Review

Applications, Challenges, and Prospects of Generative Artificial Intelligence Empowering Medical Education: Scoping Review

Yuhang Lin[1*]; Zhiheng Luo[2*]; Zicheng Ye[1]; Nuoxi Zhong[2]; Lijian Zhao[1]; Long Zhang[3]; Xiaolan Li[1], PhD; Zetao Chen[1], PhD; Yijia Chen[1], PhD

[1]Guangdong Provincial Key Laboratory of Stomatology, Hospital of Stomatology, Guanghua School of Stomatology, Sun Yat-sen University, Guangzhou, China

[2]Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China

[3]School of Government, Sun Yat-sen University, Guangzhou, China

*these authors contributed equally

Corresponding Author:

Yijia Chen, PhD
Guangdong Provincial Key Laboratory of Stomatology
Hospital of Stomatology, Guanghua School of Stomatology, Sun Yat-sen University
No. 56, Lingyuan Road West
Guangzhou 510055
China
Phone: 86 13580591020
Email: chenyij9@mail.sysu.edu.cn

Abstract

Background: Nowadays, generative artificial intelligence (GAI) drives medical education toward enhanced intelligence, personalization, and interactivity. With its vast generative abilities and diverse applications, GAI redefines how educational resources are accessed, teaching methods are implemented, and assessments are conducted.

Objective: This study aimed to review the current applications of GAI in medical education; analyze its opportunities and challenges; identify its strengths and potential issues in educational methods, assessments, and resources; and capture GAI's rapid evolution and multidimensional applications in medical education, thereby providing a theoretical foundation for future practice.

Methods: This scoping review used PubMed, Web of Science, and Scopus to analyze literature from January 2023 to October 2024, focusing on GAI applications in medical education. Following PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines, 5991 articles were retrieved, with 1304 duplicates removed. The 2-stage screening (title or abstract and full-text review) excluded 4564 articles and a supplementary search included 8 articles, yielding 131 studies for final synthesis. We included (1) studies addressing GAI's applications, challenges, or future directions in medical education, (2) empirical research, systematic reviews, and meta-analyses, and (3) English-language articles. We excluded commentaries, editorials, viewpoints, perspectives, short reports, or communications with low levels of evidence, non-GAI technologies, and studies centered on other fields of medical education (eg, nursing). We integrated quantitative analysis of publication trends and Human Development Index (HDI) with thematic analysis of applications, technical limitations, and ethical implications.

Results: Analysis of 131 articles revealed that 74.0% (n=97) originated from countries or regions with very high HDI, with the United States contributing the most (n=33); 14.5% (n=19) were from high HDI countries, 5.3% (n=7) from medium HDI countries, and 2.2% (n=3) from low HDI countries, with 3.8% (n=5) involving cross-HDI collaborations. ChatGPT was the most studied GAI model (n=119), followed by Gemini (n=22), Copilot (n=11), Claude (n=6), and LLaMA (n=4). Thematic analysis indicated that GAI applications in medical education mainly embody the diversification of educational methods, scientific evaluation of educational assessments, and dynamic optimization of educational resources. However, it also highlighted current limitations and potential future challenges, including insufficient scene adaptability, data quality and information bias, overreliance, and ethical controversies.

Conclusion: GAI application in medical education exhibits significant regional disparities in development, and model research statistics reflect researchers' certain usage preferences. GAI holds potential for empowering medical education, but widespread adoption requires overcoming complex technical and ethical challenges. Grounded in symbiotic agency theory, we advocate establishing the resource-method-assessment tripartite model, developing specialized models and constructing an integrated system of general large language models incorporating specialized ones, promoting resource sharing, refining ethical governance, and building an educational ecosystem fostering human-machine symbiosis, enabling deep tech-humanism integration and advancing medical education toward greater efficiency and human-centeredness.

JMIR Med Educ2025;11:e71125; doi: 10.2196/71125

Keywords: generative artificial intelligence; GAI; large language model; ChatGPT; medical education; human-machine symbiosis

Introduction

Background

The 21st century has seen accelerated advancement in information technology and artificial intelligence (AI), significantly altering lifestyles and work paradigms. With progress in deep learning and large-scale data processing, generative artificial intelligence (GAI) has emerged as an influential innovation. GAI rapidly expands into diverse applications, enabling content generation across text, images, and audio through the analysis of extensive datasets [1]. Its market demonstrates notable growth, with a 2024 global valuation of ~US $16.8 billion and a projected 37.6% compound annual growth rate (CAGR) from 2025 to 2030 [2], reflecting its significance in commercial and academic domains.

GAI's development is driven by advances in natural language processing (NLP), particularly the Transformer architecture, which enables the generation of complex content. Large language models (LLMs) serve as core technical implementations of GAI. Models like GPT-3, GPT-4, Copilot, and LLaMA 3 have expanded GAI applications from basic automation to sophisticated tasks including content creation, data analysis, and intelligent question-answering systems [3]. These transformer-based LLMs exemplify how conceptual GAI frameworks are operationalized via model architectures and engineering practices.

With technological advancements, GAI has gradually infiltrated more specialized fields, with medical education a prime example. This domain faces challenges due to its knowledge-intensive and highly practical characteristics: traditional teaching methods struggle to replicate clinical scenarios efficiently, and increasingly scarce clinical teaching specimens and patient resources limit the clinical practice training of medical students, all of which are not conducive to the cultivation of medical talents with both clinical thinking and practical ability [4]. In this content, GAI may empower medical education through its enhancement effects on 3 core educational elements: improving resource generation efficiency, optimizing the interactivity of pedagogical approaches, and enhancing the automation level of assessment processes [5,6]. Nevertheless, the accompanying integration risks include potential biases and inaccuracies in generated content [7] and possible inhibition of critical thinking through over-reliance [8]. Thus, optimal implementation strategies warrant further investigation.

Current GAI integration in medical education involves rapid technological iteration and shifting research paradigms [1,9,10]. Prior reviews exhibit three limitations: (L1) Overreliance on single-model analyses (predominantly ChatGPT) [9,10], (L2) insufficient examination of geographical disparities in adoption patterns, and (L3) fragmented assessment of GAI's impact across 3 core dimensions of medical education. These dimensions include resources (teaching support materials like GAI-generated clinical cases and pathological images), methods (instructional strategies like adaptive learning pathways and simulated decision-making), and assessment (automated evaluation of learner performance, such as automated short-answer scoring). Crucially, studies before 2023 were constrained by the technology's maturity, missing the recent shift from theoretical exploration to operational implementation [1]. Therefore, a new round of scoping review is urgently needed to focus on the critical evolution period between January 2023 and October, 2024 (before the completion of this scoping review), construct a multidimensional analytical framework (encompassing resources, methods, and assessment), and clarify the complex picture of the deep interaction between GAI and medical education. To guide this investigation, this study discusses the multifaceted landscape of GAI adoption in medical education through 3 interconnected lines of inquiry. First, it aims to examine whether regional disparities exist in GAI implementation and how researchers exhibit preferences for specific LLMs (eg, ChatGPT). We posit that adoption patterns will demonstrate significant stratification aligned with national development levels and reflect preferential usage of widely accessible general-purpose models. Second, it seeks to map the current state of GAI applications across educational resources, methods, and assessment dimensions. We hypothesize that effectiveness will vary substantially across these domains due to differences in technical implementation requirements and inherent task complexities. Third, it intends to identify current limitations and future challenges, positing that technical deficiencies, including ethical risks such as compromised academic integrity and data hallucinations, will constitute the most significant barriers to sustainable integration.

Theoretical Framework

Theoretical Model: The Theory of Symbiotic Agency

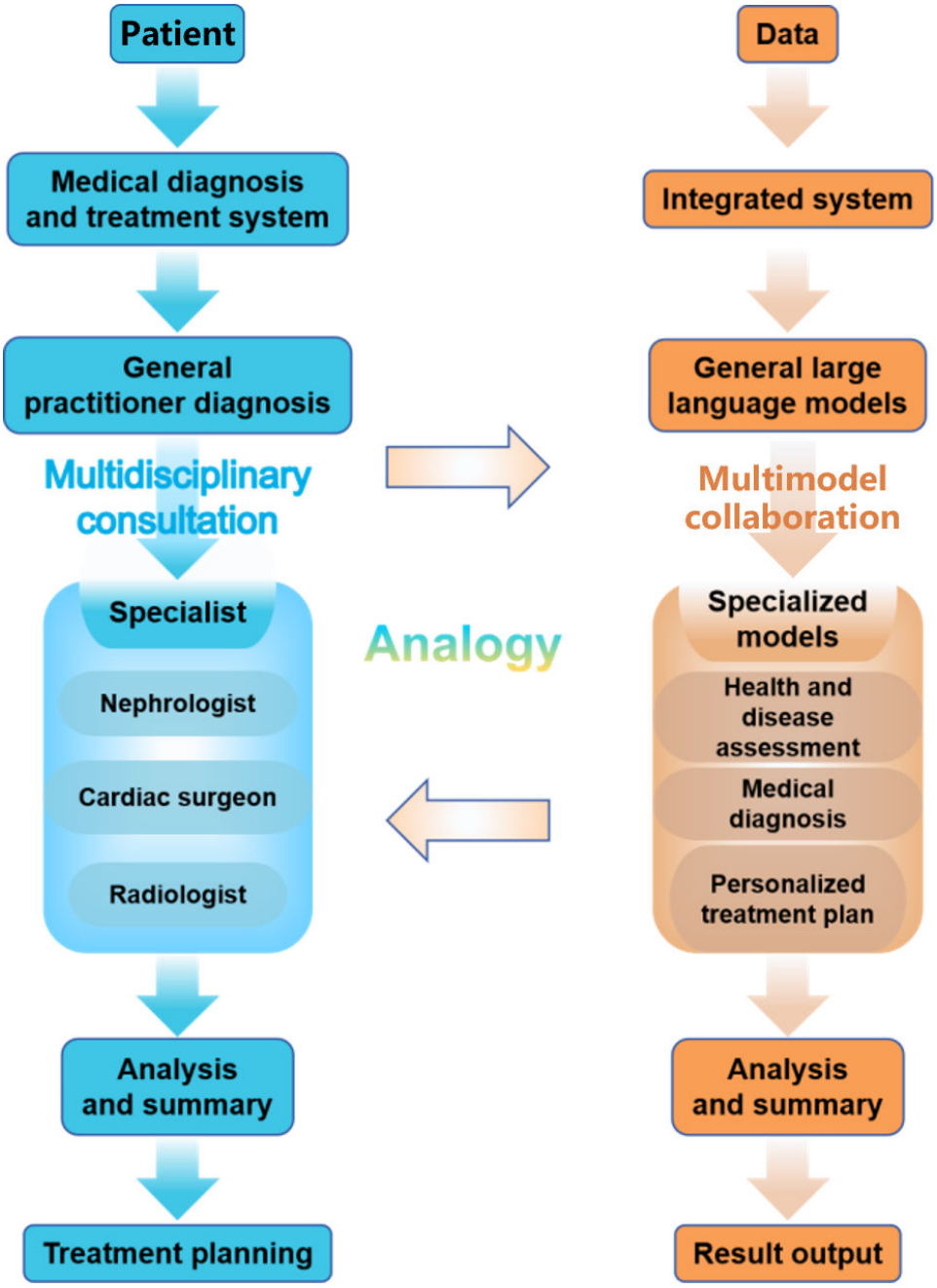
Based on the theoretical framework of symbiotic agency [11], a theory emphasizing interdependent and collaborative relationships between humans and technology, this study conceptualizes human-technology relations as a process of mutual constitution. Technology functions neither as a passive instrument dominated by humans nor as an autonomous replacement for human agency. Instead, it develops in tandem with humans through interdependent interactions: technology enhances human efficacy by expanding cognitive boundaries and enabling novel multimodal interactions, while humans legitimize technological practice by embedding ethical norms and conducting context-specific interpretations such as weighting clinical decisions. This symbiosis transcends traditional master-servant dichotomies by establishing a responsibility-sharing network. Within this network, technology acts as a co-agent in human activity systems, collectively enhancing capabilities rather than substituting human roles. This perspective provides the foundational understanding needed to maintain a dynamic balance in human-technology interdependence within medical education, forming the basis of our conceptual model.

Conceptual Model 1: Specialized Models Integrated System Based on General Large Language Models

Building upon the analytical framework established in Table S1 in Multimedia Appendix 1, which systematically compares general-purpose and domain-specialized GAI models across 3 critical dimensions (knowledge representation fidelity, task compatibility, and ethical constraint mechanisms), this study deconstructs technological heterogeneity to avoid conflating "GAI" as a homogeneous entity. The models in Multimedia Appendix 1 (see Table S1) were selected via multisource evidence synthesis, including peer-reviewed studies, industry reports (eg, Global Large Language Model (LLM) Market Research Report 2024) [12-20], and empirical validation in educational contexts, based on four criteria: (1) technological representativeness of core advancements (multimodality, reasoning, and domain adaptation); (2) broad academic and practical relevance in medical education; (3) functional diversity covering text, image, video, and domain-specific tasks; and (4) market prevalence, wide recognition, technical maturity, and development by prominent AI companies. Notably, models like Perplexity, DeepSeek, Notebook LM, and Midjourney, though used by clinicians and students in specific scenarios, were not included due to limited evaluative data and insufficient supporting information in the referenced reports.

Within this ecosystem, general LLMs serve as multitasking hubs, leveraging cross-domain adaptability and natural language interaction, while specialized models achieve context-specific efficacy through the embedding of deep medical knowledge. To resolve their complementary yet fragmented coexistence, we propose a specialized model integration system anchored to general LLMs, inspired by symbiotic agency theory and hospital diagnostic workflows [21] (see Figure 1). This architecture establishes a 3-tiered clinical analog: general LLMs serve as primary coordinators, managing task orchestration; specialized models act as domain experts, executing depth-specific processing; and protocol-based collaboration enables online consultation through knowledge distillation and output validation. This hierarchical integration embodies symbiotic agency principles: general models extend the applicability of specialized techniques by transcending domain boundaries, while specialized models enhance system depth by reinforcing medical logical rigor. Through functional complementarity and role differentiation, they form a synergistic symbiont exceeding individual capability limits, establishing an intelligent foundation for medical education characterized by adaptability, expertise, and reliability.

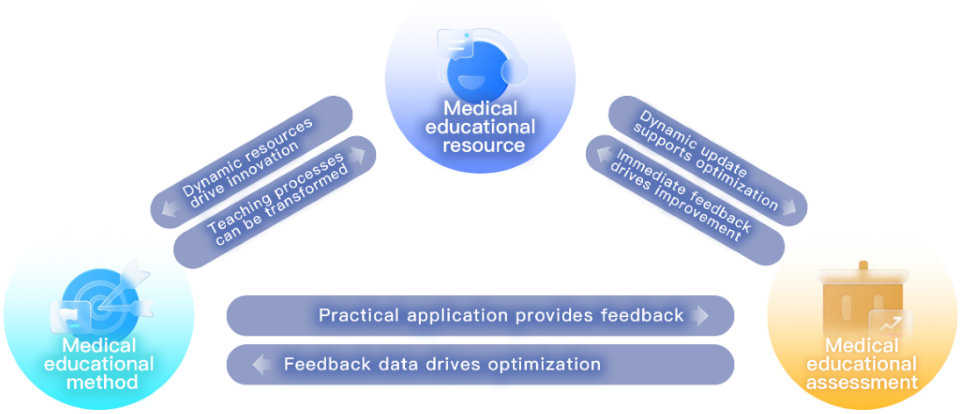
Figure 1. Specialized model integration system based on general large language models.



Conceptual Model 2: Tripartite Synergistic Integration Model for Medical Education Resources, Methods, and Assessment

The tripartite synergy paradigm, rooted in complex systems management theory and evidenced across domains from political governance to integrated health care systems (eg, the mission alignment model by Peek et al [22]), establishes our resource-method-assessment (RMA) framework (see Figure 2) as the core analytical structure [22,23]. This framework defines three interdependent dimensions: (1) resources encompassing dynamic content provisioning mechanisms, (2) methods designing knowledge-to-practice training pathways, and (3) assessment managing outcome monitoring and feedback generation. Their cyclical optimization forms an integrated whole, as resource renewal enables pedagogical innovation, method implementation yields evaluative data, and assessment outputs drive resource refinement and method calibration. Within this architecture, GAI operates as a collaborative instrument executing content generation, interaction support, and data analysis under educator-directed goal design, ethical governance, and critical intervention. The established framework provides essential categorization criteria for subsequent empirical analysis: it defines 3 dimensions—resources, methods, and assessment—directly corresponding to 3 primary research domains in GAI applications for medical education. By consolidating fragmented literature within a unified analytical structure, this framework systematically addresses cognitive limitations arising from isolated examinations of technological functions, thereby elucidating the intrinsic operational logic of technology-enabled educational transformation.

Figure 2. The model of integrated and collaborative development of medical education methods, resources, and assessment.



Methods

Review

With the rapid development of GAI, its applications in medical education have garnered considerable attention and have become a significant research focus. We conducted a preliminary search using the keyword combination of "generative artificial intelligence" and "medical education" across PubMed, Web of Science, and Scopus. Our goal was to analyze the publication trend regarding the applications and challenges of GAI in medical education over the past 5 years (from January 2020 to October 2024). The literature search was limited to sources published between January 2023 and October 2024 for the following reasons: (1) Technological progression: The 2023-2024 period coincides with a shift from theoretical proposals (pre-2023) to empirical studies on GAI implementation in medical education. (2) Scope alignment: The review prioritizes analysis of current applications, identified limitations (eg, output inaccuracies and integrity concerns), and near-future developments rather than historical trends. (3) Avoiding redundancy: Pre-2023 literature is excluded to prevent overlap with existing syntheses and focus on emergent applications (eg, automated assessment and adaptive resources) evidenced in the sampled literature (n=131). (4) Practical relevance: This timeframe reflects consolidated evidence on operational challenges and benefits relevant to contemporary pedagogical decision-making.

Search Strategy

We used Boolean operators to combine GAI and medical education keywords, creating the final search strategy (see Multimedia Appendix 2). A thorough search was conducted across 3 major databases: PubMed, Web of Science, and Scopus, focusing only on English-language articles published from January 2023 to October 2024.

Inclusion and Exclusion Criteria

This study included research articles focusing on the applications, challenges, and future development of GAI in medical education applications. Articles were excluded if they were commentaries, editorials, viewpoint, perspective, and short reports or communications with low level of evidence or did not discuss GAI within medical education. Studies focusing on non-GAI forms such as predictive analytics and natural language processing or those centered on other fields of medical education (eg, nursing) were also excluded. We excluded nursing based on fundamental educational differences. Clinical and dental education follow structured undergraduate curricula focused on acute care, diagnostics, and technical skills within hospital settings. Nursing emphasizes community practice, longitudinal relationships, and chronic disease management [24]. Including nursing would introduce significant heterogeneity in learning outcomes, GAI applications, and educational contexts. This methodological exclusion preserves thematic coherence and internal validity for analyzing GAI's role in comparable, technology-driven medical education environments.

Initially, YL and ZL conducted a preliminary screening of titles and abstracts from 3 databases. With the help of Zotero 7.0.13 (64-bit), a document management software (it is a project of Digital Scholar and developed by a global community), ZL detected duplicates of the initially screened articles according to title, author, abstract, and other information and removed duplicates. Following this initial phase, YL and ZL independently reviewed the full texts for a second round of evaluation. In cases of disagreement, ZY and NZ were consulted to mediate and make the final determination regarding inclusion.

Data Extraction Protocol

To ensure the systematicity, transparency, and reproducibility of this scoping review, a detailed data extraction protocol was developed and rigorously followed.

Data Point Definition and Protocol Development

Before comprehensive data extraction, a structured data extraction form was collaboratively developed by all authors. This iterative process was guided by our research questions and the predefined thematic framework outlined in Table 1, which focused on the applications, challenges, and prospects of GAI in medical education applications. The form was designed to systematically capture key information from each included article, encompassing: bibliographic details (eg, authors, publication year, journal, and country or

region), study characteristics (eg, research design, objectives, and population), specific GAI models used (eg, ChatGPT [OpenAI] and Gemini [Google]), application scope (single-model vs multimodel), analysis type (performance comparison across models or examination of synergistic enhancement through model integration), detailed descriptions of identified applications, challenges, and future directions of GAI application in medical education categorized exclusively through our tripartite Trinity Framework and quantitative performance metrics (reported accuracy rates, percentages, mean scores, standard deviations, and *P* values related to GAI model performance in various tasks). This granular definition of data points ensured that all relevant information pertinent to our broad research inquiry was systematically collected.

Table 1. A systematic thematic analysis of applications and challenges of generative artificial intelligence (GAI) in medical education.

Category and theme	Subtheme
Medical educational assessment	
Scoring short answers automatically.	—[a]
Evaluating articles.	—
Medical educational resources	
Providing standard answers.	• The performance of different question types. • The performance of different difficulty questions. • The performance of questions at different cognitive levels.
Generating diverse clinical cases.	—
Digital interaction and communication training.	—
Sharing educational resources.	—
Generating clinical images.	—
Medical educational methods	
Curriculum design.	—
Generating customized teaching aids.	—
Generating explanations for MCQ[b].	—
Personalized learning support.	—
Medical decision aid.	—
Multidisciplinary knowledge acquisition.	—
Academic writing optimization.	—
Existing defects at this stage	
Insufficient scene adaptability.	• Poor ability to handle complex clinical scenarios. • Lack of local background in specific regions. • Language adaptability issues. • Lack of nontextual information analysis skills.
Data quality and information bias	• Hallucination phenomena. • Lack of details on output content. • Lack of personalization. • Dataset dependency.
Potential issues in the future	
Overreliance	• Impaired critical thinking. • Decreased creativity. • Decreased teamwork ability. • Decreased practical problem-solving ability.
Ethical controversy	• Authenticity of the test results. • Academic misconduct. • Lack of clinical interaction and emotional resonance. • Resource inequality. • Ownership of intellectual property rights. • “Black box” problem and attribution of responsibility.

[a]Not available.
[b]MCQ: multiple choice question.

To better understand the global research landscape in this field, we analyzed the countries or regions of origin for the 131 selected articles. For those without a precise location, we assigned them according to the country or region of the

corresponding author's institution. To analyze the distribution of research based on the countries or region's development level, we used the Human Development Index (HDI) classification. The latest HDI data categorizes countries or regions into 4 tiers: very high, high, medium, and low human development with higher HDI scores correlating with greater national development. We also investigated cross-level HDI collaborations, which refer to partnerships between countries from different HDI categories [25].

Protocol Testing and Quality Control

To validate the comprehensiveness and clarity of the data extraction form, a pilot test was independently conducted by 2 reviewers, YL and ZL, on a randomly selected subset of 10 included articles. During this pilot phase, any discrepancies in data extraction or ambiguities within the form were identified and discussed. Based on these discussions, the data extraction form underwent iterative revisions to refine categories, clarify definitions, and ensure consistent interpretation of data points among reviewers. Following this refinement, YL and ZL independently extracted data from all 131 included articles. In cases of disagreement between the 2 independent extractors, consensus was initially sought through discussion. If a consensus could not be reached, a third and fourth reviewer, ZY and NZ, were consulted to mediate and make final determinations regarding the applicability and extraction of the data.

Synthesis of Results

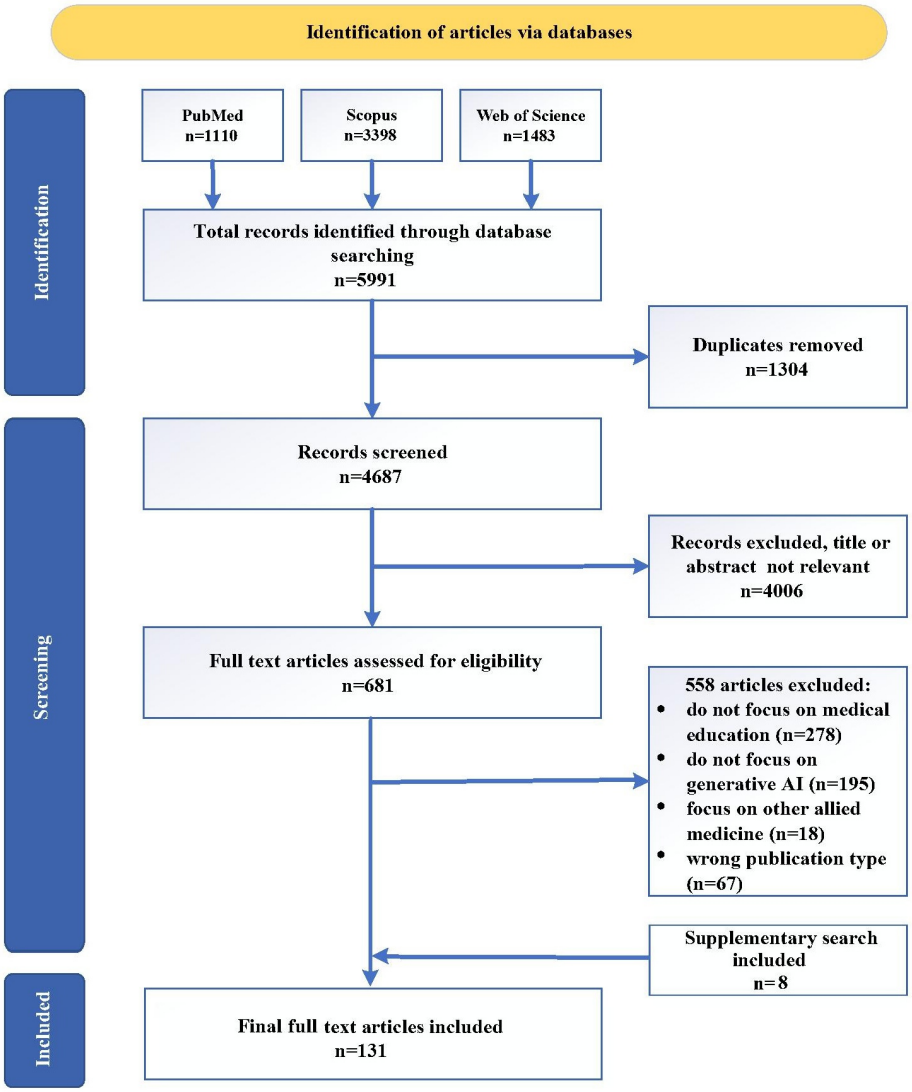
ZY subsequently compiled and reorganized the extracted data, assigning new identifiers for easier reference. This organized dataset was then categorized according to the predefined themes and subthemes (see Table 1), forming the basis for the subsequent descriptive summary and analysis. Our analysis employed a theory-driven, top-down approach anchored in a tripartite conceptual model of medical education: resource generation, method innovation, and assessment upgrade. The following sections will present a descriptive summary of the extracted data.

Results

Overview

Following our search strategy, we retrieved 5991 articles, of which 1304 were duplicates, leaving 4687 articles. In the first round of screening, 4006 irrelevant articles were excluded based on titles and abstracts, leaving 681 articles. In the second round, we excluded 558 articles after full-text review, including 278 nonmedical education articles, 195 non-GAI articles, 18 focused on other medical fields (eg, nursing), and 67 of different types (eg, commentaries). During the paper preparation, we conducted a supplementary search for 8 systematic reviews and meta-analyses. Ultimately, 131 articles were included in the final review (see Figure 3). Among the 131 included studies, the distribution of research designs was as follows: 83 cross-sectional studies, 5 randomized controlled trials (RCTs), 2 quasi-experimental studies, 1 cohort study, 1 quasi-randomized controlled trial, 8 systematic reviews and meta-analyses, 5 mixed-methods studies, and 1 case study. The remaining 25 studies were categorized as "other" with nonstandardized research designs, which were not fitting typical epidemiological or evidence-based medicine classifications. Collectively, cross-sectional studies (descriptive research designs) constituted the majority (n=83), reflecting the emerging state of GAI in medical education, where most research focuses on initial application explorations, feasibility assessments, and user experience descriptions rather than hypothesis-driven experimental designs. Other study types, including RCTs, cohort studies, and systematic reviews, provided supplementary evidence on intervention effects, longitudinal trends, and synthesized findings, respectively.

Figure 3. Article screening flow chart. AI: artificial intelligence.



Analysis of Literature Source and Human Development Levels

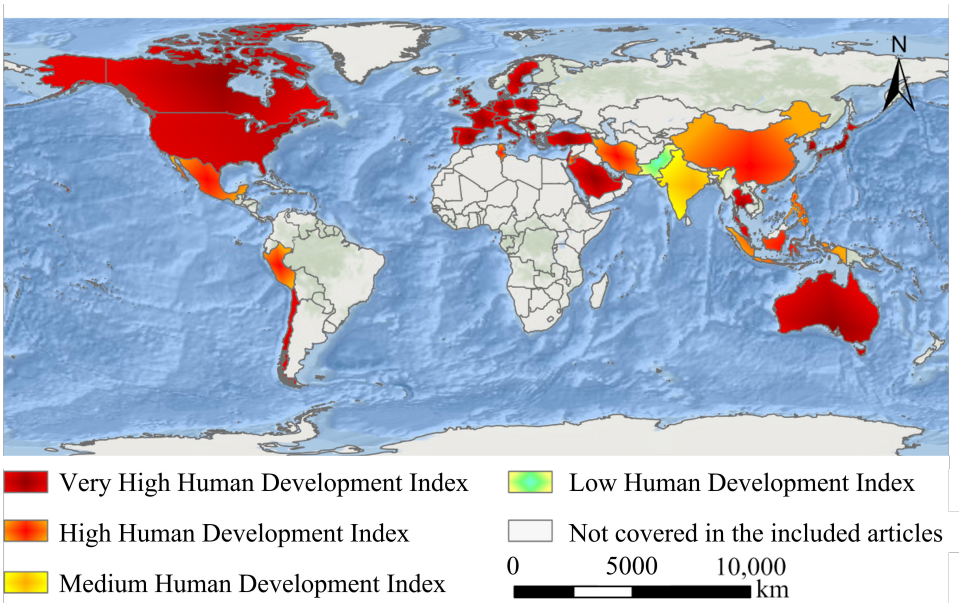
Based on the countries or regions of origin for the included articles and HDI classification, we analyzed the distribution of related studies. The results are illustrated in Table 2 and Figure 4. A significant portion (74%, n=97 articles) of the research came from countries or regions with very high human development, with the United States contributing 33 studies. High human development countries or regions produced 15% (n=19 articles), with China contributing 13 studies. Medium human development countries or regions contributed 5% (n=7 articles), mainly from India, while low human development countries or regions accounted for only 2% (n=3 articles). Furthermore, 4% of the studies (n=5 articles) involved cross-level collaborations, primarily between very high and medium or low HDI countries or regions.

Table 2. Distribution of countries or regions of origin for generative artificial intelligence (GAI) research in medical education (categorized by the HDI[a]).

HDI classification	Portion, n (%)
Very high human development	97 (74.0)
High human development	19 (14.5)
Medium human development	7 (5.3)
Low human development	3 (2.2)
Cross-level HDI collaboration	5 (3.8)

[a]HDI: Human Development Index.

Figure 4. Geographical distribution of countries or regions of origin for generative artificial intelligence (GAI) research in medical education.



Applications of GAI in Medical Education

Medical Educational Assessment

Scoring Short Answers Automatically

A recent study examined GPT-4 (OpenAI) and Gemini 1.0 Pro in automated short answer grading using 2288 student responses from 12 undergraduate medical courses across 3 languages, with instructor-provided rubrics or sample solutions as reference standards. GPT-4 showed high precision (0.91) in identifying fully correct answers, though its scores were significantly lower than human graders, while Gemini 1.0 Pro had no significant difference from human evaluations, with a mean normalized score of 0.68 (SD 0.32) and median of 0.75, similar to humans. Both models demonstrated high consistency across repeated evaluations, especially with high-quality standard responses, and these findings are specific to undergraduate medical education contexts [26].

Evaluating Articles

Liu et al [27] reported that in their study of 50 rehabilitation-related original articles, 50 sections (introductions, discussions, and conclusions) were generated by ChatGPT-3.5 and 50 were corresponding AI-rephrased versions using Wordtune Originality.ai, achieved 100% accuracy in detecting both AI-generated and AI-rephrased content. ZeroGPT correctly identified 96% of AI-generated texts and 88% of rephrased ones. The study focused specifically on rehabilitation medicine with analyzed content limited to partial article sections rather than full texts. It is notable that such high detection rates have not been widely observed across other disciplines or with newer large language model versions. The specialized nature of medical writing, including technical terminology use, may also influence these outcomes in ways not seen in broader academic contexts, which should be considered when evaluating the generalizability of these findings. Another study comparing automatic scoring systems (ChatGPT-3.5 and ChatGPT-4) with manual scoring for article quality assessment found no significant difference between GPT-4-based scoring and human grading. This demonstrates the considerable potential of GAI to enhance the quality evaluation of articles [28].

Medical Educational Resources

Providing Standard Answers

The performance of different question types: The studies encompassed a range of question types, including multiple-choice questions (MCQs), single-choice questions, short-answer questions (SAQs), true or false questions, open-ended short-answer questions (SOAQs), long-answer questions, clinical case analysis questions (CAQs), and image-text integrated questions [29-37]. An exploratory study conducted by a research team from Qatar University evaluated ChatGPT's performance across various assessment formats relevant to undergraduate dental education. The study included 50 assessment items covering 50 different learning outcomes, with 10 items for each of the 5 formats: MCQs, SAQs, short essay questions (SEQs), single true or false questions, and fill-in-the-blank items. These items were based on core clinical topics in dental education, such as restorative dentistry, periodontics, endodontics, and oral surgery, aligned with the learning outcomes expected of undergraduate dental students. In this study, ChatGPT demonstrated 90% accuracy for SAQs, SEQs, and fill-in-the-blank items and notably achieved 100% accuracy in single true or false questions [31]. However, other studies have revealed a significant decline in accuracy for CAQs, as low as 17%, which require strong logical reasoning and lack predefined options [34].

Regarding MCQs, a study reported that GPT-4 and Microsoft Bing achieved top scores (76%) on the University of Antwerp medical licensing MCQ exam, outperforming medical students. However, ChatGPT's accuracy fell considerably when tackling Chinese-language medical MCQs with an accuracy of 37%. In addition, another study reported

that in the Chinese Master's Degree Entrance Examination, ChatGPT's accuracy for single-choice questions (A1 type) was 56%, whereas for MCQs, it dropped to 33% [15]. These findings suggest that GAI's performance is not uniformly robust across all MCQ types and is influenced by factors such as question structure, subject domain, difficulty, language, and the presence of clinical vignettes or images.

The performance of different difficulty questions: In terms of difficulty, questions were generally classified as "easy," "medium," or "difficult." For example, the difficulty levels are defined based on the performance indicators of the historical question bank: "Difficult" (P<.30; less than 30% of the students answered correctly), "Medium" (P=.30 to .80), and "Easy" (P>.80) [35]. ChatGPT-4 demonstrated strong performance on easy questions, with accuracy rates reaching 97.4%. Yet, even in this category, ChatGPT-4's performance lagged behind that of residents [35,38-42]. In contrast, ChatGPT-4 excelled on medium and difficult questions, outperforming residents by 25.4 and 24.4 percentage points, respectively [38]. Across all models, performance tended to decline with increased difficulty, especially for higher-level questions that required multistep reasoning, where accuracy dropped markedly [39-45].

The performance of questions at different cognitive levels: 2 studies investigated the performance of ChatGPT-4 on questions categorized by Bloom's taxonomy, which includes 6 cognitive levels: remembering, understanding, applying, analyzing, evaluating, and creating [46]. These studies found that ChatGPT-4 consistently performed well across all cognitive levels, with an average correct answer rate of 71.96% for each cognitive level [47,48].

Generating Diverse Clinical Cases

By collaborating with instructors, GAI can quickly generate comprehensive clinical cases, including patient history, physical examination results, lab data, and differential diagnoses tailored to predefined learning objectives (eg, chest pain and joint pain). This reduces the time instructors spend developing such cases [49,50]. Furthermore, GAI-generated cases can integrate various contextual factors such as race, occupation, and lifestyle, significantly enriching the diversity of teaching materials [51]. For example, when creating a case based on a disease profile specific to a region, the ethnicity of the generated patient can be adjusted accordingly. In the context of type 2 diabetes, modifications can be made to the age range and weight distribution. In addition, randomized prompts for urine analysis may be included in urinary tract infection cases. Both patient presentations and examination findings can be randomized, and symptom expression can be customized to meet specific learning needs [51].

In Smith and colleagues' [52] study, GAI was assigned the task of creating a case of an immigrant with mental health concerns, as this group may require specialized social psychiatry interventions. The results indicated that GAI was able to produce a case that met fundamental educational objectives. However, it included several signs of emotional disorders, highlighting a need for further refinement.

Digital Interaction and Communication Training

Studies have shown that GAI is effective in promoting interactive learning and providing practice in communication skills. GAI-powered simulation tools simulate changes in clinical conditions in scenarios such as advanced cardiac life support (ACLS) and intensive care unit (ICU) sepsis, prompting students to critically analyze whether their decisions are correct [53]. In addition, conversational GAI-created digital patients provide anesthetists with valuable training for patient interactions, reducing reliance on human actors while enhancing the flexibility and consistency of the training process [54]. These digital interactions create a safe space for repeated practice, providing dynamic learning experiences that traditional textbooks cannot match [52,55]. Furthermore, conversational GAI models, such as chatbots, can simulate the role of a professor, offering critical evaluations of literature and distilling complex research into easily understandable key findings, thus fostering simulated discussions between students and experts in the field [56]. However, besides experiences and qualitative observations, formal evaluations of the reliability and validity of such GAI-generated information in a professor-like capacity are still needed.

Sharing Educational Resources

By generating accessible public health information, GAI enhances the public's understanding of essential health issues, such as infectious disease prevention and vaccination, ultimately leading to improved health literacy [30]. Furthermore, GAI-generated clinical cases can be disseminated as Open Educational Resources (OERs), providing medical educators with globally adaptable teaching materials that are customized to local contexts [51].

Generating Clinical Images

GAI tools such as Adobe Firefly, DALL·E 2 (OpenAI), Bing Image Creator, and generative adversarial networks (GANs) can create clinical images displaying various pathological features based on textual descriptions, potentially addressing the shortage of authentic pathological images in traditional medical education due to medical confidentiality and patient privacy restrictions [51,57-59]. For instance, images of retinal disease generated by the stable diffusion model enhance students' learning opportunities in ophthalmic pathology, greatly enhancing the availability of visual teaching resources [57]. However, their reliability and accuracy vary significantly across models and tasks. For example, DALL·E 2 demonstrated an overall clinical accuracy rate of 22.2% in aligning generated images with textual prompts across 15 semantic relations (eg, spatial and action-based relationships), with only 3 relationships (touching, helping, and kicking) achieving moderate consistency above 25%. In a medical education context, DALL·E 2 achieved 78% accuracy for soft-tissue tumor images but produced inconsistent results for wound images, with 65% of generated wound images containing anatomical inaccuracies or irrelevant elements [58]. A comparative study of DALL·E 2, Midjourney, and Blue Willow for generating skin ulcer images showed

DALL·E 2 performed best with an average score of 3.2/5 (scale 1-5) but still produced irrelevant content (eg, X-rays instead of pressure ulcers) in 20% of cases. Midjourney generated stylized, exaggerated features in 40% of images, while Blue Willow produced images with little relevance to prompts in 70% of attempts [59].

Medical Educational Methods

Curriculum Design

GAI shows potential in the early stages of curriculum development, aiding in quickly creating course objectives, learning strategies, and frameworks. For instance, in a study on integrated pharmacotherapy of infectious disease education modules, ChatGPT helped design curriculum goals (eg, "describe mechanisms of antibiotic resistance") with an average expert rating of 92% for appropriateness and accuracy, supporting educators—especially in designing foundational courses [60].

Generating Customized Teaching Aids

Researchers have developed derivative applications based on classical models, such as Glass AI (a powerful AI-driven knowledge management system developed by Glass Health, focusing on organizing and retrieving health-related information efficiently). It integrates GPT-4 with evidence-based, peer-reviewed clinical guidelines to generate differential diagnoses and clinical plans based on textual input of clinical cases, enabling students to interact with it and experience the GAI-driven diagnostic process for cases [61]. Similarly, an MCQ generator based on ChatGPT-generated cases offers a dynamic platform for personalized learning assessments [62].

Generating Explanations for Multiple-Choice Questions

Research shows that when GAI is used to answer MCQs, the explanations generated by GAI can better convey key knowledge points and achieve good accuracy and degree of matching with teachers' explanations. Of the 81 questions explained by the teacher and correctly answered by ChatGPT, 92.6% of the explanations were accurate and included at least part of the teacher's explanation. However, the research also highlights that if an initial response is incorrect, the likelihood of subsequent errors increases significantly (P<.001), indicating that an early mistake may lead to systematic inaccuracies in later explanations [63]. Complementing this, in a systematic review, the broader literature reviewed showed that the majority of studies (5/8, 62.5%) indicate the effectiveness of AI in generating valid MCQs, with a preference for the latest GPT-4 models (6/8, 75%) [64].

Personalized Learning Support

Studies demonstrate that GAI boosts students' learning efficiency across multiple stages by offering personalized feedback and customized content. This includes support for exam preparation [55,65-68], optimizing learning paths and review strategies [52,69-71], clarifying medical concepts [68,72-76], and assisting in the development of tailored career plans [77]. For instance, in physiological case analysis, GAI offers precise responses and contextually relevant feedback. A cross-sectional study tested 77 physiology case vignettes (covering diverse physiological and pathophysiological scenarios, designed for undergraduates) on ChatGPT 3.5, Google Bard, and Microsoft Bing. Rated by two physiologists on a 0-4 scale, ChatGPT scored highest at 3.19 (SD 0.3), outperforming Bard (2.91, SD 0.5) and Bing (2.15, SD 0.6) with P<.001. ChatGPT's precision accelerates task completion, helping students grasp medical knowledge in practical scenarios more effectively [78]. Furthermore, a study found that in cases of initial incorrect responses, GPT-4 was able to self-correct and provide accurate answers after simple follow-up questions or hints, mimicking pedagogical interactions observed in residency programs. This dynamic learning approach, coupled with rapid information processing, positions GPT-4 as an important asset for personalized learning [79].

Medical Decision Aid

GAI uses its ability to analyze complex, domain-specific knowledge to support the diagnosis of rare and intricate diseases. In addition to diagnosis, it can generate differential diagnoses tailored to the unique characteristics of each disease, providing health care professionals with precise decision-making support [80-83]. For common pathological issues and basic data analysis, GAI tools are efficient and accurate, helping pathologists organize their thought processes and expedite the initial diagnostic phases [84]. The impact of domain-specific training is profound. For instance, refined datasets in the surgical and anesthesiology fields enhance GAI's clinical decision-making capabilities. In scenarios such as a "30-year-old pregnant woman requiring an emergency appendectomy," GAI suggests not only tailored surgical strategies but also factors in critical anesthesia protocols [85]. Furthermore, in the field of traditional Chinese medicine, when combined with such tools, GAI can effectively create knowledge maps that organize entities, attributes, and their relationships to traditional Chinese medicine through graphical structures. GAI provides unique support for teaching traditional Chinese medicine and disease diagnosis and treatment decisions [86].

Multidisciplinary Knowledge Acquisition

GAI demonstrates potential in multidisciplinary knowledge acquisition within medical education by providing high-quality knowledge across various medical subfields [87-94]. GAI demonstrates adaptability across disciplines, including shoulder and elbow surgery, sports medicine, and oncology [91]. Research further indicates that GAI models such as ChatGPT-4 excel in internal medicine, pediatrics, obstetrics and gynecology, surgery, emergency care, and public health [88-90,92-94]. Notably, a study assessing ChatGPT-4's performance in the American Board of Family Medicine (ABFM) certification examination demonstrated its significant proficiency, with both the custom robot version (embedded in a specialized subenvironment designed to mimic examination conditions and given extensive

preparation resources) and the regular version (standard ChatGPT-4) achieving high correct response rates of 88.67% and 87.33% respectively, well above the passing threshold. This further highlights GAI's value in enhancing medical education within a multidisciplinary framework, making it a powerful learning support tool across a wide range of fields, including family medicine [95]. A meta-analysis of ChatGPT-3.5/4 across medical, pharmacy, dentistry, and nursing licensing exams revealed an overall accuracy of 70.1% (95% CI 65%-74.8%; *P*<.001). Performance varied significantly by field (Q=15.334; *P*=.002), with pharmacy having the highest rate (71.5%, 95% CI 66.3%-76.2%) and nursing having the lowest rate (61.8%, 95% CI 58.7%-64.9%). These results demonstrate GAI's potential to provide multidisciplinary learning support in health professions [96]. It is crucial to note that the evidence presented in this section highlights the individual learner's ability to access and comprehend information across disciplines. This review's existing evidence has not yet extensively covered GAI's direct support for complex interdisciplinary teamwork, closed-loop communication, or the cultivation of specific professional behaviors within collaborative learning environments.

Academic Writing Optimization

A study shows that GAI excels in creating article outlines and editing formatting, which alleviates common writing challenges related to poor organization and grammatical mistakes [28]. In addition, GAI can significantly improve the quality and standardization of academic writing, allowing medical educators and students to express their ideas more accurately and clearly [28,55,97]. Furthermore, GAI assists students in organizing and generating literature content while writing their thesis [98]. The content produced by GAI maintains consistency in language and includes appropriate academic terminology and logical structure, helping students present themselves more professionally in their academic writing [55]. Furthermore, GAI supports many non-native English speakers in overcoming language barriers during the academic writing process, which enables them to engage more confidently in academic communication [71].

Statistical Analysis of the Application of GAI Models

The models discussed in 131 articles include ChatGPT, Gemini (formerly known as Bard), Copilot (formerly known as Bing), Claude, and LLaMA, as well as other types of models such as StyleGAN2-ADA, Stable Diffusion, and customized chatbots.

Among the various models studied, ChatGPT stands out due to its advanced natural language processing capabilities. Of the 131 articles, 119 (89.5%) focused on ChatGPT, which was applied in diverse educational contexts, including simulating doctor-patient conversations, generating exam questions, and providing personalized learning support. These applications highlight their flexibility and adaptability in medical education. Notably, research had shown that as versions have iterated, ChatGPT-4 has significantly improved in both performance and scope compared to ChatGPT-3.5 [26,94,99-101].

Gemini was mentioned in 22 articles, accounting for 16.5% of the total. Copilot was mentioned in 11 articles, primarily due to its integration with the Microsoft ecosystem, making it ideal for educational management and resource development. Claude was cited in 6 articles. LLaMA, referenced in 4 articles, stands out for its ability to run locally, making it suitable for environments with limited resources. In addition, StyleGAN2-ADA, Stable Diffusion, and Convai were discussed in individual studies, mainly for their use in image generation and visualizing doctor-patient interactions.

The performance assessment of two or more models was compared in 26 articles. In comparative studies within the articles, numerous models have undergone head-to-head research, including ChatGPT-4 with Gemini 1.0 Pro [26], ChatGPT-4 with ChatGPT-3.5 [96], ChatGPT 3.5 with Google Bard and Microsoft Bing [78], DALL-E 2 with Midjourney and Blue Willow [59], and Originality.ai with ZeroGPT [27]. Based on these head-to-head investigations, different models demonstrate proficiency in specific tasks: ChatGPT-4 performs better in handling complex tasks, providing accurate medical knowledge, generating exam questions, and offering personalized learning support, especially in English-language medical licensing examinations; Gemini 1.0 Pro is noted for its strong contextual understanding and multimodal capabilities; ChatGPT-3.5 excels in simulating doctor-patient conversations, generating exam questions, and providing personalized learning support; Microsoft Bing achieved top scores alongside GPT-4 in medical licensing MCQ exams; DALL-E 2 shows potential in creating clinical images with specific pathological features from textual descriptions; and Originality.ai achieves high accuracy in detecting both AI-generated and AI-rephrased medical writing.

Challenges of GAI in Medical Education

Existing Defects at This Stage

Insufficient Scene Adaptability

Insufficient scene adaptability is due to the following factors.

First is the poor ability to handle complex clinical scenarios. GAI faces substantial limitations when handling complex clinical scenarios, particularly in cases requiring multistep reasoning, intricate calculations, and recognition of atypical clinical symptoms [45,87,102,103]. For instance, studies have shown that GAI struggles with MCQs, X-type problems, and tasks demanding deep reasoning. This underscores its limited ability to perform the nuanced decision-making required in medical judgments [29,30,39,41,44,47,48,66,67,84,89-92,104-107]. Furthermore, GAI-generated clinical scenarios often lack flexibility and fail to replicate the diversity and complexity of real-life clinical environments, thereby limiting learners' exposure to the spectrum of challenging cases [108,109]. GAI also faces technical limitations in generating simulated images for complex diseases, resulting in images that fail to depict

atypical manifestations accurately [57]. Furthermore, GAI models demonstrate uneven knowledge depth, exemplified by an ophthalmology meta-analysis: accuracy was 78% in "Pathology" but significantly lower in foundational or clinical areas, such as "Ophthalmology fundamentals" (52%), "Clinical ophthalmology" (57%), and "Refractive surgery" (59%) [110].

Second is the lack of local background in specific regions. Numerous studies have shown that GAI often struggles to adapt effectively to a specific region's unique background and needs when dealing with medical content related to that region, thereby undermining its universal applicability in multicultural settings [33,38,102,111]. For example, ChatGPT often responds to public health issues in India with a Western-centric perspective, overlooking local situations and cultural differences [33]. Similarly, ChatGPT struggles to accurately comprehend and adapt to the local regulatory environment when addressing medical policies specific to China, largely due to the limited representation of Chinese data in its training set [102,112].

Third is language adaptability issues. Currently, GAI exhibits significant limitations in processing languages, particularly in non-English medical education environments. The accuracy of GAI models like ChatGPT often varies greatly when handling languages such as Chinese, Korean, and Polish, resulting in incorrect outcomes in these contexts [29,30,34,38,106,113-115]. A meta-analysis quantified disparity: GPT-3.5 achieved 57% accuracy (95% CI 52%-62%; $P<.01$) in English-speaking countries and 58% (95% CI 52%-64%; $P<.01$) in non-English-speaking countries ($P=.72$). GPT-4 scored 86% (95% CI 82%-89%; $P<0.01$) in English-speaking countries versus 80% (95% CI 76%-83%; $P<.01$) in non-English-speaking countries ($P=.02$), demonstrating the adaptability issues of GAI models across different linguistic and regional contexts [116].

Fourth is a lack of nontextual information analysis skills. Current GAI tools like ChatGPT and Bard struggle to handle image-based queries, limiting their application in fields such as dentistry, neurosurgery, and nuclear medicine, where visual analysis of images and tissue samples is crucial for clinical decision-making [31,36,42,67,73-75,117].

Data Quality and Information Bias

Data quality issues and information bias occur due to the following factors.

First is the hallucination phenomenon. In GAI applications, hallucinations occur when the content generated by GAI diverges from factual accuracy or contradicts itself, remaining a prevalent issue. In total, 3 primary types of hallucinations have been identified: input-conflicting hallucination, context-conflicting hallucination, and fact-conflicting hallucination. Input-conflicting hallucination occurs when the GAI-generated content contradicts the initial information provided by the user. This can mislead learners and hinder their understanding of specific concepts [51,65,118]. Context-conflicting hallucination arises when the GAI offers contradictory responses to the same or similar questions. This inconsistency is particularly evident in complex case analyses [71,90,119,120]. Fact-conflicting hallucination occurs when the GAI reports facts that contradict established information, often with a high confidence level, which can easily mislead learners [54,121-137].

Second is the lack of details on output content. Numerous studies have highlighted that GAI often generates overly simplified or vague responses, lacking essential details and knowledge necessary for a comprehensive understanding [31,53,56,60,63,73,114,118,120,135,136,138,139]. For instance, evaluations of GAI in cardiology have revealed that it fails to specify the types of heart murmurs associated with valve diseases. In addition, GAI-generated descriptions of pathophysiology and epidemiology tend to be overly general, often including vague statements such as "certain age groups are at higher risk" without specifying the specific conditions. Furthermore, GAI often produces incomplete or inaccurate information when generating case study materials, which can lead to misleading students. For example, GAI-generated learning materials on melanoma have been known to omit crucial tumor markers like S-100 or the latest treatment for BRAF (B-Raf proto-oncogene, serine, or threonine kinase) mutations [63,138]. The same problems are evident in academic writing assistance, where GAI may create basic article structures but often lacks the depth, detail, and critical citations found in human-generated content [120].

Third is the lack of personalization. The content generated by GAI lacks personalization tailored to individual needs. This limitation mainly manifests in the generated text, which often adopts similar writing patterns and standardized language, struggling to incorporate personalized perspectives or creative expressions [28]. In a medical environment, GAI-generated treatment plans, although generally reasonable, often fail to consider individual patient characteristics, such as the severity of the disease, lifestyle, and personal preferences [105].

Fourth is dataset dependency. The performance of GAI is significantly influenced by the quality and diversity of its training data. If the data is insufficient or skewed, it may lead to potential biases and limitations in practical applications, causing underperformance in less-represented areas [33,59,73,82,85,86,111,117,122,126,140-143]. In addition, the cutoff date for the training data means that GAI may lack knowledge of the latest research, leading to outdated or inaccurate recommendations [26,32,41,66,67,74,80,89,92,94,109,129,136,141,144]. For example, when advising on treatment for bipolar disorder in pregnant women, ChatGPT-4 failed to incorporate the latest studies and instead suggested outdated methods [89]. Furthermore, the data bias present in GAI during the training process cannot be overlooked [51-53,58,61,71,72,78,107,108,132,139,145]. Such biases often arise from the intrinsic imbalances within the dataset, which subsequently permeate the generated content. These biases manifest as stereotypes, mainly depicting certain professions or physical attributes. For instance, some occupations may be associated with higher BMIs, while the French ethnicity is

stereotypically linked to the profession of "wine connoisseur" [51].

Potential Issues in the Future

Overreliance

Overreliance can be caused due to the following factors.

First is impaired critical thinking. The rapid feedback provided by GAI may reduce students' time for deep thinking, weakening their ability to analyze problems and independently engage in critical learning. This phenomenon is particularly evident in medical education, where students often rely on the answers provided by GAI when solving complex problems rather than relying on their logical reasoning and knowledge accumulation for analysis and resolution [35,39,40,50,55,69,70,72,74,75,77,98,99,124,136,146-152].

Second is decreased creativity. When students use GAI tools like ChatGPT, they often receive writing suggestions that lack the creativity and depth of human-generated content. Thus, prolonged reliance on such tools may weaken their independent writing skills and hinder their ability to engage with complex topics that require critical thinking and practical expertise [28]. Similarly, educators who overly depend on GAI for content creation may stifle their curricular innovation, limit diversity and depth in teaching materials, and ultimately diminish the overall quality of education [60].

Third is decreased teamwork ability. Overreliance on GAI tools such as ChatGPT can weaken students' communication skills and ability to engage actively in collaborative teamwork [72,152]. Furthermore, the frequent use of these tools limits opportunities for meaningful interpersonal interaction with peers and mentors, hindering the development of essential teamwork and communication skills [152].

Fourth is decreased practical problem-solving ability. Practical problem-solving is essential for clinical decision-making and patient management. However, the convenience of GAI tools may lead students to rely on preexisting solutions, neglecting the deeper analysis and logical reasoning necessary to develop personalized answers [52,55,74,75,77,87,151-153]. Furthermore, using these tools may reduce interaction with mentors and peers, limiting students' opportunities to gain diverse perspectives through collaborative discussions and approach problems from multiple angles [147].

Ethical Controversy

Ethical controversies can occur due to the following factors.

First is the authenticity of the test results, as the integration of GAI in testing and assessment may compromise the accuracy and effectiveness of traditional methods used to evaluate students' actual capabilities. GAI-generated responses or GAI-assisted evaluations risk reflecting the performance of GAI itself rather than students' authentic abilities. This issue is evident in various exams, such as medical licensing and specialty exams, presenting new ethical challenges in medical education [27,50,78,99,144,146,154].

Second is academic misconduct, since GAI-generated content often evades traditional plagiarism detection tools, making it easier for students to exploit GAI tools to complete assignments or write papers without being detected, thus jeopardizing academic integrity [31,154]. In addition, the ease of using GAI to generate answers cultivates students' mindset of overreliance on such tools for academic tasks, which may increase their future likelihood of academic misconduct [70-72,124,155]. This issue extends beyond individual students and poses a broader threat to academic ethics, as GAI-generated content can be misinterpreted as original work, distorting academic evaluations [125,150].

Third is a lack of clinical interaction and emotional resonance. When addressing complex ethical or emotional medical issues, GAI lacks the empathy and emotional responsiveness inherent in human physicians, potentially undermining trust in the doctor-patient relationship [98,131]. This limitation is supported by a General Medicine In-Training Examination (GM-ITE) study comparing GPT-4 and Japanese residents. In the GM-ITE, "medical interview and professionalism" category assesses patient communication, ethics, and professionalism. It uses scenario-based questions (eg, addressing a terminally ill patient's anxiety or resolving treatment ethics). Responses are scored 0-10 based on communication appropriateness, empathy depth, and ethical application, with top marks for nuanced, human-centric judgment. Notably, GPT-4 scored 8.6 points lower here than residents [38]. Furthermore, because GAI tools do not provide an authentic, interactive experience or situational awareness, they may struggle to simulate the behavior and reactions of real patients accurately. This limitation makes it challenging for students to fully appreciate the importance of empathy and its application in doctor-patient interactions, which affects their development of communication and empathy skills development [38,54,76,77,101,107,147,152].

Fourth is resource inequality, which is most evident in the unequal access to technology and data. Datasets used for training GAI often exhibit biases, particularly involving data from different racial or socioeconomic backgrounds. This can worsen existing health care disparities. Furthermore, developing high-quality LLMs requires substantial computational resources, creating significant access barriers, especially for educational institutions or students with limited financial means. Hence, subscription fees and hardware limitations restrict their access to these GAI tools [67,74,85,134].

Fifth is the ownership of intellectual property rights. The widespread use of GAI in medical education raises numerous intellectual property concerns, particularly regarding copyright disputes related to the medical data used during AI training [113]. In addition, the legal status of GAI-generated content remains unclear, as current copyright laws do not adequately address the ownership of GAI-generated images and texts. This leaves the ownership of such content unclear, complicating the determination of whether the rights belong to the user, the developer, or other stakeholders [27,50,58,59,124].

Sixth is the "black box" problem and the attribution of responsibility. The application of GAI in medical education faces a significant challenge known as the "black box" problem. This issue arises from the lack of transparency and interpretability of GAI models, which directly affects the safety and reliability of these applications in medical settings. This lack of transparency makes it hard to understand how GAI reaches specific conclusions, especially when results are erroneous or biased, complicating efforts to trace and correct mistakes [86,88,148]. Furthermore, when GAI is used for diagnostic or clinical decision support, any errors or biases in its generated results can make it difficult to establish accountability. Trust in the doctor-patient relationship is built on clear responsibility. However, the lack of transparency in GAI models undermines this trust, leaving patients and physicians uncertain about the safety and reliability of GAI-driven decisions [36,114].

Discussion

Principal Findings

This scoping review systematically identifies 3 core characteristics of GAI in medical education through an analysis of 131 included studies: pronounced regional disparities, empowerment potential via RMA synergy, and unresolved technical and ethical challenges. These findings must be contextualized within the field's evolving landscape: Our initial screening retrieved 5991 articles, a striking number reflecting both the opportunities and challenges of this emerging domain. This vast volume can be attributed to GAI's rapid evolution as a nascent technology, where relevant concepts remain loosely defined and inconsistent. Consequently, keyword usage lacks standardization, often resulting in the inclusion of tangentially related cross-disciplinary studies. Furthermore, GAI's inherently interdisciplinary nature broadens the scope of relevant literature. While this abundance highlights widespread interest and diverse applications, it also emphasizes the lack of conceptual clarity and consistency in frameworks. Therefore, although research is progressing, the field remains in a transitional stage, moving from "conceptual standardization" to "unified frameworks." To propel the field forward, the academic community needs to reach a consensus on GAI-related definitions and application structures. Achieving this standardization will enable better tracking of emerging trends and facilitate the effective use of new insights.

Against this backdrop, regional distribution analysis reveals marked concentration of GAI research in very high HDI regions (74%), with minimal contributions from low-HDI regions (2%) and scarce cross-regional collaborations (4%), highlighting structural inequities in global technology diffusion. Model use patterns further demonstrate ChatGPT's dominant adoption (89.5%), driven by its superior performance in multifaceted educational tasks: (1) iterative version advancements (eg, GPT-4's significant improvements in reasoning accuracy and error reduction over GPT-3.5); (2) proven efficacy across diverse applications including clinical simulation, exam question generation, and personalized tutoring; and (3) robust multilingual support despite variability in non-English contexts. This technical versatility explains its preferential adoption by researchers. The disproportionately high usage rate of general LLMs over specialized models, coupled with a predominant focus on cross-model comparisons rather than synergistic integration, reflects insufficient exploration of technical adaptability and system interoperability within current research.

Within the RMA tripartite framework established in this study, GAI reshapes medical education through coordinated optimization across 3 dimensions. In resource provisioning, it effectively mitigates traditional constraints of specimen scarcity and privacy limitations through the efficient generation of diverse clinical cases and pathological images. Methodologically, it facilitates the transition from standardized instruction to personalized education through interdisciplinary knowledge integration and targeted learning support. For assessment, high concordance in automated scoring and academic integrity monitoring provides scalable solutions for educational quality assurance. This closed-loop optimization mechanism, which encompasses resource allocation, pedagogical implementation, and evaluative feedback, validates the framework's explanatory power for technology-enabled educational transformation.

Nevertheless, profound barriers impede deeper GAI integration. Current technical deficiencies manifest as: inadequate contextual adaptation (eg, limitations in complex clinical reasoning and MCQ processing), data quality flaws (including hallucinatory outputs and deficient nontextual information analysis), and linguistic or regional biases (particularly performance degradation in non-English contexts). Long-term risks include erosion of critical thinking and creativity due to overreliance, alongside ethical governance dilemmas that encompass ambiguous accountability, inequitable resource distribution, and deficient clinician-patient emotional engagement. These dual challenges constitute fundamental barriers to implementing human-AI collaboration paradigms.

Comparison With Existing Literature

This scoping review specifically focuses on the period between January 2023 and October 2024, a critical transitional phase where GAI in medical education shifted from theoretical exploration to practical implementation. By capturing this transformative era, it addresses the gap in previous reviews [1,9] that lacked coverage of the latest advancements. While building on the foundational insights of earlier studies, this review extends their scope by identifying emerging trends and practical applications that have emerged with GAI's maturation in educational contexts.

Our observation of pronounced regional disparities starkly aligns with and quantifies the well-documented "digital divide" prevalent in global health technology diffusion [156]. However, this study provides concrete, GAI-specific evidence within medical education, highlighting the extreme concentration and the critical scarcity of cross-tier collaboration, thereby reinforcing concerns about equity

in accessing transformative educational technologies and potentially exacerbating global health workforce inequities.

Regarding model use, the overwhelming dominance of ChatGPT mirrors its widespread popularity in GAI application studies [157]. Yet, our analysis delves deeper than mere prevalence reports or bibliometric study [158-160], specifically attributing this dominance to its rapid iteration (eg, GPT-4's improvements), proven versatility across key educational tasks (clinical sim, QG, and tutoring), and relatively robust (though imperfect) multilingual support, which are crucial for adoption in the diverse contexts of medical education research.

Our development of the RMA tripartite framework represents a key theoretical departure. While existing research acknowledges GAI's impact on discrete educational facets (resource provision, teaching methodologies, and evaluative processes), a unifying framework that binds these elements into a synergistic, closed-loop optimization mechanism is conspicuously absent from the current discourse [1,9,10]. Such a framework uniquely conceptualizes these three dimensions as an interdependent, dynamic closed-loop system essential for understanding GAI's holistic transformative potential. Crucially, the empirical identification of significant RMA imbalance (robust exploration of educational methods and resources vs sparse focus on learner assessment) does not imply that assessment is under-prioritized in education broadly, but rather reflects a current skew in GAI-medical education integration—with research disproportionately focusing on resource enrichment and methodological optimization, while lagging in the development of learner assessment applications [161]. This imbalance, viewed through our novel integrative lens, offers a structured diagnostic for the systemic gap in aligning GAI capabilities with the specific needs of learner assessment within medical education.

The unresolved technical-ethical challenges documented (eg, contextual limitations, hallucinations, biases, erosion of critical thinking, and concerns about empathy) resonate strongly with growing critiques of LLMs in healthcare [162,163]. Our review explicitly maps these well-recognized limitations onto the sensitive context of medical education, highlighting their manifestation and potential impact in shaping future clinicians. This reinforces concerns raised elsewhere but grounds them firmly in the educational domain.

Another distinctive contribution of this review lies in revealing a critical technological imbalance: the overwhelming focus on general-purpose LLMs like ChatGPT contrasts sharply with the lack of systematic development of specialized medical models and the near absence of research on multimodel collaborative mechanisms within medical education [10,164]. This finding highlights a gap in the current technological approach, which hinders depth and clinical authenticity. While previous studies used available tools, our synthesis highlights this specific limitation as a barrier to deeper integration.

Implications of the Findings

Implications for Educational Practice

This study makes a key contribution to pedagogical practice by establishing the RMA tripartite framework and revealing its developmental imbalances, thereby providing a practical paradigm for the integration of GAI into medical education. The core value of this framework lies in elucidating the dynamic closed-loop nature of technology-enhanced education, wherein resource provision establishes the pedagogical foundation, methodological innovation activates knowledge transformation, and assessment feedback drives systemic evolution; these 3 components constitute an interlocking educational mechanism [165].

As evidenced in the results section, the current imbalance, characterized by rich exploration in GAI-supported educational resources and teaching methods yet relatively limited progress in GAI-driven automated evaluation of learner performance, stems from an overemphasis on short-term efficiency in early technology adoption. This has led to systemic neglect of assessment's role as an optimization tool. For example, GAI is widely used to generate diverse clinical cases and pathological images to enrich educational resources and design adaptive learning pathways to innovate teaching methods. However, in learner assessment, most GAI tools still rely on simple automated scoring of knowledge-based quizzes, with few leveraging GAI to evaluate higher-order competencies such as clinical reasoning or diagnostic accuracy [26]. Another instance is that many researchers use GAI to create interactive simulation scenarios as a methodological advancement but fail to integrate automated assessment features that track learners' decision-making processes in these scenarios [53]. This misses opportunities to use assessment data to refine the scenarios themselves. Overreliance on GAI for resource and method innovation without matching progress in automated learner assessment risks disconnecting what is taught or provided from what learners need to master, ultimately limiting GAI's ability to drive meaningful change in medical education.

Achieving optimal integration requires establishing a bidirectional enhancement cycle centered on assessment. Automated assessment data capturing learning bottlenecks should guide the real-time expansion of clinical case libraries' pathological spectra and difficulty calibration [166], shifting resource provision from one-size-fits-all to demand responsiveness. Simultaneously, the focus on core competencies (such as clinical reasoning and problem-solving) emphasized in teaching methods must be integrated into new assessment dimensions [167], driving teaching methods to evolve from mere knowledge transmission to competency development. Within this cycle, assessment functions not merely as a quality monitoring tool, but as the central nexus for the co-evolution of resources and methods.

Realizing this vision necessitates educators reconceptualizing operational logic [165]. This involves using assessment data to inform the development of educational resources, specifically leveraging insights into learners' knowledge gaps

and skill deficiencies to dynamically adjust the complexity of clinical cases [168], embedding real-time, practical, and contextual feedback mechanisms within high-order teaching activities like simulated diagnostics to optimize pedagogical strategies [169], and establishing adaptive rules enabling cross-dimensional interaction to facilitate systemic iteration [170,171]. Collectively, this structural transformation elevates the tripartite framework into an organic educational operating system.

However, technological integration inherently presents dual challenges, highlighting the importance of upholding core principles of human-AI collaboration. Generating educational resources without clinical context review risks reinforcing data biases [172]; methodological innovation overly reliant on algorithmic decisions may erode critical thinking [9]; and automated assessment replacing human judgment may overlook students' psychological needs, reducing course engagement and well-being scores [173]. These manifestations of technological alienation arise from the partial ceding of human agency. Resolution lies in upholding a human-AI symbiotic vision: recognizing GAI as a collaborator, not a replacement, in educational evolution. Specifically at the resource layer, clinicians and educators must oversee the development of educational resources (eg, clinical cases) to balance efficiency, ethics, and clinical authenticity [174,175]. At the method layer, educators should direct learning path design to integrate technological augmentation with pedagogical wisdom [176]. At the assessment layer, institutions should implement verification systems that combine human evaluation with machine automation, ensuring assessments balance efficiency with humanistic dimensions [173,177]. This reconfiguration of responsibilities positions technology as a tool and reaffirms human stewardship of education.

Implications for Technological Development

This study identifies a technological imbalance in the application of GAI within medical education. This imbalance is characterized by the dominance of large general language models, while the development of specialized models for specific medical disciplines has lacked systematic progress. This limitation restricts the depth of technology-enabled education and indicates a neglect of multimodel collaborative mechanisms within current research paradigms.

The study proposes an integrated system using general LLMs alongside specialized medical models, employing a hierarchical collaborative architecture to reshape the technological ecosystem of medical education. The core operational logic establishes a 3-tiered functional division: general models act as the central hub for teaching interactions, handling basic task parsing and process orchestration; medical specialized models, drawing on vertical domain knowledge bases, execute high-complexity core teaching tasks such as clinical reasoning and medical image generation; and a cross-model validation mechanism forms a closed-loop quality control system. This architecture adapts the hospital's multidisciplinary team approach to AI in education, aligning technological capabilities with the requirements of medical education for expertise, reliability, and contextual authenticity.

Within medical education, this integrated system can facilitate 3 key changes. First, it addresses limitations in specialized knowledge depth inherent in traditional general models, improving training efficacy for advanced clinical reasoning. Second, it leverages GAI's multimodal capabilities, which integrate text and image data, to address key issues in medical imaging education including shortages of teaching resources like rare pathological images and the limits of static materials in showing dynamic anatomical relationships. This support helps evolve pathology visualization from static atlases to interactive 3D simulations, letting students explore spatial structures and pathological changes more intuitively [178]. Third, it establishes a cross-model knowledge validation chain to automatically identify and correct typical logical inconsistencies and factual errors in general models, ensuring the academic rigor of teaching content. These changes collectively represent a paradigm shift from tool-assisted learning to intelligent teaching partnership systems [179].

Supporting the effective operation of this system requires targeted solutions to key technical challenges. The primary task involves developing specialized models with medical context adaptive capabilities, specifically enhancing their semantic parsing of unstructured clinical texts to address performance variability in complex case analysis [180,181]. Concurrently, it is necessary to construct dynamically evolving medical education datasets that incorporate cross-regional case spectra and multilingual clinical literature to systematically mitigate cultural biases and time-lag effects in training data [182]. Integrating privacy-preserving computation techniques like federated learning can enable secure data collaboration among institutions, continuously optimizing model localization and adaptation while safeguarding patient information security [183,184].

Implications for Policies and Governance

This study reveals a pronounced regional disparity in the application of GAI within the field of medical education. Specifically, regions with a very high HDI dominate research output in this domain, while contributions from the low-HDI areas account for only 2%. The scarcity of cross-tier collaboration between very high- and low-HDI areas further exacerbates this structural inequity in resource distribution. This imbalance epitomizes systemic inequalities within global knowledge production systems, rooted in 3 compounding barriers: inadequate computational infrastructure in resource-constrained settings impedes technological localization, proprietary restrictions on core models under patent regimes limit feasible technology transfer, and excessive reliance on clinical data from high-income countries compromises model adaptability to regional health care priorities. Without deliberate intervention, this self-reinforcing Matthew Effect cycle risks intensifying the global fragmentation of medical educational resources [185].

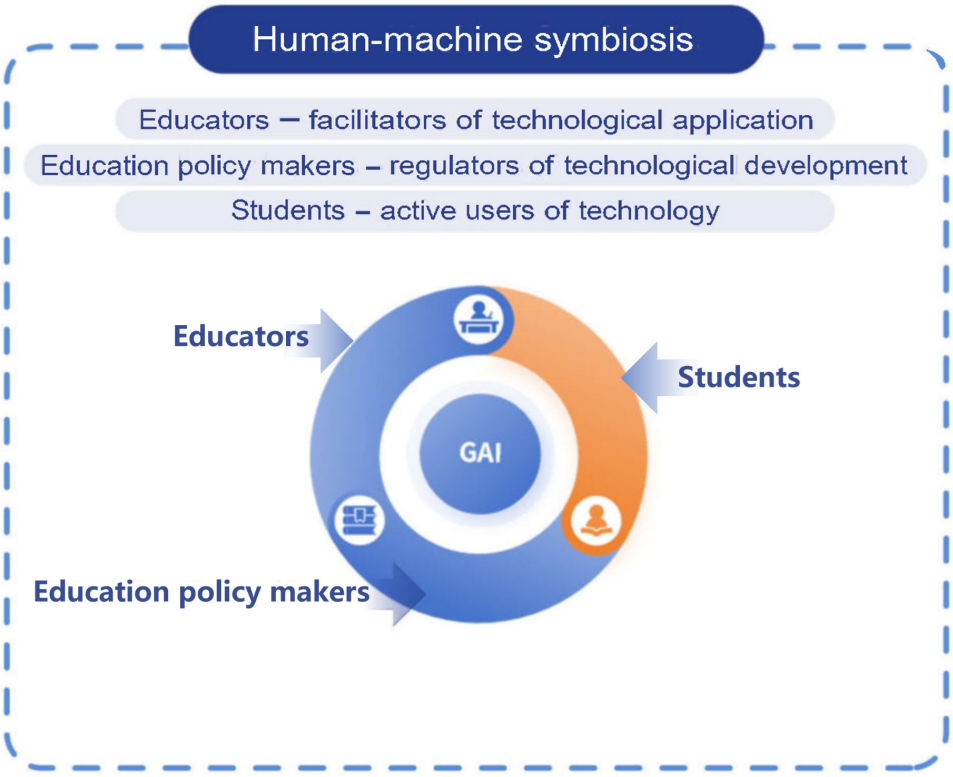
Addressing this complex challenge necessitates a multitiered governance framework. At the international level, binding technology-sharing agreements should request that holders of advanced models provide architectural access under fair-use principles, emphasizing the need to balance innovation with equitable access, while emulating open-source paradigms as a reference model [186]. Concurrently, the World Health Organization could coordinate multinational efforts to develop nonprofit medical corpora incorporating disease spectra prevalent in low-HDI regions, such as tropical and endemic diseases [187]. Nationally, ministries of education should integrate computational infrastructure into public medical education budgets [188] and similar to the Medical Education Partnership Initiative (MEPI) [189], establish dedicated funds for cross-border institutional partnerships to co-develop localized pedagogical tools that address specific regional educational needs. Institutionally, medical schools should adopt algorithmic transparency protocols, requiring deployed GAI tools to provide auditable model documentation that details the demographics and geographical coverage of training data. Fairness assessments of these tools should be carried out by multidisciplinary committees, which include clinicians, ethicists, and community representatives [190].

Simultaneously, institutional responses must address secondary risks through integrated technical, educational, and regulatory safeguards. To counter academic misconduct, educational institutions should implement dual-track verification systems that require GAI-assisted submissions to be accompanied by generation logs and validated through detection tools [191]. Academic journals must establish clear authorship standards declaring proportional human-GAI contributions [192]. Mitigating critical thinking erosion requires curriculum committees to incorporate GAI-free clinical reasoning assessments, such as on-site case analyses evaluating independent diagnostic and management planning capabilities as prerequisites for professional certification [193].

Technical deficiencies demand targeted interventions. Reducing model hallucinations requires dynamic fact-checking systems linking GAI outputs to authoritative medical knowledge bases, with confidence levels displayed during teaching platform usage [194]. To address the opacity of algorithms, where the process by which GAI models derive conclusions remains unclear, it is necessary to document the diagnostic reasoning processes of these models. Such documentation allows instructors to review the reasoning, helps determine accountability when inconsistencies occur, and can be integrated into resident training evaluations to strengthen oversight of GAI-assisted decision-making [195].

Fundamentally, governance paradigms must transition from a technocentric approach to symbiotic development. Compared to the commonly used “human in the loop” [196], which mainly emphasizes humans overseeing or making final decisions in AI systems, symbiotic agency theory goes further: it highlights mutual shaping between humans and AI. Humans guide AI development through ethical norms and clinical experience, while AI enhances human capabilities by expanding cognitive boundaries, forming a dynamic, mutually reinforcing relationship [11]. Policies should affirm human primacy in medical education, exemplified by reserving clinical empathy training exclusively for human instructors while limiting GAI to standardized case supplementation. An effective return to the essence of symbiotic agency means building collaborative mechanisms as shown in Figure 5: educators lead in setting teaching goals and ensuring ethical alignment (eg, reviewing GAI-generated cases to match real clinical logic); GAI supports personalized learning; students provide feedback to refine GAI tools; and policies clarify rights and responsibilities in this interaction. This human-centered approach ensures technological advancement aligns with pedagogical integrity and global equity imperatives.

Figure 5. Vision of human-machine symbiosis: a schematic diagram. GAI: generative artificial intelligence.



Limitations and Future Direction

This scoping review has several limitations that should be acknowledged. First, the rapidly evolving nature of GAI means our findings primarily reflect the landscape captured up to the search date; newer models and applications emerging subsequently may shift current patterns. Second, the inherent conceptual breadth and interdisciplinary nature of GAI pose challenges for exhaustive literature capture, potentially leading to omissions despite broad search parameters. Third, and most critically, while this study proposes 3 key conceptual frameworks (the RMA tripartite model, the hierarchical collaborative architecture, and the symbiotic agency principle) and argues for their feasibility based on synthesized evidence, it has not empirically tested their implementation or efficacy in authentic educational settings. Finally, reliance on published literature may underrepresent real-world implementation challenges and grassroots innovations.

Future research must bridge this critical gap by translating these frameworks into practice. Priority should be given to: (1) implementing and evaluating the RMA balancing strategies and the integrated system combining general and specialized medical GAI models in specific medical education contexts to assess their impact on learning outcomes and operational feasibility; (2) conducting longitudinal studies to track the dynamic evolution of GAI integration over time, observing its long-term empowerment effects on educational processes and outcomes; and (3) operationalizing the symbiotic agency framework to guide the design, deployment, and assessment of these interventions. This framework is essential for ensuring that human-AI collaboration in practice genuinely augments educator and learner agency, fosters critical competencies, and upholds pedagogical integrity, thereby realizing the envisioned synergistic educational ecosystem.

Conclusion

The application of GAI in medical education exhibits significant regional inequities, reflecting structural disparities in technological diffusion. Statistical findings from the model research reflect that researchers have certain preferences in its usage. The emergence of GAI has revitalized medical education, which is manifested in its promotion of the diversification of educational methods, the scientific evaluation of education assessment, and the dynamic optimization of education resources. However, these innovations are accompanied by current limitations and potential future challenges. By establishing the RMA tripartite model as a dynamic closed-loop system for educational optimization, proposing an integrated multimodel architecture to reconcile general and specialized GAI capabilities, and advancing the symbiotic agency principle to safeguard human primacy, this study provides foundational frameworks for navigating GAI integration. These contributions collectively address critical gaps in conceptual standardization and collaborative design, while delineating actionable pathways for pedagogical innovation, equitable technology development, and governance reform, which ultimately steer the field toward responsible human-AI collaboration that enhances clinical education without compromising pedagogical integrity or global equity.

Acknowledgments

This study was financially supported by the Funding of Medical Science and Technology Research in Guangdong Province, China (A2023363), the Industry-University-Research Collaborative Education Program of Ministry of Education, China (230905518284433), and the Teaching Reform Research Project of Clinical Teaching Base in Guangdong Province, China (2023-30).

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YL, ZL
Methodology: YL, ZL
Formal analysis: YL, ZL, ZY, NZ
Investigation: YL, ZL, ZY, NZ
Data curation: ZY, NZ
Writing – original draft: YL, ZL
Writing – review & editing: YC, ZC, XL
Visualization: ZY, NZ
Supervision: L Zhao, L Zhang
Project administration: L Zhao, L Zhang
Resources: YC, ZC, XL

Conflicts of Interest

None declared.

Multimedia Appendix 1

Technical features and application comparison of mainstream generative artificial intelligence (GAI) models.
[DOCX File (Microsoft Word File), 18 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Search strategy.
[DOC File (Microsoft Word File), 59 KB-Multimedia Appendix 2]

Checklist 1

PRISMA-ScR checklist.
[DOCX File (Microsoft Word File), 68 KB-Checklist 1]

References

1. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ. Oct 20, 2023;9:e48785. [doi: 10.2196/48785] [Medline: 37862079]
2. Generative AI market (2025 - 2030). Grand View Research. URL: https://www.grandviewresearch.com/industry-analysis/generative-ai-market-report [Accessed 2025-03-03]
3. Stretton B, Kovoor J, Arnold M, Bacchi S. ChatGPT-based learning: generative artificial intelligence in medical education. Med Sci Educ. Feb 2024;34(1):215-217. [doi: 10.1007/s40670-023-01934-5] [Medline: 38510403]
4. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
5. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ. Jun 6, 2023;9:e48163. [doi: 10.2196/48163] [Medline: 37279048]
6. Totlis T, Natsis K, Filos D, et al. The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. Surg Radiol Anat. Aug 16, 2023;45(10):1321-1329. [doi: 10.1007/s00276-023-03229-1]
7. Hanna JJ, Wakene AD, Lehmann CU, Medford RJ. Assessing racial and ethnic bias in text generation for healthcare-related tasks by ChatGPT1. medRxiv. Aug 28, 2023:2023.08.28.23294730. [doi: 10.1101/2023.08.28.23294730] [Medline: 37693388]
8. Densen P. Challenges and opportunities facing medical education. Trans Am Clin Climatol Assoc. 2011;122:48-58. [Medline: 21686208]

9. Xu T, Weng H, Liu F, et al. Current status of ChatGPT use in medical education: potentials, challenges, and strategies. J Med Internet Res. Aug 28, 2024;26:e57896. [doi: 10.2196/57896] [Medline: 39196640]
10. Temsah O, Khan SA, Chaiah Y, et al. Overview of early ChatGPT's presence in medical literature: insights from a hybrid literature review by ChatGPT and human experts. Cureus. Apr 2023;15(4):e37281. [doi: 10.7759/cureus.37281] [Medline: 37038381]
11. Neff G, Nagy P. Agency in the digital age: using symbiotic agency to explain human–technology interaction. In: Papacharissi Z, editor. A Networked Self and Human Augmentics, Artificial Intelligence, Sentience. 1st ed. Routledge; 2018:97-107. [doi: 10.4324/9781315202082-8] ISBN: 978-1-315-20208-2
12. The 22 best generative AI tools for SMBs to stay competitive in 2025. WebFX. URL: https://www.webfx.com/blog/marketing/best-generative-ai-tools/ [Accessed 2025-07-19]
13. Temsah MH, Alhuzaimi AN, Almansour M, et al. Art or artifact: evaluating the accuracy, appeal, and educational value of AI-generated imagery in DALL·E 3 for illustrating congenital heart diseases. J Med Syst. May 23, 2024;48(1):54. URL: https://sciety.org/articles/activity/10.21203/rs.3.rs-3895175/v1 [Accessed 2025-07-19] [doi: 10.1007/s10916-024-02072-0] [Medline: 38780839]
14. Claude 2: reviews, prices & features. Appvizer. URL: https://www.appvizer.com/artificial-intelligence/llm/claude-2 [Accessed 2025-07-19]
15. Global large language model (LLM) market research report 2024. QYResearch; 2024. URL: https://www.qyresearch.com/reports/2212992/large-language-model--llm [Accessed 2025-10-09]
16. OpenAI's o3 - AI model details. DocsBot AI. URL: https://docsbot.ai/models/o3 [Accessed 2025-07-19]
17. OpeneVidence. AITop10. URL: https://aitop10.tools/zh/detail/openevidence [Accessed 2025-07-19]
18. Sora Turbo: OpenAI's enhanced video generation model goes public. Neurohive. URL: https://neurohive.io/en/ai-apps/sora-turbo-openai-s-enhanced-video-generation-model-goes-public/ [Accessed 2025-07-19]
19. AI tools for medical education and research. Macon & Joan Brock Virginia Health Sciences at Old Dominion University. URL: https://www.evms.edu/about_us/ai_resources/resources_and_ai_tools/ai_tools_for_medical_education_and_research/ [Accessed 2025-07-26]
20. Cho J, Puspitasari FD, Zheng S, et al. Sora as an AGI world model? A complete survey on text-to-video generation. arXiv. Preprint posted online on Mar 8, 2024. [doi: 10.48550/ARXIV.2403.05131]
21. Hu H, Liang H, Wang H. Longitudinal study of the earliest pilot of tiered healthcare system reforms in China: will the new type of chronic disease management be effective? Soc Sci Med. Sep 2021;285:114284. [doi: 10.1016/j.socscimed.2021.114284]
22. Peek CJ, Allen M, Loth KA, et al. Harmonizing the tripartite mission in academic family medicine: a longitudinal case example. Ann Fam Med. 2024;22(3):237-243. [doi: 10.1370/afm.3108] [Medline: 38806264]
23. Geenens R, De Schutter H. A tripartite model of federalism. Philos Soc Crit. Sep 2023;49(7):753-785. [doi: 10.1177/01914537211066850]
24. Windak A, Rochfort A, Jacquet J. The revised European definition of general practice/family medicine. a pivotal role of one health, planetary health and sustainable development goals. Eur J Gen Pract. Dec 2024;30(1):2306936. [doi: 10.1080/13814788.2024.2306936] [Medline: 38334099]
25. Human development report 2023-24. United Nations Development Programme; Mar 2024. URL: https://hdr.undp.org/content/human-development-report-2023-24 [Accessed 2024-12-05]
26. Grévisse C. LLM-based automatic short answer grading in undergraduate medical education. BMC Med Educ. Sep 27, 2024;24(1):1060. [doi: 10.1186/s12909-024-06026-5] [Medline: 39334087]
27. Liu JQJ, Hui KTK, Al Zoubi F, et al. The great detectives: humans versus AI detectors in catching large language model-generated medical writing. Int J Educ Integr. May 20, 2024;20(1):8. [doi: 10.1007/s40979-024-00155-6]
28. Li J, Zong H, Wu E, et al. Exploring the potential of artificial intelligence to enhance the writing of english academic papers by non-native english-speaking medical students - the educational application of ChatGPT. BMC Med Educ. Jul 9, 2024;24(1). [doi: 10.1186/s12909-024-05738-y]
29. Li KC, Bu ZJ, Shahjalal M, et al. Performance of ChatGPT on Chinese master's degree entrance examination in clinical medicine. Grewal HS, editor. PLoS ONE. 2024;19(4):e0301702. [doi: 10.1371/journal.pone.0301702] [Medline: 38573944]
30. Cherif H, Moussa C, Missaoui AM, Salouage I, Mokaddem S, Dhahri B. Appraisal of ChatGPT's aptitude for medical education: comparative analysis with third-year medical students in a pulmonology examination. JMIR Med Educ. Jul 23, 2024;10:e52818. [doi: 10.2196/52818] [Medline: 39042876]
31. Ali K, Barhom N, Tamimi F, Duggal M. ChatGPT—A double-edged sword for healthcare education? Implications for assessments of dental students. Eur J Dental Education. Feb 2024;28(1):206-211. [doi: 10.1111/eje.12937]

32. Panthier C, Gatinel D. Success of ChatGPT, an AI language model, in taking the French language version of the European Board of Ophthalmology examination: A novel approach to medical knowledge assessment. J Fr Ophtalmol. Sep 2023;46(7):706-711. [doi: 10.1016/j.jfo.2023.05.006] [Medline: 37537126]
33. Gandhi AP, Joesph FK, Rajagopal V, et al. Performance of ChatGPT on the India undergraduate community medicine examination: cross-sectional study. JMIR Form Res. Mar 25, 2024;8:e49964. [doi: 10.2196/49964] [Medline: 38526538]
34. Yu P, Fang C, Liu X, et al. Performance of ChatGPT on the Chinese postgraduate examination for clinical medicine: survey study. JMIR Med Educ. Feb 9, 2024;10:e48514. [doi: 10.2196/48514] [Medline: 38335017]
35. Morreel S, Verhoeven V, Mathysen D. Microsoft Bing outperforms five other generative artificial intelligence chatbots in the Antwerp University multiple choice medical license exam. Banerjee I, editor. PLOS Digit Health. Feb 2024;3(2):e0000349. [doi: 10.1371/journal.pdig.0000349] [Medline: 38354127]
36. Guerra GA, Hofmann H, Sobhani S, et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. World Neurosurg. Nov 2023;179:e160-e165. [doi: 10.1016/j.wneu.2023.08.042] [Medline: 37597659]
37. Huang RS, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung FH. Assessment of resident and AI chatbot performance on the University of Toronto family medicine residency progress test: comparative study. JMIR Med Educ. Sep 19, 2023;9:e50514. [doi: 10.2196/50514] [Medline: 37725411]
38. Watari T, Takagi S, Sakaguchi K, et al. Performance comparison of ChatGPT-4 and Japanese medical residents in the general medicine in-training examination: comparison study. JMIR Med Educ. Dec 6, 2023;9:e52202. [doi: 10.2196/52202] [Medline: 38055323]
39. Terwilliger E, Bcharah G, Bcharah H, Bcharah E, Richardson C, Scheffler P. Advancing medical education: performance of generative artificial intelligence models on otolaryngology board preparation questions with image analysis insights. Cureus. Jul 2024;16(7):e64204. [doi: 10.7759/cureus.64204] [Medline: 39130878]
40. Revercomb L, Patel AM, Fu D, Filimonov A. Performance of novel GPT-4 in otolaryngology knowledge assessment. Indian J Otolaryngol Head Neck Surg. Dec 2024;76(6):6112-6114. [doi: 10.1007/s12070-024-04935-x] [Medline: 39559040]
41. Riedel M, Kaefinger K, Stuehrenberg A, et al. ChatGPT's performance in German OB/GYN exams – paving the way for AI-enhanced medical education and clinical practice. Front Med. 2023;10. [doi: 10.3389/fmed.2023.1296615]
42. Patel EA, Fleischer L, Filip P, et al. Comparative performance of ChatGPT 3.5 and GPT4 on rhinology standardized board examination questions. OTO Open. 2024;8(2):e164. [doi: 10.1002/oto2.164] [Medline: 38938507]
43. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ. Jun 29, 2023;9:e48002. [doi: 10.2196/48002] [Medline: 37384388]
44. Haddad F, Saade JS. Performance of ChatGPT on ophthalmology-related questions across various examination levels: observational study. JMIR Med Educ. Jan 18, 2024;10:e50842. [doi: 10.2196/50842] [Medline: 38236632]
45. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. Feb 8, 2023;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
46. Anderson LW, Krathwohl DR. A Taxonomy For Learning, Teaching, And Assessing: A Revision Of Bloom's Taxonomy Of Educational Objectives. Addison Wesley Longman, Inc; 2001. ISBN: 0-321-08405-5
47. 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):17. [doi: 10.3352/jeehp.2024.21.17] [Medline: 38977032]
48. Bharatha A, Ojeh N, Fazle Rabbi AM, et al. Comparing the performance of ChatGPT-4 and medical students on MCQs at varied levels of Bloom's taxonomy. Adv Med Educ Pract. 2024;15:393-400. [doi: 10.2147/AMEP.S457408] [Medline: 38751805]
49. Wong K, Fayngersh A, Traba C, Cennimo D, Kothari N, Chen S. Using ChatGPT in the development of clinical reasoning cases: a qualitative study. Cureus. May 2024;16(5):e61438. [doi: 10.7759/cureus.61438] [Medline: 38953081]
50. Shimizu I, Kasai H, Shikino K, et al. Developing medical education curriculum reform strategies to address the impact of generative AI: qualitative study. JMIR Med Educ. Nov 30, 2023;9:e53466. [doi: 10.2196/53466] [Medline: 38032695]
51. Bakkum MJ, Hartjes MG, Piët JD, et al. Using artificial intelligence to create diverse and inclusive medical case vignettes for education. Brit J Clinical Pharma. Mar 2024;90(3):640-648. [doi: 10.1111/bcp.15977]
52. Smith A, Hachen S, Schleifer R, Bhugra D, Buadze A, Liebrenz M. Old dog, new tricks? Exploring the potential functionalities of ChatGPT in supporting educational methods in social psychiatry. Int J Soc Psychiatry. Dec 2023;69(8):1882-1889. [doi: 10.1177/00207640231178451]

53. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT interactive medical simulations for early clinical education: case study. JMIR Med Educ. Nov 10, 2023;9:e49877. [doi: 10.2196/49877] [Medline: 37948112]
54. Sardesai N, Russo P, Martin J, Sardesai A. Utilizing generative conversational artificial intelligence to create simulated patient encounters: a pilot study for anaesthesia training. Postgrad Med J. Mar 18, 2024;100(1182):237-241. [doi: 10.1093/postmj/qgad137] [Medline: 38240054]
55. Magalhães Araujo S, Cruz-Correia R. Incorporating ChatGPT in medical informatics education: mixed methods study on student perceptions and experiential integration proposals. JMIR Med Educ. Mar 20, 2024;10:e51151. [doi: 10.2196/51151] [Medline: 38506920]
56. Brennan L, Balakumar R, Bennett W. The role of ChatGPT in enhancing ENT surgical training – a trainees' perspective. J Laryngol Otol. May 2024;138(5):480-486. [doi: 10.1017/S0022215123001354]
57. Tabuchi H, Engelmann J, Maeda F, et al. Using artificial intelligence to improve human performance: efficient retinal disease detection training with synthetic images. Br J Ophthalmol. Sep 20, 2024;108(10):1430-1435. [doi: 10.1136/bjo-2023-324923] [Medline: 38485215]
58. Seth I, Lim B, Cevik J, et al. Utilizing GPT-4 and generative artificial intelligence platforms for surgical education: an experimental study on skin ulcers. Eur J Plast Surg. Jan 29, 2024;47(1):19. [doi: 10.1007/s00238-024-02162-9]
59. Fan BE, Chow M, Winkler S. Artificial intelligence-generated facial images for medical education. MedSciEduc. Nov 14, 2023;34(1):5-7. [doi: 10.1007/s40670-023-01942-5]
60. Al-Worafi YM, Goh KW, Hermansyah A, Tan CS, Ming LC. The use of ChatGPT for education modules on integrated pharmacotherapy of infectious disease: educators' perspectives. JMIR Med Educ. Jan 12, 2024;10:e47339. [doi: 10.2196/47339] [Medline: 38214967]
61. Robleto E, Habashi A, Kaplan MAB, et al. Medical students' perceptions of an artificial intelligence (AI) assisted diagnosing program. Med Teach. Sep 2024;46(9):1180-1186. [doi: 10.1080/0142159X.2024.2305369] [Medline: 38306667]
62. Kıyak YS, Kononowicz AA. Case-based MCQ generator: a custom ChatGPT based on published prompts in the literature for automatic item generation. Med Teach. Aug 2, 2024;46(8):1018-1020. [doi: 10.1080/0142159X.2024.2314723]
63. Tong L, Wang J, Rapaka S, Garg PS. Can ChatGPT generate practice question explanations for medical students, a new faculty teaching tool? Med Teach. Mar 4, 2025;47(3):560-564. [doi: 10.1080/0142159X.2024.2363486]
64. Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. BMC Med Educ. Mar 29, 2024;24(1):354. [doi: 10.1186/s12909-024-05239-y] [Medline: 38553693]
65. Kawahara T, Sumi Y. GPT-4/4V's performance on the Japanese National Medical Licensing Examination. Med Teach. Mar 2025;47(3):450-457. [doi: 10.1080/0142159X.2024.2342545] [Medline: 38648547]
66. Tran CG, Chang J, Sherman SK, De Andrade JP. Performance of ChatGPT on American Board of Surgery in-training examination preparation questions. J Surg Res. Jul 2024;299:329-335. [doi: 10.1016/j.jss.2024.04.060] [Medline: 38788470]
67. Botross M, Mohammadi SO, Montgomery K, Crawford C. Performance of Google's artificial intelligence chatbot "Bard" (now "Gemini") on ophthalmology board exam practice questions. Cureus. Mar 2024;16(3):e57348. [doi: 10.7759/cureus.57348] [Medline: 38690460]
68. Gan W, Ouyang J, Li H, et al. Integrating ChatGPT in orthopedic education for medical undergraduates: randomized controlled trial. J Med Internet Res. Aug 20, 2024;26:e57037. [doi: 10.2196/57037] [Medline: 39163598]
69. Thomae AV, Witt CM, Barth J. Integration of ChatGPT into a course for medical students: explorative study on teaching scenarios, students' perception, and applications. JMIR Med Educ. Aug 22, 2024;10:e50545. [doi: 10.2196/50545] [Medline: 39177012]
70. Favero TG. Using artificial intelligence platforms to support student learning in physiology. Adv Physiol Educ. Jun 1, 2024;48(2):193-199. [doi: 10.1152/advan.00213.2023]
71. Ganjavi C, Eppler M, O'Brien D, et al. ChatGPT and large language models (LLMs) awareness and use. A prospective cross-sectional survey of U.S. medical students. PLOS Digit Health. Sep 2024;3(9):e0000596. [doi: 10.1371/journal.pdig.0000596] [Medline: 39236008]
72. Sallam M, Salim NA, Barakat M, Al-Tammemi AB. ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. Narra J. Apr 2023;3(1):e103. [doi: 10.52225/narra.v3i1.103] [Medline: 38450035]
73. Arun G, Perumal V, Urias F, et al. ChatGPT versus a customized AI chatbot (Anatbuddy) for anatomy education: a comparative pilot study. Anatomical Sciences Ed. Oct 2024;17(7). [doi: 10.1002/ase.2502] [Medline: 39169464]

74. Deng A, Chen W, Dai J, et al. Current application of ChatGPT in undergraduate nuclear medicine education: Taking Chongqing Medical University as an example. Med Teach. Jun 3, 2025;47(6):997-1003. [doi: 10.1080/0142159X.2024.2399673]
75. Garabet R, Mackey BP, Cross J, Weingarten M. ChatGPT-4 performance on USMLE step 1 style questions and its implications for medical education: a comparative study across systems and disciplines. MedSciEduc. Dec 27, 2023;34(1):145-152. [doi: 10.1007/s40670-023-01956-z]
76. Saleem N, Mufti T, Sohail SS, Madsen DØ. ChatGPT as an innovative heutagogical tool in medical education. Cogent Education. Dec 31, 2024;11(1):2332850. [doi: 10.1080/2331186X.2024.2332850]
77. Huang H, Lin HC. ChatGPT as a life coach for professional identity formation in medical education. Educational Technology & Society. 2024;27(3):374-389. URL: https://eric.ed.gov/?q=AI%2C+AND+data&ff1=souEducational+Technology+%26+Society&id=EJ1437405 [Accessed 2025-10-09]
78. Dhanvijay AKD, Pinjar MJ, Dhokane N, Sorte SR, Kumari A, Mondal H. Performance of large language models (ChatGPT, Bing Search, and Google Bard) in solving case vignettes in physiology. Cureus. Aug 2023;15(8):e42972. [doi: 10.7759/cureus.42972] [Medline: 37671207]
79. Wang T, Mainous AG 3rd, Stelter K, O'Neill TR, Newton WP. Performance evaluation of the generative pre-trained transformer (GPT-4) on the family medicine in-training examination. J Am Board Fam Med. Oct 25, 2024;37(4):528-582. [doi: 10.3122/jabfm.2023.230433R1] [Medline: 39214695]
80. Abdullahi T, Singh R, Eickhoff C. Learning to make rare and complex diagnoses with generative AI assistance: qualitative study of popular large language models. JMIR Med Educ. Feb 13, 2024;10:e51391. [doi: 10.2196/51391] [Medline: 38349725]
81. Guastafierro V, Corbitt DN, Bressan A, et al. Unveiling the risks of ChatGPT in diagnostic surgical pathology. Virchows Arch. Apr 2025;486(4):663-673. [doi: 10.1007/s00428-024-03918-1] [Medline: 39269615]
82. Sarangi PK, Irodi A, Panda S, Nayak DSK, Mondal H. Radiological differential diagnoses based on cardiovascular and thoracic imaging patterns: perspectives of four large language models. Indian J Radiol Imaging. Apr 2024;34(2):269-275. [doi: 10.1055/s-0043-1777289] [Medline: 38549881]
83. Shukla R, Mishra AK, Banerjee N, Verma A. The comparison of ChatGPT 3.5, Microsoft Bing, and Google Gemini for diagnosing cases of neuro-ophthalmology. Cureus. Apr 2024;16(4):e58232. [doi: 10.7759/cureus.58232] [Medline: 38745784]
84. Hadi A, Tran E, Nagarajan B, Kirpalani A. Evaluation of ChatGPT as a diagnostic tool for medical learners and clinicians. Ata F, editor. PLoS ONE. Jul 31, 2024;19(7):e0307383. [doi: 10.1371/journal.pone.0307383]
85. Guthrie E, Levy D, Del Carmen G. The Operating and Anesthetic Reference Assistant (OARA): A fine-tuned large language model for resident teaching. Am J Surg. Aug 2024;234:28-34. [doi: 10.1016/j.amjsurg.2024.02.016] [Medline: 38365551]
86. Zhang Y, Hao Y. Traditional Chinese medicine knowledge graph construction based on large language models. Electronics (Basel). Jul 2024;13(7):1395. [doi: 10.3390/electronics13071395]
87. Luke W, Seow Chong L, Ban KH, et al. Is ChatGPT 'ready' to be a learning tool for medical undergraduates and will it perform equally in different subjects? Comparative study of ChatGPT performance in tutorial and case-based learning questions in physiology and biochemistry. Med Teach. Nov 2024;46(11):1441-1447. [doi: 10.1080/0142159X.2024.2308779]
88. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res. May 2023;104(5):269-273. [doi: 10.4174/astr.2023.104.5.269] [Medline: 37179699]
89. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish Medical Residency entrance examination (MIR): promising horizons for AI in clinical medicine. Clin Pract. Nov 20, 2023;13(6):1460-1487. [doi: 10.3390/clinpract13060130] [Medline: 37987431]
90. Lai UH, Wu KS, Hsu TY, Kan JKC. Evaluating the performance of ChatGPT-4 on the United Kingdom Medical Licensing Assessment. Front Med. Sep 19, 2023;10:1240915. [doi: 10.3389/fmed.2023.1240915]
91. Isleem UN, Zaidat B, Ren R, et al. Can generative artificial intelligence pass the orthopaedic board examination? J Orthop. Jul 2024;53:27-33. [doi: 10.1016/j.jor.2023.10.026]
92. Mackey BP, Garabet R, Maule L, Tadesse A, Cross J, Weingarten M. Evaluating ChatGPT-4 in medical education: an assessment of subject exam performance reveals limitations in clinical curriculum support for students. Discov Artif Intell. May 16, 2024;4(1):38. [doi: 10.1007/s44163-024-00135-2]
93. Jaworski A, Jasiński D, Jaworski W, et al. Comparison of the performance of artificial intelligence versus medical professionals in the Polish Final Medical Examination. Cureus. Aug 2024;16(8):e66011. [doi: 10.7759/cureus.66011] [Medline: 39221376]

94. Abbas A, Rehman MS, Rehman SS. Comparing the performance of popular large language models on the National Board of Medical Examiners sample questions. Cureus. Mar 2024;16(3):e55991. [doi: 10.7759/cureus.55991] [Medline: 38606229]
95. Goodings AJ, Kajitani S, Chhor A, et al. Assessment of ChatGPT-4 in family medicine board examinations using advanced AI learning and analytical methods: observational study. JMIR Med Educ. Oct 8, 2024;10:e56128. [doi: 10.2196/56128] [Medline: 39378442]
96. Jin HK, Lee HE, Kim E. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. BMC Med Educ. Sep 16, 2024;24(1). [doi: 10.1186/s12909-024-05944-8]
97. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ (Chicago Ill). Nov 2024;58(11):1276-1285. [doi: 10.1111/medu.15402]
98. Alkhaaldi SMI, Kassab CH, Dimassi Z, et al. Medical student experiences and perceptions of ChatGPT and artificial intelligence: cross-sectional study. JMIR Med Educ. Dec 22, 2023;9:e51302. [doi: 10.2196/51302] [Medline: 38133911]
99. Hersh W, Fultz Hollis K. Results and implications for generative AI in a large introductory biomedical and health informatics course. NPJ Digit Med. Sep 13, 2024;7(1):247. [doi: 10.1038/s41746-024-01251-0] [Medline: 39271955]
100. Altamimi I, Alhumimidi A, Alshehri S, et al. The scientific knowledge of three large language models in cardiology: multiple-choice questions examination-based performance. Annals of Medicine & Surgery. May 3, 2024;86(6):3261-3266. [doi: 10.1097/MS9.0000000000002120]
101. Hou Y, Guo L, Luo F. Conflict of interest the authors declare that they have no conflict of interest. SSRN Journal. 2022. [doi: 10.2139/ssrn.4258054]
102. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ. Feb 14, 2024;24(1):143. [doi: 10.1186/s12909-024-05125-7] [Medline: 38355517]
103. Bongco EDA, Cua SKN, Hernandez M, Pascual JSG, Khu KJO. The performance of ChatGPT versus neurosurgery residents in neurosurgical board examination-like questions: a systematic review and meta-analysis. Neurosurg Rev. Dec 7, 2024;47(1):892. [doi: 10.1007/s10143-024-03144-y] [Medline: 39643792]
104. Cuthbert R, Simpson AI. Artificial intelligence in orthopaedics: can Chat Generative Pre-trained Transformer (ChatGPT) pass Section 1 of the Fellowship of the Royal College of Surgeons (Trauma & Orthopaedics) examination? Postgrad Med J. Sep 21, 2023;99(1176):1110-1114. [doi: 10.1093/postmj/qgad053] [Medline: 37410674]
105. Tangadulrat P, Sono S, Tangtrakulwanich B. Using ChatGPT for clinical practice and medical education: cross-sectional survey of medical students' and physicians' perceptions. JMIR Med Educ. Dec 22, 2023;9:e50658. [doi: 10.2196/50658] [Medline: 38133908]
106. Nicikowski J, Szczepański M, Miedziaszczyk M, Kudliński B. The potential of ChatGPT in medicine: an example analysis of nephrology specialty exams in Poland. Clin Kidney J. Aug 2024;17(8):sfae193. [doi: 10.1093/ckj/sfae193] [Medline: 39099569]
107. Borchert RJ, Hickman CR, Pepys J, Sadler TJ. Performance of ChatGPT on the situational judgement test-a professional dilemmas-based examination for doctors in the United Kingdom. JMIR Med Educ. Aug 7, 2023;9:e48978. [doi: 10.2196/48978] [Medline: 37548997]
108. Hudon A, Kiepura B, Pelletier M, Phan V. Using ChatGPT in psychiatry to design script concordance tests in undergraduate medical education: mixed methods study. JMIR Med Educ. Apr 4, 2024;10:e54067. [doi: 10.2196/54067] [Medline: 38596832]
109. Agarwal M, Sharma P, Goswami A. Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. Cureus. Jun 2023;15(6):e40977. [doi: 10.7759/cureus.40977] [Medline: 37519497]
110. Wu JH, Nishida T, Liu TYA. Accuracy of large language models in answering ophthalmology board-style questions: A meta-analysis. Asia Pac J Ophthalmol (Phila). Sep 2024;13(5):100106. [doi: 10.1016/j.apjo.2024.100106]
111. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, et al. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical Examination: a cross-sectional study. J Educ Eval Health Prof. 2023;20:30. [doi: 10.3352/jeehp.2023.20.30] [Medline: 37981579]
112. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc. Aug 1, 2023;86(8):762-766. [doi: 10.1097/JCMA.0000000000000946] [Medline: 37294147]
113. Yoon SH, Oh SK, Lim BG, Lee HJ. Performance of ChatGPT in the in-training examination for anesthesiology and pain medicine residents in South Korea: observational study. JMIR Med Educ. Sep 16, 2024;10:e56859. [doi: 10.2196/56859] [Medline: 39284182]
114. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: Pave the way for medical AI. Int J Med Inform. Sep 2023;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173]

115. Keshtkar A, Atighi F, Reihani H. Systematic review of ChatGPT accuracy and performance in Iran's medical licensing exams: A brief report. J Educ Health Promot. Nov 2024;13(1):421. [doi: 10.4103/jehp.jehp_1210_24]
116. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res. Jul 25, 2024;26:e60807. [doi: 10.2196/60807] [Medline: 39052324]
117. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery. Dec 1, 2023;93(6):1353-1365. [doi: 10.1227/neu.0000000000002632] [Medline: 37581444]
118. Elias ML, Burshtein J, Sharon VR. OpenAI's GPT-4 performs to a high degree on board-style dermatology questions. Int J Dermatology. Jan 2024;63(1):73-78. [doi: 10.1111/ijd.16913]
119. Sabri H, Saleh MHA, Hazrati P, et al. Performance of three artificial intelligence (AI)-based large language models in standardized testing; implications for AI-assisted dental education. J of Periodontal Research. Feb 2025;60(2):121-133. [doi: 10.1111/jre.13323]
120. Ilgaz HB, Çelik Z. The significance of artificial intelligence platforms in anatomy education: an experience with ChatGPT and Google Bard. Cureus. Sep 2023;15(9):e45301. [doi: 10.7759/cureus.45301] [Medline: 37846274]
121. Khorshidi H, Mohammadi A, Yousem DM, et al. Application of ChatGPT in multilingual medical education: How does ChatGPT fare in 2023's Iranian residency entrance examination. Informatics in Medicine Unlocked. 2023;41:101314. [doi: 10.1016/j.imu.2023.101314]
122. Huang CH, Hsiao HJ, Yeh PC, Wu KC, Kao CH. Performance of ChatGPT on Stage 1 of the Taiwanese medical licensing exam. Digit HEALTH. 2024;10:20552076241233144. [doi: 10.1177/20552076241233144] [Medline: 38371244]
123. Apornvirat S, Namboonlue C, Laohawetwanit T. Comparative analysis of ChatGPT and Bard in answering pathology examination questions requiring image interpretation. Am J Clin Pathol. Sep 3, 2024;162(3):252-260. [doi: 10.1093/ajcp/aqae036]
124. Cross J, Robinson R, Devaraju S, et al. Transforming medical education: assessing the integration of ChatGPT into faculty workflows at a Caribbean medical school. Cureus. Jul 2023;15(7):e41399. [doi: 10.7759/cureus.41399] [Medline: 37426402]
125. Soulage CO, Van Coppenolle F, Guebre-Egziabher F. The conversational AI "ChatGPT" outperforms medical students on a physiology university examination. Adv Physiol Educ. Dec 1, 2024;48(4):677-684. [doi: 10.1152/advan.00181.2023] [Medline: 38991037]
126. Gritti MN, AlTurki H, Farid P, Morgan CT. Progression of an artificial intelligence chatbot (ChatGPT) for pediatric cardiology educational knowledge assessment. Pediatr Cardiol. Feb 2024;45(2):309-313. [doi: 10.1007/s00246-023-03385-6] [Medline: 38170274]
127. Bartoli A, May AT, Al-Awadhi A, Schaller K. Probing artificial intelligence in neurosurgical training: ChatGPT takes a neurosurgical residents written exam. Brain Spine. 2024;4:102715. [doi: 10.1016/j.bas.2023.102715] [Medline: 38163001]
128. Rasmussen ME, Akbarov K, Titovich E, et al. Potential of e-learning interventions and artificial intelligence-assisted contouring skills in radiotherapy: the ELAISA study. JCO Glob Oncol. Aug 2024;10(10):e2400173. [doi: 10.1200/GO.24.00173] [Medline: 39236283]
129. Mousavi M, Shafiee S, Harley JM, Cheung JCK, Abbasgholizadeh Rahimi S. Performance of generative pre-trained transformers (GPTs) in Certification Examination of the College of Family Physicians of Canada. Fam Med Com Health. May 2024;12(Suppl 1):e002626. [doi: 10.1136/fmch-2023-002626]
130. Temsah MH, Alhuzaimi AN, Almansour M, et al. Art or artifact: evaluating the accuracy, appeal, and educational value of AI-generated imagery in DALL·E 3 for illustrating congenital heart diseases. J Med Syst. May 23, 2024;48(1):54. [doi: 10.1007/s10916-024-02072-0] [Medline: 38780839]
131. Fang Q, Reynaldi R, Araminta AS, et al. Artificial intelligence (AI)-driven dental education: exploring the role of chatbots in a clinical learning environment. J Prosthet Dent. Oct 2025;134(4):1296-1303. [doi: 10.1016/j.prosdent.2024.03.038] [Medline: 38644064]
132. Cheung BHH, Lau GKK, Wong GTC, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions-A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS ONE. 2023;18(8):e0290691. [doi: 10.1371/journal.pone.0290691] [Medline: 37643186]
133. Ignjatović A, Stevanović L. Efficacy and limitations of ChatGPT as a biostatistical problem-solving tool in medical education in Serbia: a descriptive study. J Educ Eval Health Prof. Oct 16, 2023;20(28):28. [doi: 10.3352/jeehp.2023.20.28]
134. Agarwal M, Goswami A, Sharma P. Evaluating ChatGPT-3.5 and Claude-2 in answering and explaining conceptual medical physiology multiple-choice questions. Cureus. Sep 2023;15(9):e46222. [doi: 10.7759/cureus.46222] [Medline: 37908959]
135. Yanagita Y, Yokokawa D, Fukuzawa F, Uchida S, Uehara T, Ikusaka M. Expert assessment of ChatGPT's ability to generate illness scripts: an evaluative study. BMC Med Educ. May 15, 2024;24(1):536. [doi: 10.1186/s12909-024-05534-8] [Medline: 38750546]

136. Sauder M, Tritsch T, Rajput V, Schwartz G, Shoja MM. Exploring generative artificial intelligence-assisted medical education: assessing case-based learning for medical students. Cureus. Jan 2024;16(1):e51961. [doi: 10.7759/cureus.51961] [Medline: 38333501]
137. Hanna RE, Smith LR, Mhaskar R, Hanna K. Performance of language models on the family medicine in-training exam. Fam Med. Oct 2024;56(9):555-560. [doi: 10.22454/FamMed.2024.233738] [Medline: 39207788]
138. Takahashi H, Shikino K, Kondo T, et al. Educational utility of clinical vignettes generated in Japanese by ChatGPT-4: mixed methods study. JMIR Med Educ. Aug 13, 2024;10:e59133. [doi: 10.2196/59133] [Medline: 39137031]
139. Waikel RL, Othman AA, Patel T, et al. Recognition of genetic conditions after learning with images created using generative artificial intelligence. JAMA Netw Open. Mar 4, 2024;7(3):e242609. [doi: 10.1001/jamanetworkopen.2024.2609] [Medline: 38488790]
140. Collins BR, Black EW, Rarey KE. Introducing AnatomyGPT: A customized artificial intelligence application for anatomical sciences education. Clin Anat. Sep 2024;37(6):661-669. [doi: 10.1002/ca.24178] [Medline: 38721869]
141. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. Dagan A, editor. PLOS Digit Health. Feb 2023;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
142. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep. Nov 22, 2023;13(1). [doi: 10.1038/s41598-023-46995-z]
143. Murphy Lonergan R, Curry J, Dhas K, Simmons BI. Stratified evaluation of GPT's question answering in surgery reveals artificial intelligence (AI) knowledge gaps. Cureus. Nov 2023;15(11):e48788. [doi: 10.7759/cureus.48788] [Medline: 38098921]
144. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Reshaping medical education: performance of ChatGPT on a PES medical examination. Cardiol J. 2024;31(3):442-450. [doi: 10.5603/cj.97517] [Medline: 37830257]
145. Coşkun Ö, Kıyak YS, Budakoğlu Iİ. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: A randomized controlled experiment. Med Teach. Feb 2025;47(2):268-274. [doi: 10.1080/0142159X.2024.2327477] [Medline: 38478902]
146. Knoedler L, Alfertshofer M, Knoedler S, et al. Pure wisdom or potemkin villages? A comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE step 3 style questions: quantitative analysis. JMIR Med Educ. Jan 5, 2024;10:e51148. [doi: 10.2196/51148] [Medline: 38180782]
147. Uribe SE, Maldupa I, Kavadella A, et al. Artificial intelligence chatbots and large language models in dental education: worldwide survey of educators. Eur J Dent Educ. Nov 2024;28(4):865-876. [doi: 10.1111/eje.13009] [Medline: 38586899]
148. Jarry Trujillo C, Vela Ulloa J, Escalona Vivas G, et al. Surgeons vs ChatGPT: assessment and feedback performance based on real surgical scenarios. J Surg Educ. Jul 2024;81(7):960-966. [doi: 10.1016/j.jsurg.2024.03.012] [Medline: 38749814]
149. Meo SA, Al-Khlaiwi T, AbuKhalaf AA, Meo AS, Klonoff DC. The scientific knowledge of Bard and ChatGPT in endocrinology, diabetes, and diabetes technology: multiple-choice questions examination-based performance. J Diabetes Sci Technol. May 2025;19(3):705-710. [doi: 10.1177/19322968231203987] [Medline: 37798960]
150. Shamim MS, Zaidi SJA, Rehman A. The revival of essay-type questions in medical education: harnessing artificial intelligence and machine learning. J Coll Physicians Surg Pak. May 2024;34(5):595-599. [doi: 10.29271/jcpsp.2024.05.595] [Medline: 38720222]
151. Meo SA, Alotaibi M, Meo MZS, Meo MOS, Hamid M. Medical knowledge of ChatGPT in public health, infectious diseases, COVID-19 pandemic, and vaccines: multiple choice questions examination based performance. Front Public Health. 2024;12:1360597. [doi: 10.3389/fpubh.2024.1360597] [Medline: 38711764]
152. Ba H, zhang L, Yi Z. Enhancing clinical skills in pediatric trainees: a comparative study of ChatGPT-assisted and traditional teaching methods. BMC Med Educ. May 22, 2024;24(1). [doi: 10.1186/s12909-024-05565-1]
153. Almazrou S, Alanezi F, Almutairi SA, et al. Enhancing medical students critical thinking skills through ChatGPT: An empirical study with medical students. Nutr Health. Jul 2025;31(3):1023-1033. [doi: 10.1177/02601060241273627]
154. Crawford LM, Hendzlik P, Lam J, et al. Digital ink and surgical dreams: perceptions of artificial intelligence–generated essays in residency applications. J Surg Res. Sep 2024;301:504-511. [doi: 10.1016/j.jss.2024.06.020] [Medline: 39042979]
155. Mosleh R, Jarrar Q, Jarrar Y, Tazkarji M, Hawash M. Medicine and pharmacy students' knowledge, attitudes, and practice regarding artificial intelligence programs: Jordan and West Bank of Palestine. Adv Med Educ Pract. 2023;14:1391-1400. [doi: 10.2147/AMEP.S433255] [Medline: 38106923]
156. Western MJ, Smit ES, Gültzow T, et al. Bridging the digital health divide: a narrative review of the causes, implications, and solutions for digital health inequalities. Health Psychol Behav Med. 2025;13(1):2493139. [doi: 10.1080/21642850.2025.2493139] [Medline: 40276490]
157. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. May 4, 2023;6. [doi: 10.3389/frai.2023.1169595]

158. Liu L, Qu S, Zhao H, et al. Global trends and hotspots of ChatGPT in medical research: a bibliometric and visualized study. Front Med. May 16, 2024;11. [doi: 10.3389/fmed.2024.1406842]
159. Khan N, Khan Z, Koubaa A, Khan MK, Salleh R bin. Global insights and the impact of generative AI-ChatGPT on multidisciplinary: a systematic review and bibliometric analysis. Conn Sci. Dec 31, 2024;36(1). [doi: 10.1080/09540091.2024.2353630]
160. 100+ eye-opening ChatGPT statistics: tracing the roots of generative AI to its global dominance. Master of Code. Jan 2025. URL: https://masterofcode.com/blog/chatgpt-statistics [Accessed 2025-07-26]
161. See BH, Gorard S, Lu B, Dong L, Siddiqui N. Is technology always helpful?: A critical review of the impact on learning outcomes of education technology in supporting formative assessment in schools. Res Pap Educ. Nov 2, 2022;37(6):1064-1096. [doi: 10.1080/02671522.2021.1907778]
162. Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. Informatics (MDPI). ;11(3):57. [doi: 10.3390/informatics11030057]
163. Busch F, Hoffmann L, Rueger C, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med. Jan 21, 2025;5(1). [doi: 10.1038/s43856-024-00717-2]
164. Meyer JG, Urbanowicz RJ, Martin PCN, et al. ChatGPT and large language models in academia: opportunities and challenges. BioData Min. Jul 13, 2023;16(1). [doi: 10.1186/s13040-023-00339-9]
165. Mao J, Chen B, Liu JC. Generative artificial intelligence in education and its implications for assessment. TechTrends. Jan 2024;68(1):58-66. [doi: 10.1007/s11528-023-00911-4]
166. Turner L, Hashimoto DA, Vasisht S, Schaye V. Demystifying AI: current state and future role in medical education assessment. Acad Med. Apr 1, 2024;99(4S Suppl 1):S42-S47. [doi: 10.1097/ACM.0000000000005598] [Medline: 38166201]
167. Lakhtakia R, Otaki F, Alsuwaidi L, Zary N. Assessment as learning in medical education: feasibility and perceived impact of student-generated formative assessments. JMIR Med Educ. Jul 22, 2022;8(3):e35820. [doi: 10.2196/35820] [Medline: 35867379]
168. Machkour M, El Jihaoui M, Lamalif L, Faris S, Mansouri K. Toward an adaptive learning assessment pathway. Front Educ. 2025;10. [doi: 10.3389/feduc.2025.1498233]
169. Solis Trujillo BP, Velarde-Camaqui D, Gonzales Nuñez CA, Castillo Silva EV, Gonzalez Said de la Oliva M del P. The current landscape of formative assessment and feedback in graduate studies: a systematic literature review. Front Educ. May 12, 2025;10. [doi: 10.3389/feduc.2025.1509983]
170. Wilson C, Scott B. Adaptive systems in education: a review and conceptual unification. IJILT. Jan 3, 2017;34(1):2-19. [doi: 10.1108/IJILT-09-2016-0040]
171. Kolluru V, Mungara S, Chintakunta AN. Adaptive learning systems: harnessing AI for customized educational experiences. IJCSITY. Aug 30, 2018;6(3):13-26. URL: https://airccse.org/journal/ijcsity/Current2018.html [Accessed 2025-10-09] [doi: 10.5121/ijcsity.2018.6302]
172. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: Implications for clinical decision-making. PLOS Digit Health. Nov 2024;3(11):e0000651. [doi: 10.1371/journal.pdig.0000651]
173. Sawan M. Balancing automation and empathy: how teachers can thrive with AI. Zenodo. Preprint posted online on May 18, 2025. [doi: 10.5281/zenodo.15456225]
174. Bond M, Khosravi H, De Laat M, et al. A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. Int J Educ Technol High Educ. Jan 19, 2024;21(1). [doi: 10.1186/s41239-023-00436-z]
175. Resnik DB, Hosseini M. The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool. AI Ethics. Apr 2025;5(2):1499-1521. [doi: 10.1007/s43681-024-00493-8]
176. Tong D, Jin B, Tao Y, Ren H, Atiquil Islam AYM, Bao L. Exploring the role of human-AI collaboration in solving scientific problems. Phys Rev Phys Educ Res. May 2025;21(1):010149. [doi: 10.1103/PhysRevPhysEducRes.21.010149]
177. Yu S, Lee SS, Hwang H. The ethics of using artificial intelligence in medical research. KMJ. Dec 2024;39(4):229-237. [doi: 10.7180/kmj.24.140]
178. Web-Based Medical Teaching Using a Multi-Agent System Applications and Innovations in Intelligent Systems XIII. Springer London; 181-194. [doi: 10.1007/1-84628-224-1_14] ISBN: 978-1-84628-223-2
179. Wei H, Qiu J, Yu H, Yuan W. MEDCO: medical education copilots based on a multi-agent framework. arXiv. Preprint posted online on Aug 22, 2024. [doi: 10.48550/ARXIV.2408.12496]
180. Liu F, Zhou H, Gu B, et al. Application of large language models in medicine. Nat Rev Bioeng. 2025;3(6):445-464. URL: https://www.nature.com/articles/s44222-025-00279-5 [Accessed 2025-07-12] [doi: 10.1038/s44222-025-00279-5]
181. Zhang K, Meng X, Yan X, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res. Jan 7, 2025;27:e59069. [doi: 10.2196/59069] [Medline: 39773666]
182. Hasanzadeh F, Josephson CB, Waters G, Adedinsewo D, Azizi Z, White JA. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. NPJ Digit Med. Mar 11, 2025;8(1):154. [doi: 10.1038/s41746-025-01503-7] [Medline: 40069303]

183. Li H, Li C, Wang J, et al. Review on security of federated learning and its application in healthcare. Future Generation Computer Systems. Jul 2023;144:271-290. [doi: 10.1016/j.future.2023.02.021]
184. Hu F, Qiu S, Yang X, Wu C, Nunes MB, Chen H. Privacy-preserving healthcare and medical data collaboration service system based on blockchain and federated learning. CMC. 2024;80(2):2897-2915. [doi: 10.32604/cmc.2024.052570]
185. Ozer M. The Matthew Effect in Turkish Education System. BUJFED. Nov 13, 2024. [doi: 10.14686/buefad.1359312]
186. Lucchi N. ChatGPT: a case study on copyright challenges for generative artificial intelligence systems. Eur j risk regul. Sep 2024;15(3):602-624. [doi: 10.1017/err.2023.59]
187. Mitra A, Mawson A. Neglected tropical diseases: epidemiology and global burden. TropicalMed. Aug 5, 2017;2(3):36. [doi: 10.3390/tropicalmed2030036]
188. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ. Dec 3, 2019;5(2):e16048. [doi: 10.2196/16048] [Medline: 31793895]
189. Talib ZM, Kiguli-Malwadde E, Wohltjen H, et al. Transforming health professions' education through in-country collaboration: examining the consortia among African medical schools catalyzed by the Medical Education Partnership Initiative. Hum Resour Health. Dec 2015;13(1). [doi: 10.1186/1478-4491-13-1]
190. Ueda D, Kakinuma T, Fujita S, et al. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol. Jan 2024;42(1):3-15. [doi: 10.1007/s11604-023-01474-3]
191. Bui TTU, Tong TVA. The impact of AI writing tools on academic integrity: unveiling English-majored students' perceptions and practical solutions. acoj. Jan 27, 2025;16(1):83-110. URL: http://asiacall-acoj.org/index.php/journal/issue/view/7 [Accessed 2025-10-09] [doi: 10.54855/acoj.251615]
192. Yoo JH. Defining the boundaries of AI use in scientific writing: a comparative review of editorial policies. J Korean Med Sci. Jun 16, 2025;40(23):e187. [doi: 10.3346/jkms.2025.40.e187] [Medline: 40524628]
193. Schwartzstein RM. Clinical reasoning and artificial intelligence: Can AI really think. Trans Am Clin Climatol Assoc. 2024;134:133-145. [Medline: 39135584]
194. Kim Y, Jeong H, Chen S, et al. Medical hallucinations in foundation models and their impact on healthcare. arXiv. Preprint posted online on Feb 26, 2025. [doi: 10.48550/arXiv.2503.05777]
195. Alkhanbouli R, Matar Abdulla Almadhaani H, Alhosani F, Simsekler MCE. The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions. BMC Med Inform Decis Mak. 2025;25(1). [doi: 10.1186/s12911-025-02944-6]
196. Cohen IG, Babic B, Gerke S, Xia Q, Evgeniou T, Wertenbroch K. How AI can learn from the law: putting humans in the loop only on appeal. npj Digit Med. Aug 25, 2023;6(1). [doi: 10.1038/s41746-023-00906-8]

Abbreviations

AI: artificial intelligence
GAI: generative artificial intelligence
HDI: Human Development Index
LLM: large language model
MCQ: multiple-choice question
RMA: resource-method-assessment
SAQ: short-answer question

Edited by Blake Lesselroth; peer-reviewed by Bertalan Meskó, Changyu Wang, Ching Nam Hang, Lingxuan Zhu, Rong Yin; submitted 10.Jan.2025; final revised version received 26.Jul.2025; accepted 23.Sep.2025; published 23.Oct.2025

Please cite as:
Lin Y, Luo Z, Ye Z, Zhong N, Zhao L, Zhang L, Li X, Chen Z, Chen Y
Applications, Challenges, and Prospects of Generative Artificial Intelligence Empowering Medical Education: Scoping Review
JMIR Med Educ2025;11:e71125
URL: https://mededu.jmir.org/2025/1/e71125
doi: 10.2196/71125