stable Diffusion can generate visual materials—including realistic anatomical illustrations, radiological findings, and dermatological presentations—that may be integrated into medical teaching [16].

A notable advantage of image-generating models is their capacity to create synthetic yet realistic patient images without compromising privacy. For instance, DALL-E has been employed to produce diverse facial images displaying features of rare syndromes, thereby expanding access to visual cases while preserving patient confidentiality [12]. Similarly, histopathological images produced by generative adversarial networks have been shown to be virtually indistinguishable from real slides, enabling dataset augmentation in pathology education [17]. Furthermore, ChatGPT-4.0 has demonstrated competence in certain medical imaging evaluation tasks, suggesting potential applications in visually oriented medical fields [18]. However, its accuracy remains limited—for example, in radiographic positioning error detection, it achieved only partial recognition (mean score 2.9/5)—highlighting the need for continued radiography training despite its promise [18].

These technologies can enrich medical education by providing unlimited case variations, supporting visual diagnostic training, and mitigating the scarcity of publicly shareable clinical images.

Multimodal and simulation models

A particularly promising avenue is the development of multi-

modal AI systems that integrate textual, visual, and auditory inputs. Early examples include GPT-4 with image-input capabilities and models designed to interpret radiographs or simulate auscultatory findings. Such systems can replicate full clinical encounters, with text-based chatbots role-playing as patients [19].

In parallel, GenAI is increasingly being applied to create simulation environments and virtual patients. These include chatbot-based standardized patients (SPs), AI-driven objective structured clinical examination (OSCE) stations, and immersive scenarios that allow students to practice clinical reasoning, communication, and empathy in low-risk settings [20,21].

The scalability and flexibility of these AI-driven simulations offer particular advantages for institutions with limited access to human SPs or diverse clinical exposures. They also provide opportunities for repeated practice, real-time feedback, and adaptive difficulty adjustments based on learner performance [2].

Integration into learning management and clinical systems

Some institutions have begun piloting GenAI tools integrated directly into learning management systems (LMS) or electronic medical records. For example, GenAI can analyze student performance data in an LMS and generate personalized study recommendations [22]. Similarly, in clinical clerkships, AI-powered assistants can help students compose the Subjective, Objective, Assessment and Plan (SOAP) notes, formulate differential diagnoses, or draft patient instructions in real time—strengthening both documentation skills and clinical reasoning [3]. These applications blur the boundaries between education and clinical practice, underscoring the importance of ethical frameworks and institutional policies to ensure appropriate use [23].

The landscape of GenAI in medical education is expanding rapidly across multiple technological domains. LLMs provide a foundation for text-based tutoring and communication, image-generation models contribute to visual learning, and multimodal and simulation systems enable interactive, immersive experiences. As these technologies evolve, medical educators must remain aware of their capabilities and limitations in order to incorporate them effectively into curriculum design and instructional delivery.

Opportunities of GenAI in medical education

GenAI offers transformative opportunities across diverse aspects of medical education, from personalized learning to large-scale simulation and efficient content creation. By augmenting both teaching and learning processes, GenAI has the potential to

expand access, reduce faculty burden, and promote learner-centered education [2] (Fig. 3).

Personalized learning and tutoring

One of the most compelling applications of GenAI is its capacity to provide personalized, on-demand tutoring tailored to each learner's pace, knowledge level, and areas of weakness. LLMs such as ChatGPT can function as virtual tutors that answer questions, clarify difficult concepts, and create customized learning resources [1].

Unlike static textbooks or traditional lectures, GenAI tools interact conversationally with students, adapting in real time to individual needs. For example, a learner struggling with cardiology may request simplified explanations, analogies, or practice questions aligned with their specific knowledge gaps [2]. The ability to generate personalized quizzes, flashcards, and learning paths supports the principles of competency-based education [22].

Furthermore, these systems can deliver instant formative feedback, enabling learners to correct misconceptions before they become entrenched. Such feedback loops are particularly valuable for fostering self-directed learning and strengthening metacognitive skills [19].

Clinical simulation and virtual patients

GenAI is revolutionizing clinical skills training through the development of virtual patient simulations. LLMs can engage learners in natural language conversations that replicate history-taking, diagnostic reasoning, and management decisions [20]. Recent implementations have shown that ChatGPT-based virtual patients can be deployed at scale, at costs as low as US\$0.006 per complete conversation, with more than 75% of students reporting enhanced learning effectiveness [20].

Pilot studies have demonstrated that ChatGPT can serve as a chatbot-based standardized patient, providing learners with opportunities to practice interviewing skills in low-pressure environments. Students have reported improved confidence and communication skills, as well as reduced performance anxiety [20].

GenAI can also be applied to the design of OSCE cases, the generation of structured checklists for assessors, and the provision of real-time feedback on clinical reasoning and communication [21]. When combined with image-generation models, these simulations can incorporate synthetic photos, radiographs, or laboratory data, further enhancing realism [16,17].

Such simulations are particularly valuable in resource-limited settings where access to human SPs or diverse patient cases is constrained. They offer scalable, repeatable, and ethically safe training environments that strengthen experiential learning without compromising patient safety.

Assessment and feedback enhancement

GenAI creates new opportunities for efficient, scalable, and individualized assessment. LLMs can generate high-quality exam items, including multiple-choice questions, clinical vignettes, and open-ended prompts. These outputs can serve as foundations for exam construction or as self-assessment resources for students [24,25].

AI can also support automated grading. Natural language processing tools have been applied to assess OSCE performance and written responses, producing consistent and timely feedback. One study reported that AI-based graders achieved greater scoring consistency across OSCE domains, particularly in technical and checklist-based assessments [26].

In addition, AI-generated feedback can assist students in revising essays or reflective writing, identifying logical gaps in clinical reasoning, and visualizing learning progress over time. These features promote formative, learner-centered assessment by shifting the emphasis from punitive evaluation toward continuous improvement [2].

Educational content creation and curriculum design

GenAI also supports faculty by simplifying educational content development. Instructors can use LLMs to draft lecture slides, design case-based discussions, or create materials for problem-based learning. For example, by providing learning objectives, educators can prompt a GenAI tool to generate aligned case scenarios that include clinical data, differential diagnoses, and teaching questions [2].

Beyond content generation, GenAI can aid in curriculum mapping and optimization. AI tools can analyze curricular materials,

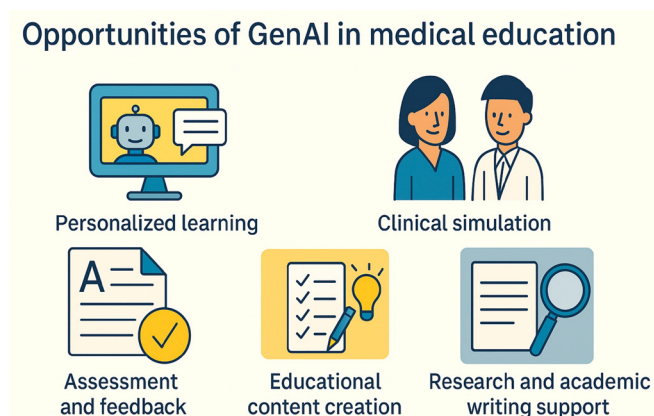


Fig. 3. Opportunities of generative artificial intelligence (GenAI) in medical education. Figure generated by the authors using ChatGPT.

identify gaps or redundancies, and recommend improvements aligned with national licensing requirements or competency frameworks [27]. Additional benefits include language translation and accessibility. AI-generated materials can be adapted for learners with varying levels of proficiency or translated into multiple languages, thereby promoting inclusivity and broadening access to educational resources [2].

Research and academic writing support

Medical students and residents frequently struggle with the workload and complexity of academic writing. GenAI tools can assist with literature summarization, research idea generation, and manuscript drafting. They can rapidly synthesize large bodies of literature, suggest relevant citations, and enhance grammar and clarity—benefits that are especially valuable for non-native English speakers [28]. Nonetheless, such support must be applied carefully. While GenAI can improve productivity, risks remain regarding hallucinated references or inappropriate authorship practices [2].

When used transparently and ethically, however, GenAI can significantly reduce barriers to academic engagement, enabling learners to focus more on critical thinking and data interpretation rather than formatting or summarization.

Challenges and risks of GenAI in medical education

Despite its transformative potential, the integration of GenAI into medical education introduces major challenges. These risks span technical, ethical, educational, and legal domains. Without thoughtful implementation, GenAI could unintentionally undermine core educational values and exacerbate existing disparities [1,2] (Fig. 4).

Ethical concerns: bias, equity, and humanistic values

One of the most pressing concerns is the inherent bias in AI-generated content. Because GenAI systems are trained on large datasets that reflect existing disparities in medical literature, they may produce biased outputs that underrepresent certain ethnicities, genders, or cultural perspectives [3]. This increases the risk that AI-generated cases or explanations may perpetuate stereotypes or deliver unequal educational value across populations.

The digital divide poses an additional equity concern. Access to high-performance GenAI tools often depends on paid subscriptions, stable internet connections, and advanced computing infrastructure. Students in resource-limited settings may therefore be disadvantaged [2].

A further ethical issue is the potential erosion of humanistic and

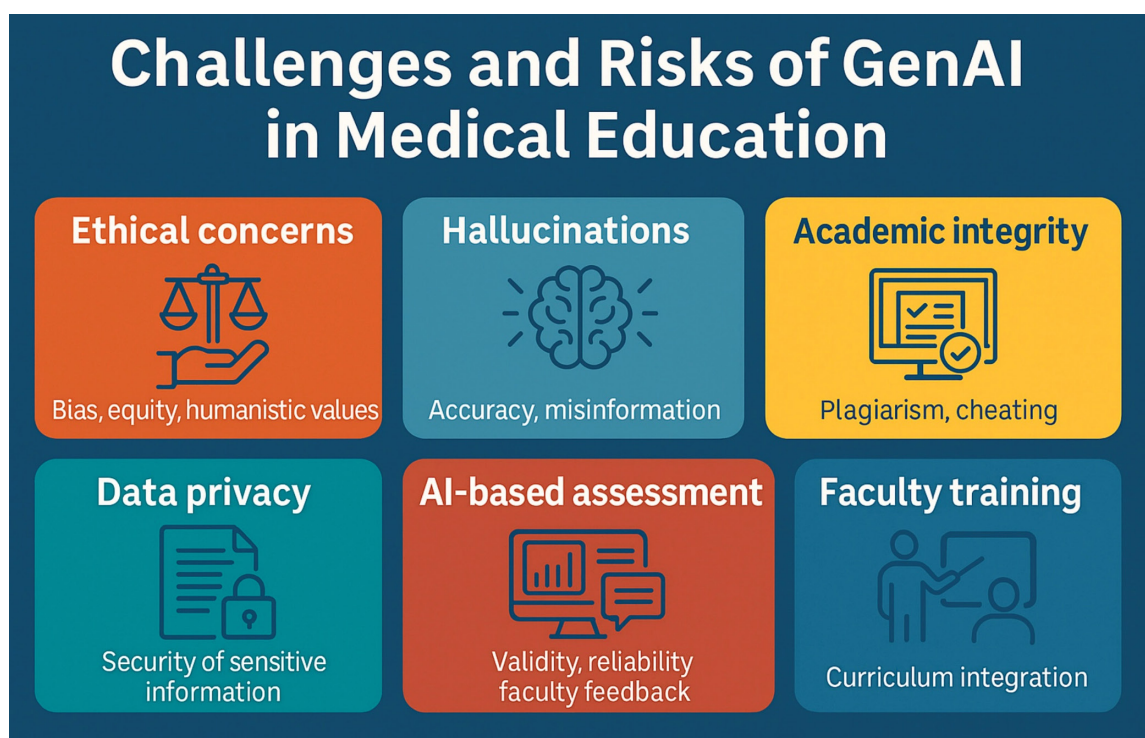


Fig. 4. Challenges and risks of generative artificial intelligence (GenAI) in medical education. Figure generated by the authors using ChatGPT.

professional values. Overreliance on AI tutors and simulations could reduce opportunities for real patient interaction and mentorship, both of which are critical for developing empathy, communication skills, and ethical judgment [20]. Thus, GenAI must be implemented in ways that augment, rather than replace, the human elements of medical education.

Hallucinations, accuracy, and misinformation

A well-recognized limitation of current LLMs is their tendency to hallucinate, producing fabricated or inaccurate information presented with unwarranted confidence. In medical education, this risk is especially concerning, as learners may accept such content uncritically [2]. For instance, ChatGPT has been shown to invent plausible-sounding references or provide incorrect clinical recommendations [10]. Moreover, models such as GPT-4 are trained on data with fixed cut-off dates, making them unaware of the most recent clinical guidelines or discoveries [1]. These limitations make human oversight essential. GenAI should be treated as a supplementary resource rather than an authoritative source, and students must be trained to critically appraise AI outputs, verify them against trusted references, and seek expert validation when necessary [2].

Academic integrity and plagiarism

Another pressing challenge involves academic integrity. GenAI tools can generate essays, solve problems, or compose reflective statements, raising concerns about “technological plagiarism” [2]. Students may misuse AI outputs in assignments or assessments, undermining the validity of evaluations.

Medical schools are beginning to adapt their academic integrity policies in response. Some now explicitly forbid the submission of AI-generated content without disclosure, while others are experimenting with AI-detection tools. However, such detectors remain imperfect and prone to false positives [2].

AI-assisted cheating during exams—particularly in remote or open-book formats—also poses a significant risk. Given that GenAI models perform well on licensing-style questions, assessment formats and proctoring strategies will need to be reevaluated [25].

Data privacy and security

The use of cloud-based GenAI platforms raises important privacy and security concerns, especially when students or faculty input clinical cases or patient information. Even de-identified data may inadvertently breach confidentiality if processed on third-party servers [3]. Educational institutions must therefore establish strict guidelines that prohibit entering protected health information into public GenAI tools. Some schools are developing en-

crypted or locally hosted AI systems to ensure that data remain within institutional firewalls, though such solutions require considerable technical investment [1].

Another unresolved issue involves the governance of student data. Many GenAI platforms store user interactions, yet policies regarding how this data is used, how long it is retained, and whether it contributes to future model training remain unclear. Transparent agreements between educational institutions and AI providers are essential to address these concerns.

Limitations in AI-based assessment and feedback

While GenAI can enhance efficiency in grading and feedback, concerns persist regarding validity and reliability. AI graders may overestimate performance in structured domains while failing to capture nuanced competencies such as empathy, ethical reasoning, or reflective writing [26].

Excessive reliance on AI-driven feedback also risks reducing faculty–student interaction, depriving learners of mentorship and contextual insight that only human educators can provide. In addition, students may receive inconsistent or oversimplified feedback if AI systems lack sufficient contextual awareness of local curricular objectives [2]. For these reasons, AI-based assessment should be employed as a supplement rather than a substitute, with continuous monitoring and quality assurance to ensure fairness and accuracy.

Faculty training and curriculum integration

Successful integration of GenAI requires targeted faculty development and deliberate curricular reform. Many educators lack familiarity with GenAI tools, leading to inconsistent implementation, skepticism, or undue reliance on unverified outputs [23]. To address these gaps, workshops, continuing education programs, and collaborative pilot projects are necessary to strengthen faculty AI literacy. Equally important is the thoughtful integration of AI-related content into medical curricula. Topics such as prompt engineering, ethical usage, and critical appraisal of AI outputs should be aligned with accreditation standards and professional competencies [20].

Without institutional support, AI tools risk being underused or misapplied, resulting in wasted resources or unintended consequences. Administrative leadership, resource allocation, and cross-disciplinary collaboration are therefore essential for effective adoption [1].

Current implementations and case studies

GenAI adoption in medical education has accelerated since

2022, largely due to the accessibility of tools such as ChatGPT and DALL-E. While most institutions remain in exploratory or pilot phases, several pioneering medical schools have already begun integrating GenAI into curricula, assessments, and clinical simulations [29].

Institutional adoption trends

A recent survey of US Osteopathic Medical Schools reported that 93% lacked formal GenAI policies and 73% had no plans to introduce mandatory AI-related education [29]. Nonetheless, adoption is evolving rapidly, with a clear divide emerging between early adopters and institutions still in preliminary planning. Whereas some schools experiment within individual courses, leading institutions have established comprehensive, structured programs.

Notable examples include:

Harvard Medical School

HMS launched a formal course, AI in healthcare, for incoming students and provided innovation grants to support faculty-led GenAI initiatives [2]. Designed collaboratively with faculty from HMS, MIT, and the Harvard T.H. Chan School of Public Health, the course covers technical foundations, ethical considerations, bias mitigation, and clinical integration strategies [8]. Students engage with case studies, hands-on programming, and discussions on health equity. HMS also established the AI in medicine program, serving as a hub for AI research, education, and innovation across the institution [30].

Mount Sinai's Icahn School of Medicine

Mount Sinai became the first medical school to integrate ChatGPT Edu across all levels of education—from preclinical coursework to clinical clerkships—under a formal agreement with OpenAI ensuring HIPAA compliance and data protection. Their approach emphasizes training students to use AI as an assistant rather than a replacement, with applications spanning clinical reasoning, case analysis, research, and curriculum development [31].

NYU Grossman School of Medicine

NYU introduced a precision education initiative that uses AI to personalize curricula and study aids according to students' learning styles and goals. Early pilot efforts involve first-year students, with AI operating within the student portal to deliver seamless, unobtrusive support. Applications extend to integration with electronic health record data, predictive analytics, and virtual learning tools [32].

These examples highlight early institutional willingness to explore GenAI's potential in enhancing both teaching and learning.

Applications in the classroom and beyond

In classroom settings, faculty have reported using LLMs to co-develop lecture slides, generate interactive quizzes, and facilitate real-time Socratic questioning via chatbots. In anatomy courses, AI-generated images and 3-dimensional visualizations have been used to supplement cadaver-based learning, particularly in remote or hybrid environments [16].

During clinical rotations, students have employed AI assistants to draft SOAP notes, review clinical guidelines, and simulate differential diagnoses in real time—transforming GenAI into a bedside learning companion [3].

In a notable simulation study, ChatGPT was deployed as a virtual standardized patient during mock OSCEs. The AI system simulated diverse emotional tones and symptom narratives during history-taking interviews. Learners reported reduced anxiety, improved confidence, and greater flexibility in practicing complex cases [20].

Faculty have also leveraged GenAI to create OSCE case checklists, rubrics, and feedback forms, thereby streamlining the development of standardized assessments and reducing examiner workload [21].

Policy gaps and institutional challenges

Despite the rapid adoption described earlier, policy development has not kept pace. Key concerns include: (1) Whether students must disclose AI assistance in assignments; (2) How to prevent and detect AI-generated plagiarism; (3) Who is accountable for AI-generated errors in student submissions; and (4) Whether AI output can be cited in academic writing.

In the absence of clear institutional guidance, usage patterns vary widely. Some faculty permit unrestricted use of GenAI as a writing or brainstorming tool, while others strictly prohibit it, citing concerns over fairness and academic integrity [2].

Student and faculty perspectives

Emerging qualitative studies indicate that students are eager adopters of GenAI, particularly Generation Z learners who value productivity, personalization, and efficiency [26]. Common student uses include (1) summarizing complex lectures, (2) practicing board-style questions, and (3) drafting structured reflections or research outlines.

At the same time, many students report uncertainty about how far they can ethically use these tools. Some have requested clearer institutional policies and formal training on appropriate AI usage

[2]. Faculty perspectives have been more diverse, ranging from enthusiastic experimentation to strong skepticism. While some educators regard GenAI as a productivity enhancer, others worry that it could erode fundamental clinical reasoning skills or be misused in high-stakes assessments [23].

Notably, educators who use GenAI most effectively often benefit from institutional support, training, or collaborative platforms, underscoring professional development as a key enabler of responsible adoption [33].

Future directions and recommendations

Responsible integration of GenAI into medical education requires a multi-layered approach, summarized in our framework (Fig. 5). This framework consists of 4 interconnected layers: core principles that guide all activities, a foundation of AI literacy and ethical guidelines, implementation through pilot programs and curricular integration, and outcomes such as enhanced learning and ethical AI use. Recommendations are organized according to these layers to provide a coherent roadmap for sustainable integration of GenAI. For example, “developing AI literacy and competency” aligns with the foundation, “reimagining assessment” with

implementation, and “human–AI interaction” with outcomes.

Developing AI literacy and competency

Ensuring the safe and effective use of GenAI in medical education is essential. Both students and faculty must build AI literacy—the ability to understand how these systems function, recognize their limitations, and apply them ethically [20]. As a cornerstone of the foundation layer, AI literacy provides the knowledge base upon which all subsequent implementation strategies and desired outcomes depend. Without this foundation, neither educators nor learners can deploy, evaluate, or benefit from GenAI tools effectively.

Educational initiatives should establish structured programs that cover fundamental principles of AI and machine learning, ensuring that both students and faculty understand the concepts driving these technologies. Curricula should also include practical training in prompt engineering for LLMs, equipping learners with the skills to interact effectively with AI systems. Equally important are evaluation skills—learners must be able to assess AI-generated outputs against trusted sources to foster a culture of verification and scholarly rigor. Case-based discussions that highlight ethical use and misuse scenarios provide valuable opportunities for learn-

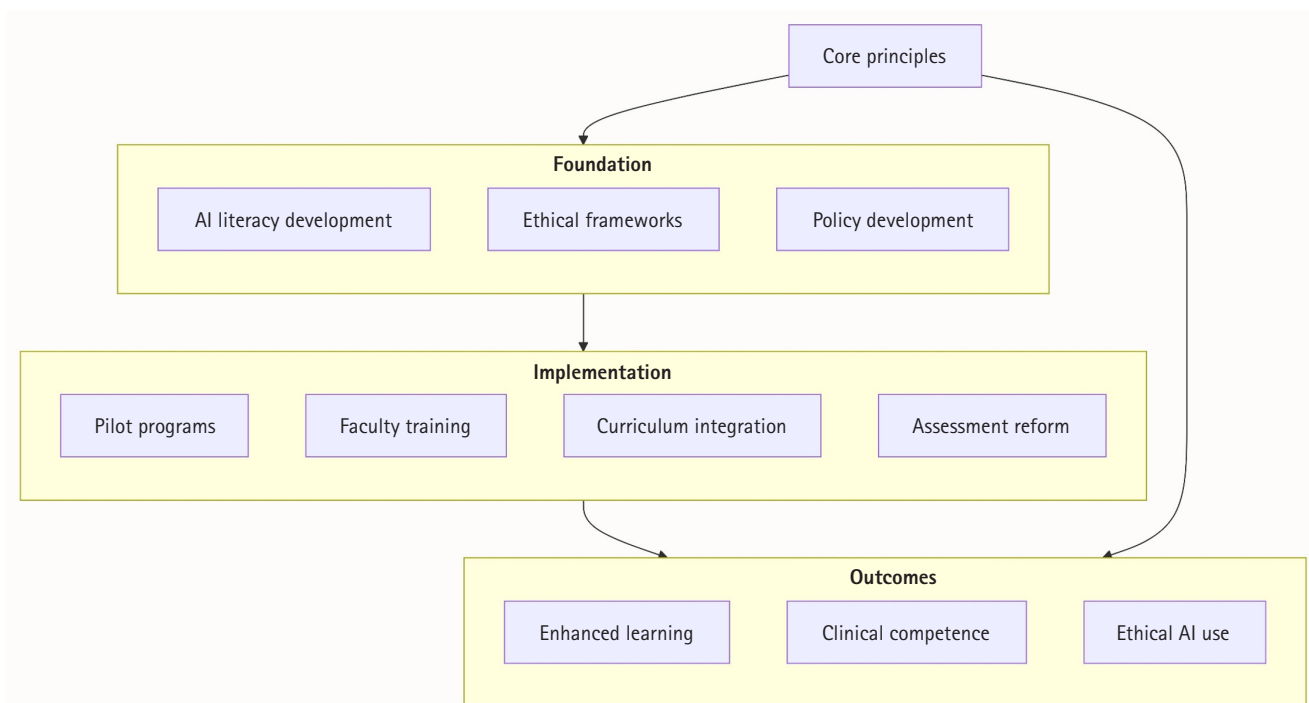


Fig. 5. Framework for responsible generative artificial intelligence (GenAI) integration in medical education. This framework illustrates the multi-layered approach required for responsible GenAI integration, emphasizing the interconnected nature of foundational elements, implementation strategies, and desired outcomes. All layers are guided by core principles and sustained through stakeholder engagement and continuous quality improvement.

ers and educators to practice navigating complex decisions about appropriate AI applications in medical contexts. Future research should explore the relative impact of stand-alone AI courses versus embedded instruction throughout the curriculum to determine the most effective teaching strategies [1].

Reimagining assessment in the AI era

As AI becomes increasingly capable of answering standard medical exam questions, educators must reconsider traditional assessment models. This transformation represents the Implementation layer of our framework, where foundational AI literacy must translate into concrete curricular reforms—particularly in assessment methods that safeguard academic integrity while leveraging AI's educational potential. Future assessments must emphasize higher-order cognitive skills, including complex clinical reasoning, ethical decision-making, and the demonstration of empathy in patient interactions. To this end, institutions should incorporate oral examinations and OSCEs that assess real-time thinking and adaptive problem-solving, providing authentic evaluations of clinical competence while reducing the risk of AI manipulation. The design of AI-resilient assessment formats is also crucial. These should focus on domains where GenAI tools cannot effectively substitute for personal reflection, experiential learning, and hands-on performance requiring human judgment and interpersonal skills [2]. At the same time, research is needed to validate whether AI-generated test items predict student performance and whether AI-assisted grading aligns with human judgment [26].

Human-AI interaction and educational outcomes

Understanding human-AI interaction in medical education requires rigorous empirical investigation. This research directly connects the implementation layer to the outcomes layer of our framework, measuring whether strategic changes in curriculum and assessment actually produce the enhanced learning, improved clinical competence, and ethical AI use that constitute our ultimate goals.

A central recommendation for guiding learner-AI interactions is the DEFT-AI framework (Diagnosis, Evidence, Feedback, Teaching, and recommendations for AI use). This framework provides supervisors with a structured method to promote critical thinking during learner-AI interactions [34]. By distinguishing among delegation, evaluation, and collaborative use, DEFT-AI encourages learners to verify AI outputs before relying on them, reinforcing the “verify before trust” paradigm. Incorporating this framework ensures that AI functions as a tool to strengthen, rather than weaken, clinical reasoning capabilities.

Mitigating algorithmic bias and ensuring equity

Addressing algorithmic bias and ensuring equitable access to GenAI technologies is a critical priority for medical education institutions. Equity, as a core principle of our framework, must permeate every layer—from ensuring diverse representation in AI literacy programs, to piloting initiatives in underserved contexts, to measuring disparities in educational outcomes across different student populations. Systematic efforts are needed to audit and mitigate biases embedded in GenAI training datasets. Collaboration between developers and educators will be essential to fine-tune models using diverse, representative data that encompass a broad range of medical conditions, populations, and cultural contexts. Integrating bias-detection mechanisms into AI systems can provide ongoing monitoring and quality assurance, while involving students in bias-identification exercises can cultivate the critical consciousness required for ethical medical practice [23]. These educational approaches provide learning opportunities, fostering awareness of inequities in GenAI processes and promoting culturally competent healthcare delivery. Equally important is equitable access. Pilot programs in underserved regions should explore lightweight or offline GenAI solutions and assess their effects on learning outcomes [2].

Establishing ethical and regulatory frameworks

Robust institutional policies and ethical guidelines must underpin all applications of GenAI. These frameworks embody the core principles of transparency and responsibility, while also forming part of the Foundation layer that enables subsequent implementation and ensures sustainable outcomes. Recent scholarship suggests that competency-based medical education frameworks may provide useful analogues for AI regulation, given similar challenges in governing adaptive systems with opaque processes [35]. Institutional policies must address critical aspects of disclosure and attribution of AI assistance, ensuring transparency in academic work while upholding scholarly honesty and intellectual property standards. Clear distinctions between acceptable and prohibited uses of GenAI in coursework and assessments are essential for maintaining educational validity and preventing misconduct. Equally important are comprehensive data privacy protections for AI-generated materials to ensure compliance with regulatory requirements. Furthermore, accountability structures must be established to address potential harms associated with AI use. These should include defined responsibility chains and remediation processes to protect both learners and institutions from negative consequences [1]. Professional organizations and accrediting bodies should also consider establishing competency standards for AI literacy in medical education and recommending shared

ethical usage norms across institutions.

Advancing domain-specific medical AI tools

To improve accuracy and safety, future GenAI applications should prioritize the development of medical-specific LLMs trained exclusively on curated, peer-reviewed clinical data [3]. These technological advances apply across all layers of the framework—strengthening the foundation with more reliable educational tools, enabling implementation through sophisticated clinical simulations, and contributing to outcomes of enhanced learning and clinical competence. Specialized models offer superior factual reliability and domain relevance compared with general-purpose systems that may lack medical precision. Integrating multimodal capabilities—such as combining voice, image, and text inputs—holds transformative potential for creating immersive simulations that closely replicate real clinical scenarios. Such innovations are particularly valuable in specialized fields like radiology, dermatology, and procedural training [19]. These advancements will require sustained collaboration among AI developers, clinicians, and educators to ensure that tools are pedagogically appropriate. Trials of instructor-facing AI systems that deliver cohort-level performance analytics or targeted teaching recommendations can further improve educational effectiveness and guide curriculum development [22].

Conclusion

GenAI presents vast opportunities for medical education, including personalized learning, advanced simulations, and efficient content development, thereby enhancing engagement and scalability. At the same time, it introduces significant challenges, such as hallucinations, inherent biases, threats to academic integrity, and data privacy concerns, all of which may undermine critical thinking and empathy. Successful integration depends on preserving the humanistic values of medical education. This requires clear institutional guidelines, comprehensive faculty development, essential AI literacy for learners, and robust ethical frameworks. AI should serve as a vital collaborator rather than a replacement for human expertise, helping cultivate physicians with strong clinical judgment, sound ethical reasoning, and profound empathy. Aligning technological innovation with foundational medical competencies will ensure that GenAI strengthens, rather than diminishes, the mission of medical education.

ORCID

Junseok Kang: <https://orcid.org/0009-0000-6179-6098>

Jihyun Ahn: <https://orcid.org/0000-0001-5592-6079>

Authors' contribution

Conceptualization: JA. Data curation: not applicable. Formal analysis: JK, JA. Investigation: JK, JA. Writing—original draft: JK, JA. Writing—review & editing: JK, JA.

Conflict of interest

No potential conflict of interest relevant to this article was reported.

Funding

None.

Data availability

Not applicable.

Acknowledgments

Figs. 1–4 in this manuscript were generated with assistance from ChatGPT-5 (OpenAI), based on author-provided prompts.

Supplementary materials

None.

References

1. Tran M, Balasooriya C, Jonnagaddala J, Leung GK, Mahboobani N, Ramani S, Rhee J, Schuwirth L, Najafzadeh-Tabrizi NS, Semmler C, Wong ZS. Situating governance and regulatory concerns for generative artificial intelligence and large language models in medical education. *NPJ Digit Med* 2025;8:315. <https://doi.org/10.1038/s41746-025-01721-z>
2. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. *JMIR Med Educ* 2023;9:e48785. <https://doi.org/10.2196/48785>
3. Janumpally R, Nanua S, Ngo A, Youens K. Generative artificial intelligence in graduate medical education. *Front Med (Lausanne)* 2024;11:1525604. <https://doi.org/10.3389/fmed.2024.1525604>
4. Ahn J, Kim B. Application of generative artificial intelligence in dyslipidemia care. *J Lipid Atheroscler* 2025;14:77-93. <https://doi.org/10.12997/jla.2025.14.1.77>
5. Rabbani SA, El-Tanani M, Sharma S, Rabbani SS, El-Tanani Y, Kumar R, Saini M. Generative artificial intelligence in health-care: applications, implementation challenges, and future directions. *BioMedInformatics* 2025;5:37. <https://doi.org/10.3390/biomedinformatics5030037>
6. Saowaprut P, Wabina RS, Yang J, Siriwat L. Performance of large

- language models on Thailand's national medical licensing examination: a cross-sectional study. *J Educ Eval Health Prof* 2025;22:16. <https://doi.org/10.3352/jeehp.2025.22.16>
7. Paulus VH, Ravi A. AI is changing healthcare: Harvard Medical School is following suit. *The Harvard Crimson Inc.*; 2024.
8. Chang BS. Transformation of undergraduate medical education in 2023. *JAMA* 2023;330:1521-1522. <https://doi.org/10.1001/jama.2023.16943>
9. Harvard Medical School. AI in Medicine PhD Track: program description. Harvard Medical School; 2025.
10. Garcia-Lopez IM, Trujillo-Linan L. Ethical and regulatory challenges of generative AI in education: a systematic review. *Front Educ* 2025;10:1565938. <https://doi.org/10.3389/feduc.2025.1565938>
11. Hale J, Alexander S, Wright ST, Gilliland K. Generative AI in undergraduate medical education: a rapid review. *J Med Educ* 2024;11:1-15. <https://doi.org/10.1177/23821205241266697>
12. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. *J Educ Eval Health Prof* 2024;21:6. <https://doi.org/10.3352/jeehp.2024.21.6>
13. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepano C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2:e0000198. <https://doi.org/10.1371/journal.pdig.0000198>
14. Ebel S, Ehrengut C, Denecke T, Gossmann H, Beeskow AB. GPT-4o's competency in answering the simulated written European Board of Interventional Radiology exam compared to a medical student and experts in Germany and its ability to generate exam items on interventional radiology: a descriptive study. *J Educ Eval Health Prof* 2024;21:21. <https://doi.org/10.3352/jeehp.2024.21.21>
15. Yudovich MS, Makarova E, Hague CM, Raman JD. Performance of GPT-3.5 and GPT-4 on standardized urology knowledge assessment items in the United States: a descriptive study. *J Educ Eval Health Prof* 2024;21:17. <https://doi.org/10.3352/jeehp.2024.21.17>
16. Fan BE, Chow M, Winkler S. Artificial intelligence-generated facial images for medical education. *Med Sci Educ* 2024;34:5-7. <https://doi.org/10.1007/s40670-023-01942-5>
17. Levine AB, Peng J, Farnell D, Nursey M, Wang Y, Naso JR, Ren H, Farahani H, Chen C, Chiu D, Talhouk A, Sheffield B, Riaz M, Ip PP, Parra-Herran C, Mills A, Singh N, Tessier-Cloutier B, Salisbury T, Lee J, Salcudean T, Jones SJ, Huntsman DG, Gilks CB, Yip S, Bashashati A. Synthesis of diagnostic quality cancer pathology images by generative adversarial networks. *J Pathol* 2020;252:178-188. <https://doi.org/10.1002/path.5509>
18. Arruzza ES, Evangelista CM, Chau M. The performance of ChatGPT-4.0o in medical imaging evaluation: a cross-sectional study. *J Educ Eval Health Prof* 2024;21:29. <https://doi.org/10.3352/jeehp.2024.21.29>
19. Chu SN, Goodell AJ. Synthetic patients: simulating difficult conversations with multimodal generative ai for medical education. *arXiv [Preprint]* 2024 May 30. <https://doi.org/10.48550/arXiv.2405.19941>
20. Cross J, Kayalackakom T, Robinson RE, 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 JE. Assessing ChatGPT's capability as a new age standardized patient: qualitative study. *JMIR Med Educ* 2025;11:e63353. <https://doi.org/10.2196/63353>
21. Misra SM, Suresh S. Artificial intelligence and objective structured clinical examinations: using ChatGPT to revolutionize clinical skills assessment in medical education. *J Med Educ Curric Dev* 2024;11:23821205241263475. <https://doi.org/10.1177/23821205241263475>
22. Mitra NK, Chitra E. Glimpses of the use of generative AI and ChatGPT in medical education. *Educ Med J* 2024;16:155-164. <https://doi.org/10.21315/eimj2024.16.2.11>
23. Franco D'Souza R, Mathew M, Mishra V, Surapaneni KM. Twelve tips for addressing ethical concerns in the implementation of artificial intelligence in medical education. *Med Educ Online* 2024;29:2330250. <https://doi.org/10.1080/10872981.2024.2330250>
24. 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;100:858-865. <https://doi.org/10.1093/postmj/qgae065>
25. Qiu Y, Liu C. Capable exam-taker and question-generator: the dual role of generative AI in medical education assessment. *Global Med Educ* 2025 Jan 14 [Epub]. <https://doi.org/10.1515/gme-2024-0021>
26. Tekin M, Yurdal MO, Toraman C, Korkmaz G, Uysal I. Is AI the future of evaluation in medical education?: AI vs. human evaluation in objective structured clinical examination. *BMC Med Educ* 2025;25:641. <https://doi.org/10.1186/s12909-025-07241-4>
27. Shimizu I, Kasai H, Shikino K, Araki N, Takahashi Z, Onodera M, Kimura Y, Tsukamoto T, Yamauchi K, Asahina M, Ito S, Kawakami E. Developing medical education curriculum reform strategies to address the impact of generative AI: qualitative study. *JMIR Med Educ* 2023;9:e53466. <https://doi.org/10.2196/63353>

- 10.2196/53466
28. Lie M, Rodman A, Crowe B. Harnessing generative artificial intelligence for medical education. *Acad Med* 2025;100:116. <https://doi.org/10.1097/ACM.0000000000005760>
29. Ichikawa T, Olsen E, Vinod A, Glenn N, Hanna K, Lund GC, Pierce-Talsma S. Generative artificial intelligence in medical education-policies and training at US osteopathic medical schools: descriptive cross-sectional survey. *JMIR Med Educ* 2025;11:e58766. <https://doi.org/10.2196/58766>
30. Harvard Medical School. AI in health care: from strategies to implementation: course description [Internet]. Harvard Medical School; 2025 [cited 2025 Aug 10]. Available from: <https://learn.hms.harvard.edu/programs/ai-health-care-strategies-implementation>
31. Mount Sinai. Icahn School of Medicine at Mount Sinai expands AI innovation with OpenAI's ChatGPT Edu rollout [Internet]. Mount Sinai; 2025 [cited 2025 Aug 10]. Available from: <https://www.mountsinai.org/about/newsroom/2025/icahn-school-of-medicine-at-mount-sinai-expands-ai-innovation-with-openais-chatgpt-edu-rollout>
32. NYU Langone Health. Artificial intelligence supercharges learning for students at NYU Grossman School of Medicine. NYU Langone Health; 2023.
33. McCoy L, Ganesan N, Rajagopalan V, McKell D, Nino DF, Swaim MC. A training needs analysis for AI and generative AI in medical education: perspectives of faculty and students. *J Med Educ Curric Dev* 2025;12:23821205251339226. <https://doi.org/10.1177/23821205251339226>
34. Abdalnour RE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. *N Engl J Med* 2025;393:786-797. <https://doi.org/10.1056/NEJMr2503232>
35. Vokinger KN, Soled DR, Abdalnour R. Regulation of AI: learnings from medical education. *NEJM AI* 2025;2:AIP2401059. <https://doi.org/10.1056/AIP2401059>