Advances in Medical Education and Practice

Dovepress
Taylor & Francis Group

Open Access Full Text Article

REVIEW

Delving into the Practical Applications and Pitfalls of Large Language Models in Medical Education: Narrative Review

Rui Li[1], Tong Wu[2–4]

[1]Emergency Department, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, People's Republic of China; [2]National Clinical Research Center for Obstetrical and Gynecological Diseases, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, People's Republic of China; [3]Key Laboratory of Cancer Invasion and Metastasis, Ministry of Education, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, People's Republic of China; [4]Department of Obstetrics and Gynecology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, People's Republic of China

Correspondence: Tong Wu, Department of Obstetrics and Gynecology, Tongji Hospital, No. 1095, Jiefang Avenue, Wuhan, 430030, People's Republic of China, Email tongwu66@tjh.tjmu.edu.cn

Abstract: Large language models (LLMs) have emerged as valuable tools in medical education, attracting substantial attention in recent years. They offer educators essential support in developing instructional plans, generating interactive materials, and facilitating efficient feedback mechanisms. Furthermore, LLMs enhance students' language acquisition, writing proficiency, and creativity in educational activities. This review aims to examine the practical applications of LLMs in enhancing the educational and academic performance of both teachers and students, providing specific examples to demonstrate their effectiveness. Additionally, we address the inherent challenges associated with LLM implementation and propose viable solutions to optimize their use. Our study lays the groundwork for the broader integration of LLMs in medical education and research, ensuring the highest standards of medical learning and, ultimately, patient safety.

Keywords: large language models, medical education, artificial intelligence, educator, automation bias, hallucination

Advances in Medical Education and Practice downloaded from https://www.dovepress.com/
For personal use only.

Background

Large language models (LLMs), known as a cutting-edge artificial intelligence (AI) technology, have garnered significant attention since the release of the Chat Generative Pre-trained Transformer (ChatGPT) in November 2022.[1] In contrast to traditional tools like AI generators, which typically rely on predefined algorithms and datasets, LLMs leverage vast datasets and advanced neural network architectures to perform a wide range of tasks. These models possess a remarkable ability to understand textual subtleties and generate more nuanced and contextually relevant outputs in real-time conversations, blurring the distinction between human and machine generated content.[2]

LLMs are increasingly recognized as essential instruments in medical education to equip future physicians and healthcare professionals with adequate training.[3] With the help of LLMs, teachers may develop multiple-choice questions that more accurately reflect real-world clinical scenarios,[4] and educational institutions can make data-driven decisions regarding curriculum development, student assessment, and resource allocation. For learners, LLMs offer the capability to generate content tailored to their individual interests, cognitive abilities, and learning styles, thereby enhancing the creation of adaptive learning environments.[5] Collectively, LLMs have made substantial strides in transforming medical education practices.

Despite the existing literature highlighting the potential of LLMs to transform educational paradigms, it often falls short of providing concrete, real-world examples of implementation within medical curricula.[6–9] It may be attributed to the fact that the audience for this topic is typically broad, encompassing medical educators, researchers, students, and policymakers, among others. Due to the significant differences in the specific needs and focuses, providing concrete

Received: 19 September 2024
Accepted: 10 April 2025
Published: 18 April 2025

© 2025 Li and Wu. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at https://www.dovepress.com/terms.php and incorporate the Creative Commons Attribution – Non Commercial (unported, v4.0) License (http://creativecommons.org/licenses/by-nc/4.0/). By accessing the work you hereby accept the Terms. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5 of our Terms (https://www.dovepress.com/terms.php).

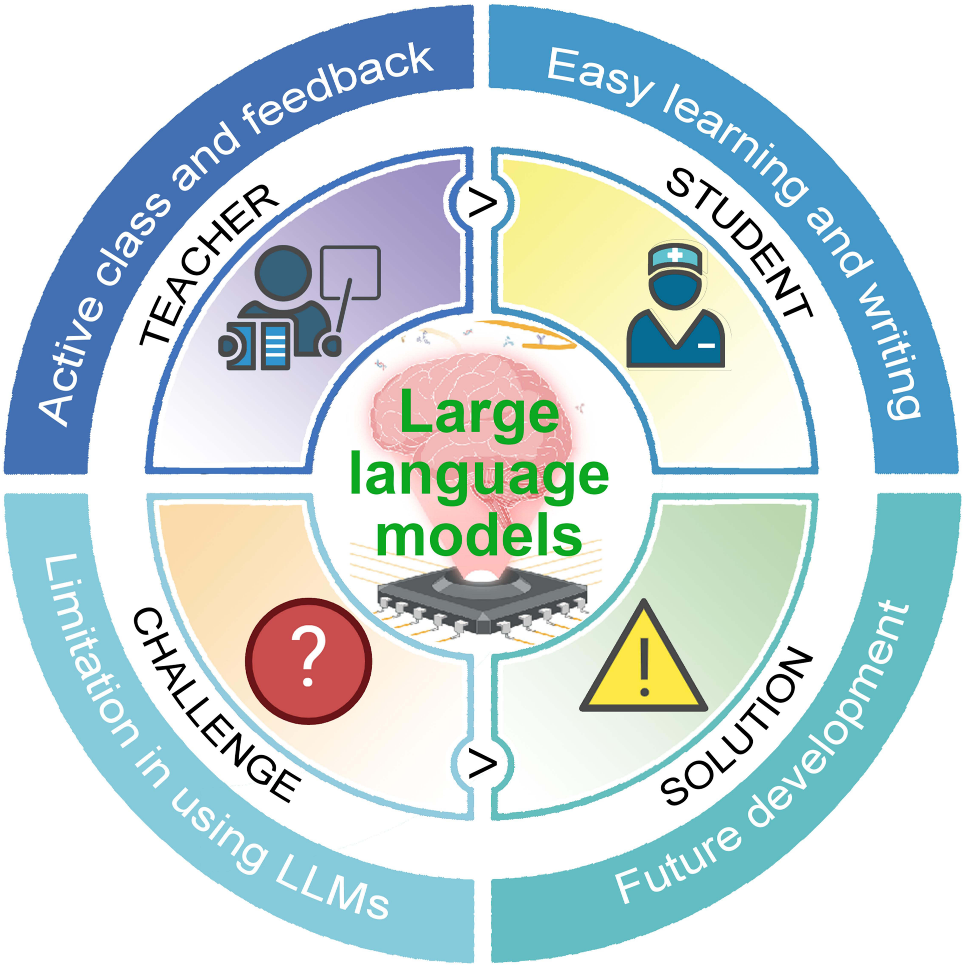


Figure 1 The landscape of large language models (LLMs) in medical education domain. Both of teachers and students can benefits this AI tools, however, some challenges need to be solved.

examples may not meet the demands of all readers and could potentially compromise the universality and applicability. Furthermore, much of the current discourse relies on grey literature, including websites, blogs, and news outlets, which may propagate misinformation and lack the rigor of peer-reviewed research.[10] The challenges associated with LLMs necessitate careful consideration as well. Given these gaps, there is a pressing need, particularly for most medical educators, to explore and evaluate the implementation of LLMs in real-world contexts and to generate robust, peer-reviewed evidence.

In this study, we endeavor to offer concrete examples of LLMs bridge the gap between theoretical frameworks and practical applications, with a primary focus on insights derived from peer-reviewed literature. We begin by discussing the application scenarios of LLMs before, during, and after class. Following this, we delineate the challenges encountered in employing LLMs and present corresponding possible solutions (Figure 1). This research contributes to advancing the application of artificial intelligence language models within the context of medical education.

Method

Given the primary objective of this review is integrative rather than summative, we determined that a narrative review would be more appropriate than a systematic review. Although systematic reviews are often preferred for their reproducibility, they typically have a narrower scope compared to narrative reviews.[11] Narrative reviews are particularly advantageous when the aim is to provide a comprehensive overview of a topic. This approach permits a more flexible examination of the literature, accommodating diverse methodologies and study designs.[11] While systematic reviews adhere to stringent reporting standards, the standards for narrative reviews are less well established. In formulating our methodological approach, we referred to the Scale for the Assessment of Narrative Review Articles to ensure the quality assessment of narrative review articles.[12]

Articles were identified via the Biomed Central, Web of Science, and Medline databases, using the key words "large language model", "ChatGPT", "medical education" and "surgical training" separately and in various combinations. In addition, we examined the references cited and citing articles. Studies were included if they: (1) utilized empirical data (eg, randomized controlled trials, cohort studies, case reports), (2) were peer-reviewed, and (3) were published from 2010 onwards. Studies were excluded if they: (1) were non-empirical works (eg, editorials, opinion pieces, theoretical frameworks), (2) were not related to the intersection of medical education and LLMs, (3) reported duplicate records, or (4) were written in a language other than English. This approach allowed us to identify and include studies that provided practical applications of LLMs. Both authors were actively involved in the literature search process.

Transformation of Teaching Strategy

Medical educators are encouraged to creatively incorporate LLMs with routine curriculum (Figure 2). Before class, LLMs can assist in formulating teaching plans tailored to diverse learning styles of students. For visual learners, LLMs can transform radiologic images into 3D models within virtual reality interfaces, which allows students to explore anatomical structures in a virtual environment during anatomy classes.[13] The integration of LLMs enhances virtual simulations by making them more dynamic and offering real-time feedback and personalized guidance as students navigate through virtual scenarios.[14] Auditory learners can benefit from lectures, discussions, and podcasts generated by LLMs. Research on pharmacology podcasts indicates that audio resources improve comprehension and empower students to revisit lectures, prepare for examinations, and clarify complex topics.[15] Similarly, in the field of anesthesiology and intensive care, podcasts show significant positive impacts on knowledge acquisition and clinical skills development.[16] For kinesthetic learners, LLMs in virtual surgery training offer detailed, context-aware instructions at each procedural step, and even interpret and respond to user inputs.[5] This interaction allows learners to ask questions and

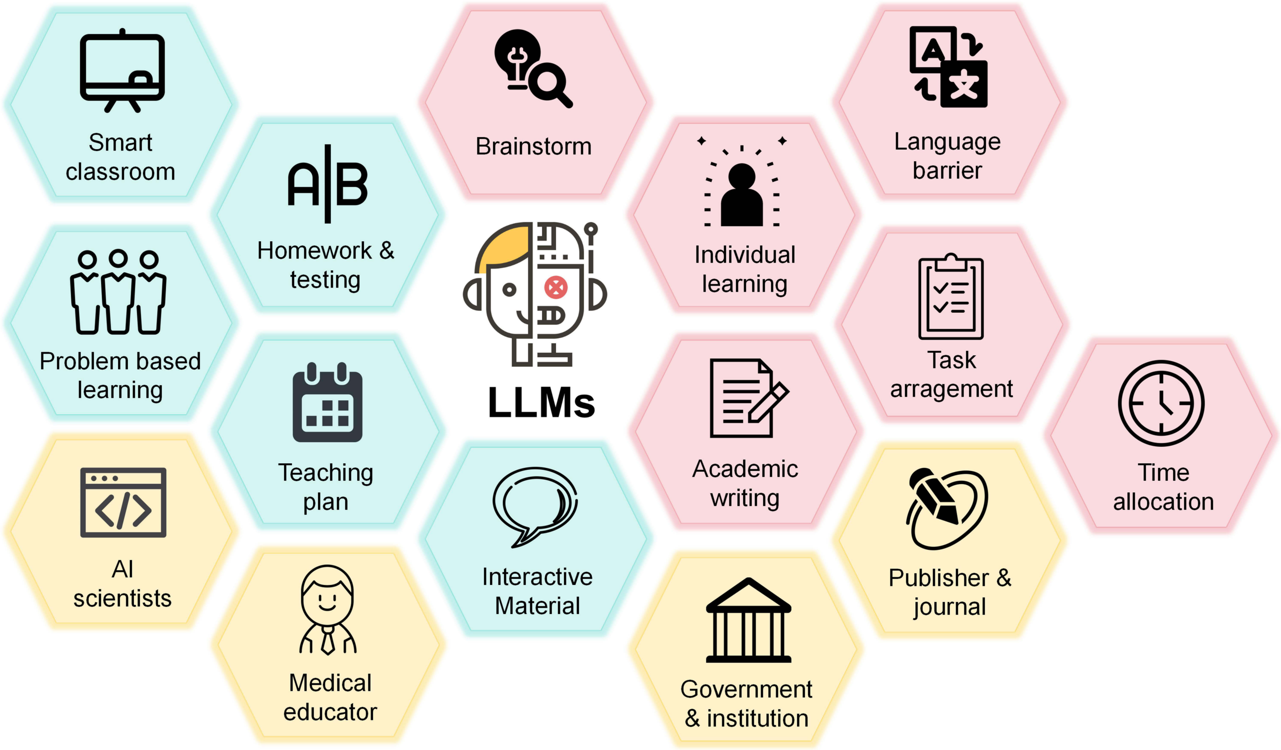
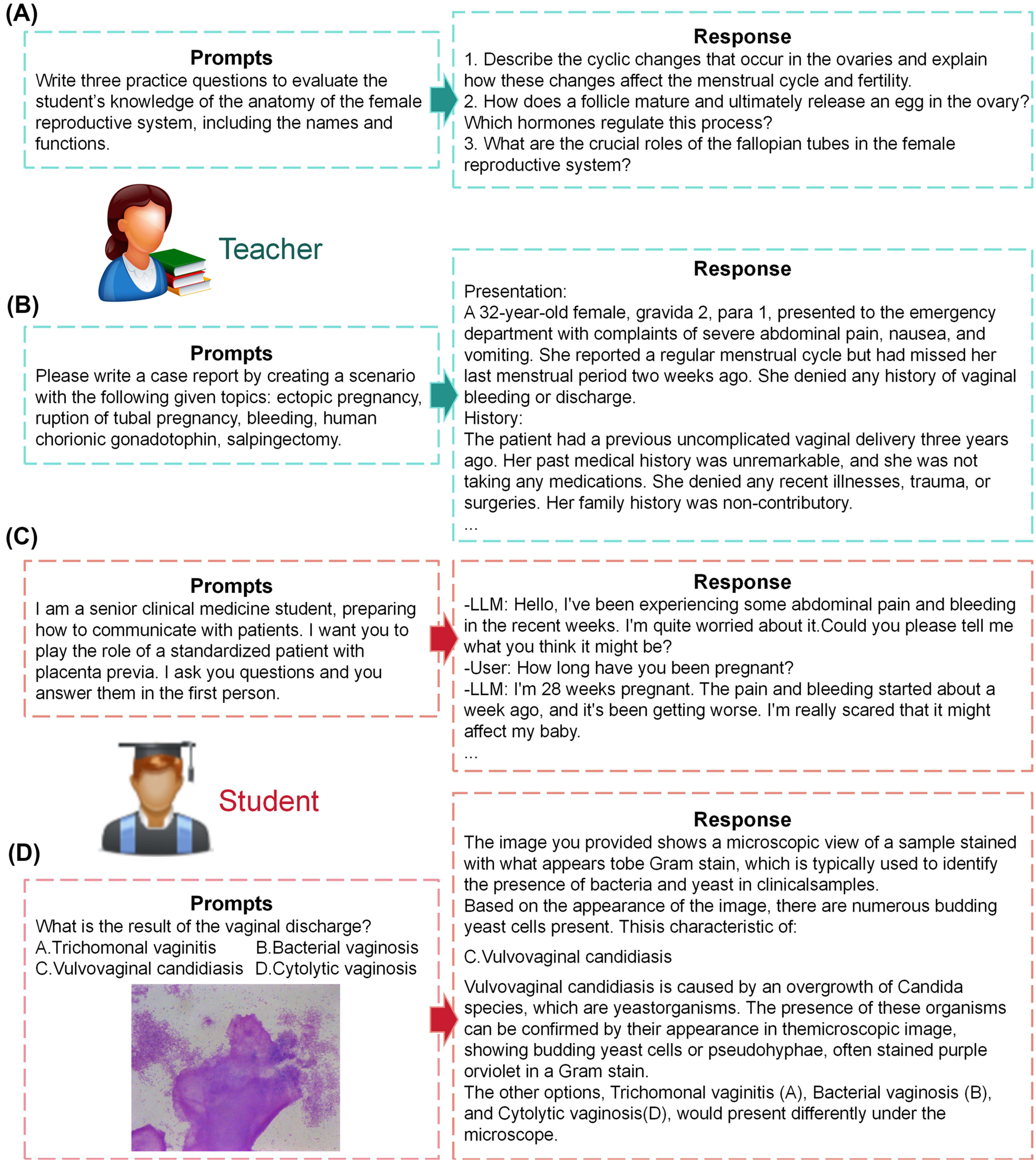


Figure 2 Application areas for teachers (green) and students (red) are shown. Key participants that are relevant to the development of LLMs powered medical education include medical educator, AI scientists, school regulations, and publishers.

https://doi.org/10.2147/AMEP.S497020

receive immediate, customized feedback, which is essential for reinforcing learning and ensuring a comprehensive understanding of complex surgical techniques.[17] Such ChatGPT-assisted surgical training was implemented in a two-week training program, where significant improvement in practical clinical skills was observed.[18]

During class, LLMs allows for the rapid generation of interactive educational materials, such as questions, task cards, and clinical scenarios (Figure 3A–C). Case-based learning scenarios designed by LLMs may represent diverse patient demographics, thereby ensuring that students encounter a broad spectrum of clinical situations.[19] The creation of culturally sensitive



Prompts
Write three practice questions to evaluate the student's knowledge of the anatomy of the female reproductive system, including the names and functions.

Response
1. Describe the cyclic changes that occur in the ovaries and explain how these changes affect the menstrual cycle and fertility.
2. How does a follicle mature and ultimately release an egg in the ovary? Which hormones regulate this process?
3. What are the crucial roles of the fallopian tubes in the female reproductive system?

Prompts
Please write a case report by creating a scenario with the following given topics: ectopic pregnancy, ruption of tubal pregnancy, bleeding, human chorionic gonadotophin, salpingectomy.

Response
Presentation:
A 32-year-old female, gravida 2, para 1, presented to the emergency department with complaints of severe abdominal pain, nausea, and vomiting. She reported a regular menstrual cycle but had missed her last menstrual period two weeks ago. She denied any history of vaginal bleeding or discharge.
History:
The patient had a previous uncomplicated vaginal delivery three years ago. Her past medical history was unremarkable, and she was not taking any medications. She denied any recent illnesses, trauma, or surgeries. Her family history was non-contributory.
...

Prompts
I am a senior clinical medicine student, preparing how to communicate with patients. I want you to play the role of a standardized patient with placenta previa. I ask you questions and you answer them in the first person.

Response
-LLM: Hello, I've been experiencing some abdominal pain and bleeding in the recent weeks. I'm quite worried about it.Could you please tell me what you think it might be?
-User: How long have you been pregnant?
-LLM: I'm 28 weeks pregnant. The pain and bleeding started about a week ago, and it's been getting worse. I'm really scared that it might affect my baby.
...

Prompts
What is the result of the vaginal discharge?
A.Trichomonal vaginitis B.Bacterial vaginosis
C.Vulvovaginal candidiasis D.Cytolytic vaginosis

Response
The image you provided shows a microscopic view of a sample stained with what appears tobe Gram stain, which is typically used to identify the presence of bacteria and yeast in clinicalsamples.
Based on the appearance of the image, there are numerous budding yeast cells present. Thisis characteristic of:

C.Vulvovaginal candidiasis

Vulvovaginal candidiasis is caused by an overgrowth of Candida species, which are yeastorganisms. The presence of these organisms can be confirmed by their appearance in themicroscopic image, showing budding yeast cells or pseudohyphae, often stained purple orviolet in a Gram stain.
The other options, Trichomonal vaginitis (A), Bacterial vaginosis (B), and Cytolytic vaginosis(D), would present differently under the microscope.

Figure 3 Example prompts to ask LLM questions and its response about (**A**) teaching questions, (**B**) writing clinical cases, (**C**) simulating conversation, and (**D**) testing.

educational cases is essential for training healthcare professionals who can cater to the needs of diverse patient populations.[19] In our examples, LLMs could generate self-check quizzes and clinical cases with detailed answer explanations. This aligns with the finding that the quality and readability of clinical contents produced by ChatGPT are comparable to those authored by human experts.[20] Zeng et al utilized open-ended questions generated by ChatGPT for a cohort of urology interns within problem-based learning sessions. They found that students assisted by ChatGPT outperformed the traditional group in exams, and had significant gains in medical interviewing, clinical judgment, and overall clinical competence.[21] Moreover, LLMs can effectively create realistic and clinically relevant scenarios that facilitate the application of theoretical knowledge to practical situations.[22] A recent study demonstrated that engaging students in the diagnosis and treatment of "virtual patients" significantly enhanced compliance and satisfaction.[23] In diagnostic course, LLMs can serve as powerful tools for generating differential diagnoses and supporting clinical decision-making processes. GPT-4, in particular, can produce differential diagnosis lists with a high degree of accuracy, thereby exposing medical students and trainees to a wider array of potential diagnoses.[24]

After class, LLMs function as a valuable supplementary resource to construct medical questions for homework assignments and assessments. For instance, LLMs can be used to generate questions that align with Entrustable Professional Activities, which are crucial for assessing medical students' readiness for clinical practice.[25] Another example is the generation of context-rich short answer questions (CR-SAQs). While traditional multiple-choice questions fail to assess higher cognitive skills effectively, CR-SAQs encourage students to engage more deeply with the material, and foster a better understanding of the subject matter.[26] It was found that the overall quality of questions generated by LLMs and human are comparable.[27] Overall, the incorporation of LLMs into the design and execution of medical education assessments signifies a substantial advancement in educational technology.

To enhance teachers' proficiency in working with LLMs, some academic guidelines have been developed.[28] A structured digital teaching competency framework has been proposed to ensure that medical teachers acquire the requisite skills for effective digital teaching.[29] This framework encompasses general digital competencies, specific digital teaching skills, and expertise in employing innovative digital technologies. Another significant pipeline involves dynamic classrooms, intelligent lesson plans, and personalized learning experiences.[30] As a whole, these guidelines focus on integrating LLMs into the teaching process to facilitate better learning experiences for students.

Customization for Student Learning

LLMs offer advantages in facilitating students' learning process in various dimensions. The language barriers and reading comprehension challenges can be easily addressed by LLMs. One study highlights the importance of linguistic clues in reading comprehension among medical students, suggesting that language interventions could improve their comprehension skills.[31] Additionally, LLMs can be designed to incorporate metacognitive strategies, so as to guide medical students in self-monitoring and employing effective reading techniques.[32] For students with specific language impairments, LLMs provide personalized language trainings, and ensure they can fully engage with medical curricula.[33]

In other cases, the integration of LLMs into clinical decision support systems has shown promising outcomes in improving the interpretation of medical guidelines.[34] In a double-blind randomized study conducted by Emilia, a control group engaging in conversations with AI patients was compared to an intervention group receiving additional feedback from LLMs. After four sessions, the feedback group demonstrated superior performance in creating context and securing information, indicating that AI-simulated conversations with feedback can aid clinical decision-making.[35] Furthermore, LLMs increase engagement in self-regulated learning in student-centered curricula, where students are encouraged to plan, monitor, and evaluate their learning strategies.[36] Through a randomized controlled trial involving 129 undergraduate medical students, researchers observed that students utilizing ChatGPT not only performed better in short-term orthopedics tests but also achieved higher scores in the final exams of surgery and obstetrics and gynecology.[37] Moreover, in collaborative learning environments, LLMs serve as a convenient platform for students to discuss and share knowledge. The use of social media platforms, such as WeChat, alongside LLMs, has been shown to support interactive and participatory teaching methodologies.[38]

Existing research has evaluated the performance of ChatGPT in advanced Cardiovascular Life Support Training exams, the United State Medical Licensing Examination, and Japanese national medical exams. These findings demonstrated that

ChatGPT reached a high level of proficiency so they could also be used to improve the students' performance.[39] However, in our example the answer is incorrect, partly because of the difficulty in identifying images (Figure 3D). In these cases, students should avoid relying on LLMs for every task, like some work requiring students to reflect and document their personal feelings. Engaging in reflective practices allows students to critically evaluate their experiences, potentially disrupting negative perceptions and enhancing their learning outcomes. It is found that nursing students' perceptions of working with older adults are significantly influenced by their personal experiences and the social contexts they encounter.[40]

In terms of the student feedback, the integration of LLMs into educational practices is demonstrated to significantly enhance their learning experience. One study reported a high level of satisfaction (7.9/10) among students when incorporating LLMs into medical education.[41] Park found that a significantly higher percentage of students supported the use of ChatGPT in class (75.6% vs 17.1%).[42] For residency education, ChatGPT received positive perceptions regarding ease of use (4.48/5), and helpfulness (4.61/5). Nevertheless, the reasonability (4.00/5) required further improvement.[43] Additionally, research indicated that positive openness to LLMs correlated with increased satisfaction and higher applicability.[44] Therefore, by providing access to a vast array of information and facilitating interactive learning environments, LLMs are revolutionizing traditional educational paradigms and improving student outcomes across diverse disciplines.

Promotion of Academic Performance

The efficacy of LLMs in processing and synthesizing information is particularly beneficial in academic contexts, where time constraints and the necessity for comprehensive data analysis are prevalent. LLMs are capable of creating drafts based on users' prompts, allowing students to request abstracts, bibliographies, tables of contents, data analysis, literature reviews, and rapid summaries of key points within minutes – tasks that previously required days or even weeks to complete. As an example, four LLMs, namely FlanT5, OpenHermes-NeuralChat, Mixtral and Platypus 2, were used to screen titles and abstracts in systematic reviews, demonstrating promising results in the automation of publications.[45] In addition, LLMs can conduct thematic analysis of qualitative data, and achieve substantial similarity to human-generated themes.[46]

The integration of LLMs into academic writing and publishing is gaining traction. PubReCheck guides researchers in preparing text-mining-ready articles, thereby improving the discoverability and impact of their work.[47] Likewise, BioVisReport creates interactive websites for visualizing published data, thus enhancing the reproducibility and accessibility of research findings.[48] Given the prevalence of grammatical and scientific errors in existing literature, LLMs offer valuable feedback on revisions, including grammar and spelling corrections, as well as suggest alternative phrases to enhance the overall quality of the text.[49] The styles of a manuscript can be adjusted based on the user's needs and receive a score. Furthermore, the role of LLMs in academic writing extends to the citation and reference management. The OpCitance project, which involves identifying citation contexts in PubMed Central articles, illustrates the potential to facilitate reference mapping and ensure that citations are accurately formatted and linked to their corresponding sources.[50] In summary, LLMs provide a multifaceted approach to improving academic manuscripts by offering feedback on grammar and spelling, suggesting alternative phrases, aiding in citation management, and supporting educational initiatives.

By offering direct access to a wide array of academic resources, LLMs enable the exploration of more thoroughly, nuanced, and well-informed perspectives. This, in turn, leads to improved assignment outcomes and a stronger understanding of researchers. LLMs can also be leveraged to automatically generate machine-readable protocols from scientific publications. ProtoCode can interpret and curate knowledge from complex scientific literature, converting literature-based protocols into operational files suitable for laboratory equipment.[51]

Challenges and Solutions About LLMs

Although LLMs show great values in multiple aspects, they still face several challenges in the context of medical education, such as automation bias, hidden hallucination problems, and the model interpretability.

Automation Bias

Automation bias arises when students rely excessively on LLMs. It means students are easy to accept responses, or even wrong answers, without question, and their critical thinking, problem-solving, and innovation abilities are therefore restricted.[52,53] Young medical students, as digital natives, are comfortable with internet information, and may assume that LLMs are always accurate, due to limited expertise, and lower overall confidence. Like the response in Figure 3D, despite LLMs provided explicit reasoning, their answer is incorrect. Research on testing the answers of ChatGPT, ERNIE Bot, and ChatGLM about breast cancer reveals that while LLMs can provide comprehensive answers, their accuracy is notably lower for specialized topics.[54] Importantly, it remains the responsibility of physicians to verify the accuracy of LLM's answers when submitting their assignments.[55] Otherwise, a physician exhibiting automation bias may apply flawed conclusions in patient care, potentially causing significant and irreversible harm.[56]

Solutions: (1) It is crucial to underscore the necessity of not relying solely on LLMs but engaging in meaningful learning exercises. Educators should teach students to critically assess the accuracy and reliability of information from LLMs. These include fostering a student-centric evaluation approach to ensure students can discern and process information effectively, instilling a strong sense of responsibility and ethical awareness in their learning journey. With proper guidance, knowledge-based chatbots may benefit medical students' academic performance, critical thinking, and learning satisfaction.[57]

(2) Facing LLMs, students are encouraged to question results rather than passively accept them. It is consistent with the Illusion of Explanatory Depth, where students are encouraged to delve deeper into the information provided by LLMs.[58] The concept of active engagement with LLMs is further supported by research in cognitive enhancement, where students are encouraged to make deliberate choices rather than impulsive decisions.

Hallucination

LLMs prioritize generating the most suitable response based on input, and they cannot verify the accuracy of statements. LLMs may produce references and clinical cases that appear legitimate but are fabricated or incorrect, and this condition is called hallucinations. Hallucinations can present as fabricated bibliographic citations that do not correspond to genuine scholarly works. An investigation into ChatGPT-3.5 and ChatGPT-4 revealed that a considerable proportion of the citations generated by these models were fabricated, with 55% of GPT-3.5 citations and 18% of GPT-4 citations being non-existent.[59] Another study assessing ChatGPT's performance in managing distal radius fractures found that a significant portion of the references provided by the model were fabricated.[60] In certain cases, Bard even failed to retrieve any relevant papers and had a hallucination rate of 91.4%.[61] This potential for misinformation can undermine the academic integrity of institutions and may lead to disciplinary actions. Articles that include hallucinations has happened in journals like Medical Teacher.[62] In severe conditions, plagiarism may happen, which is a more pressing concern.

Solutions: (1) To prevent the misuse of LLMs, it is critical for educators to cultivate self-discipline and ethical awareness among students. This can be achieved by guiding students to establish realistic goals, manage their time efficiently, and devise strategies to overcome challenges within supportive learning environments.[63] Additionally, comprehensive ethics education programs should be developed to teach medical students to apply these principles in practical, real-world situations.[64]

(2) The existence of hallucinations necessitates the establishment of robust guidelines, frameworks and tools to ensure the responsible utilization of LLMs.[65] Plagiarism detection systems, AI output detectors, and human reviewers are crucial in identifying hallucinations. Turnitin and iThenticate are common plagiarism detection software. These tools work by comparing submitted texts against extensive databases of academic publications, internet sources, and previously submitted documents to identify similarities.[66] AI output detectors assess the structure and style of the text to determine the likelihood of AI generation, with some tools achieving high accuracy while others performed less effectively.[67,68] However, despite the capabilities of these tools, human judgment remains essential for the accurate interpretation and validation of these findings.

(3) Most journals and publishers have implemented new strategies to detect and avoid inappropriate use of the AI. Submitting authors are required to disclose any relevant AI technologies used in their manuscripts.[62] Such transparency is vital for sustaining trust within the scientific community and guaranteeing the ethical application of LLM technologies.[69] The journal *Medical Teacher* proposed eight key lessons, emphasizing the importance of carefully verifying citations and references, as even page numbers can provide critical clues.[62]

(4) Institutions can adopt a multifaceted approach to ensure the accuracy, reliability, and ethical standards of LLM-generated content. By reviewing existing frameworks and criteria from other fields, institutions can develop robust evaluation frameworks to assess the external validity, applicability, and transferability.[70] This enables institutions to better determine whether the content is applicable and useful in various contexts. Furthermore, institutions can enhance the content and structure of educational resources, such as LibGuides, to improve the usability and accessibility, as demonstrated in the field of dentistry.[71]

Incorrect Response

A significant risk associated with ChatGPT 3.0 stems from its inability to access the most up-to-date information. ChatGPT was found to provide incorrect or incomplete response concerning pharmacological data, posing a high risk of harm to patients.[72] It has also led to the dissemination of incorrect details concerning teratogenic drugs, potentially resulting in unwarranted pregnancy terminations.[73,74] Moreover, there have been instances where LLMs inappropriately suggested that a simulated mental health patient should engage in self-harm.[75] Consequently, the inability of ChatGPT to access the most recent data warrants careful scrutiny.

Inaccurate responses can manifest in geographical data and exhibit discriminatory tendencies. As LLMs are primarily trained on data from resource-rich regions, there is a consequent scarcity of research and applications in underrepresented areas, such as Aotearoa New Zealand, where distinct cultural and social contexts necessitate customized debiasing approaches.[76] On this condition, LLMs may propagate race-based medicine, perpetuating harmful and inaccurate medical practices based on race.[77] LLMs may discriminate against certain groups, including women and individuals of Black descent, even when utilizing ostensibly unbiased prompts.[78] Such biases underscore the imperative for meticulous curation and specialized fine-tuning of training data to mitigate these adverse effects.[30]

Solutions: (1) Addressing biases effectively requires a multifactorial framework that engages raters from diverse backgrounds and expertise. Recently, Stephen et al conducted a large-scale empirical study with the Med-PaLM, and developed a toolbox for identifying health equity harms and biases.[79] The study highlighted the importance of employing diverse assessment methodologies and involving raters with varying backgrounds.[79] It is strongly recommended that LLMs associated ethical implications necessitate a systematic review and should involve diverse stakeholders, including researchers, healthcare professionals, and ethicists, to ensure that the deployment of LLMs is guided by comprehensive ethical frameworks and human oversight.[80]

(2) Embedding systemic equity throughout AI applications is imperative for addressing unfair biases. As bias can be introduced during data collection, algorithm design, and even during the deployment of AI systems, a critical aspect lies in the recognition and mitigation of biases at various stages of AI development and implementation. Researchers can delineate and mitigate unfair biases by employing tailored and comprehensive questionnaires. It helps identify domain-specific socioecological inequities and the selection of appropriate mitigation strategies.[81] Furthermore, there is a need for a regulatory framework to address health inequity in LLM applications. For example, the European Union's regulatory framework on medical devices has set rules for performance and data quality, with continuously evolving measures to address the full spectrum of biases in AI systems.[82]

"Black Box" Effect

Since LLMs use advanced techniques to obtain results, and do not rely on predefined rules, the results are sometimes difficult for users to verify. LLMs cannot explain the sources of their advice, the rationale behind it, or the extent to which ethical considerations have been weighed. This is especially relevant in areas like cardiovascular imaging, where AI systems are used for diagnosis but often lack explain ability, posing legal and ethical challenges.[83] Decision making in clinical cases testing have numerous confounders, including patient age, history, culture conditions, and variations in diagnostic and treatment procedures. Whether will LLMs adapt to these modifications is essential for teachers and students to assess the results.

Solutions: It is of great significance to develop more readable AI systems that can provide more clear and understandable outputs. The Transparency and Interpretability for Understandability framework has been suggested to ensure that LLMs in mental health are comprehensible, thereby fostering trust and usability.[84,85] The development of explainable

AI techniques, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), has been proposed to enhance the interpretability of AI systems in drug discovery, making the decision-making process more transparent and understandable.[84]

Although being difficult, Durán and Jongsma argue for the necessity of design transparency to enhance the trustworthiness of medical AI systems. They propose a framework called computational reliabilism, which aims to provide epistemic warrants for the reliability of AI outputs. However, they acknowledge that justified knowledge is not sufficient for normative justification, emphasizing the need for a more transparent design and implementation of AI systems to address these challenges.[86] Yang et al developed BlastAssist to measure interpretable features, including size, symmetry, fragmentation etc., to replace time-consuming and subjective manual procedure. Readable LLMs in medical education will be powerful resources to promote the development of this field.

Taken together, developing an effective medical education system that incorporates LLMs necessitates collaborative efforts among medical educators, AI scientists, ethicists, regulatory bodies, and publishers is urgent (Figure 2). Medical educators must understand both the strengths and limitations of LLMs, as well as the diverse needs of students. AI scientists play a vital role by providing technical expertise and addressing potential challenges. Ethicists' duty is to ensure that practices across various domains adhere to ethical standards. Regulatory bodies are responsible for ensuring the evaluation system adheres to relevant legal and ethical standards, including the formulation and enforcement of rules to prevent data misuse. Additionally, the involvement of publishers is crucial for addressing scientific opinions and concerns related to academic ethics and writing.[87]

Conclusion

As technology advances, medical education will increasingly rely on technological tools. Teaching becomes more efficient and captivating through utilizing LLMs to design engaging lessons, providing instant feedback, and simulating real-life scenarios. On the other hand, students can carry out individualized learning more effectively based on their own pace and learning styles, and conduct in-depth discussions deepen their understanding. Furthermore, LLMs have demonstrated a vast potential and broad performance in the academic field. However, it is important to recognize that LLM is not inherently powerful, and it may also generate bias, hallucination, and opaque response. Looking ahead, it is essential to educate medical students on scholarly integrity, patient and public safety, and the ethical use of LLMs. Such education will equip them with the necessary skills and knowledge to excel in a rapidly changing healthcare environment. Future research needs to focus on the long-term effects of LLM integration. Longitudinal studies spanning multiple academic years should be designed to evaluate how LLMs impacts students' medical knowledge acquisition, clinical decision-making skills, and professional ethics. Only through such comprehensive efforts can we ensure that future doctors are well-equipped to harness the power of LLMs while safeguarding the quality and ethics of medical education.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This work is financially supported by the grants from the Teaching and Research Project of the Second Clinical College of Tongji Medical College, Huazhong University of Science and Technology (TJXYJ2023016; TJSZ2024016).

Disclosure

The authors declare that they have no competing interests in this work.

References

1. Wong RS, Ming LC, Raja Ali RA. The intersection of ChatGPT, clinical medicine, and medical education. *JMIR Med Educ*. 2023;9:e47274. doi:10.2196/47274
2. Zernikow J, Grassow L, Gröschel J, et al. Clinical application of large language models: does ChatGPT replace medical report formulation? An experience report. *Inn Med*. 2023;64(11):1058–1064.
3. Castonguay A, Farthing P, Davies S, et al. Revolutionizing nursing education through Ai integration: a reflection on the disruptive impact of ChatGPT. *Nurse Educ Today*. 2023;129:105916. doi:10.1016/j.nedt.2023.105916
4. Artsi Y, Sorin V, Konen E, et al. Large language models for generating medical examinations: systematic review. *BMC Med Educ*. 2024;24(1):354. doi:10.1186/s12909-024-05239-y
5. Mannekote A, Davies A, Pinto JD, et al. Large language models for whole-learner support: opportunities and challenges. *Front Artif Intell*. 2024;7:1460364. doi:10.3389/frai.2024.1460364
6. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med Educ*. 2023;9:e48291. doi:10.2196/48291
7. Altintas L, Sahiner M. Transforming medical education: the impact of innovations in technology and medical devices. *Expert Rev Med Devices*. 2024;21(9):797–809. doi:10.1080/17434440.2024.2400153
8. 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*. 2024;26:e60807. doi:10.2196/60807
9. Tozsin A, Ucmak H, Soyturk S, et al. The role of artificial intelligence in medical education: a systematic review. *Surg Innov*. 2024;31(4):415–423. doi:10.1177/15533506241248239
10. Shoureshi PS, Rajasegaran A, Kokorowski P, et al. Social media engagement, perpetuating selected information, and accuracy regarding CA SB-201: treatment or intervention on the sex characteristics of a minor. *J Pediatr Urol*. 2021;17(3):372–377. doi:10.1016/j.jpurol.2021.01.047
11. Torraco R. Writing integrative literature reviews: using the past and present to explore the future. *Human Resource Development Review*. 2016;15.
12. Baethge C, Goldbeck-Wood S, Mertens S. SANRA-a scale for the quality assessment of narrative review articles. *Res Integr Peer Rev*. 2019;4:5. doi:10.1186/s41073-019-0064-8
13. Ammanuel S, Brown I, Uribe J, et al. Creating 3D models from radiologic images for virtual reality medical education modules. *J Med Syst*. 2019;43(6):166. doi:10.1007/s10916-019-1308-3
14. Neyem A, Cadile M, Burgos-Martínez SA, et al. Enhancing medical anatomy education with the integration of virtual reality into traditional lab settings. *Clin Anat*. 2024.
15. Meade O, Bowskill D, Lymn JS. Pharmacology podcasts: a qualitative study of non-medical prescribing students' use, perceptions and impact on learning. *BMC Med Educ*. 2011;11:2. doi:10.1186/1472-6920-11-2
16. Bjurström MF, Borgquist O, Kander T, et al. Audio podcast and procedural video use in anaesthesiology and intensive care: a nationwide survey of Swedish anaesthetists. *Acta Anaesthesiol Scand*. 2024;68(7):923–931. doi:10.1111/aas.14433
17. Zhao Y, Zhang Y, Zhang Y, et al. LEVA: using large language models to enhance visual analytics. *IEEE Trans Vis Comput Graph*. 2024.
18. 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*. 2024;24(1):558. doi:10.1186/s12909-024-05565-1
19. Lopez M, Goh PS. Catering for the needs of diverse patient populations: using ChatGPT to design case-based learning scenarios. *Med Sci Educ*. 2024;34(2):319–325. doi:10.1007/s40670-024-01975-4
20. Dunn C, Hunter J, Steffes W, et al. Artificial intelligence-derived dermatology case reports are indistinguishable from those written by humans: a single-blinded observer study. *J Am Acad Dermatol*. 2023;89(2):388–390. doi:10.1016/j.jaad.2023.04.005
21. Hui Z, Zewu Z, Jiao H, et al. Application of ChatGPT-assisted problem-based learning teaching method in clinical medical education. *BMC Med Educ*. 2025;25(1):50. doi:10.1186/s12909-024-06321-1
22. Smith PE, Mucklow JC. Writing clinical scenarios for clinical science questions. *Clin Med*. 2016;16(2):142–145. doi:10.7861/clinmedicine.16-2-142
23. Huang Y, Xu -B-B, Wang X-Y, et al. Implementation and evaluation of an optimized surgical clerkship teaching model utilizing ChatGPT. *BMC Med Educ*. 2024;24(1):1540. doi:10.1186/s12909-024-06575-9
24. Ríos-Hoyo A, Shan NL, Li A, et al. Evaluation of large language models as a diagnostic aid for complex medical cases. *Front Med*. 2024;11:1380148. doi:10.3389/fmed.2024.1380148
25. Salzman DH, McGaghie WC, Caprio TW, et al. A mastery learning capstone course to teach and assess components of three entrustable professional activities to graduating medical students. *Teach Learn Med*. 2019;31(2):186–194. doi:10.1080/10401334.2018.1526689
26. Bird JB, Olvet DM, Willey JM, et al. Patients don't come with multiple choice options: essay-based assessment in UME. *Med Educ Online*. 2019;24(1):1649959. doi:10.1080/10872981.2019.1649959
27. Laupichler MC, Rother JF, Grunwald Kadow IC, et al. Large language models in medical education: comparing ChatGPT- to human-generated exam questions. *Acad Med*. 2024;99(5):508–512. doi:10.1097/ACM.0000000000005626
28. Raja Indran I, Paranthaman P, Gupta N, et al. Twelve tips to leverage AI for efficient and effective medical question generation: a guide for educators using Chat GPT. *Medical Teacher*. 2023;1–6.
29. Saaiq M, Khan RA, Yasmeen R. Digital teaching: developing a structured digital teaching competency framework for medical teachers. *Med Teach*. 2024;46(10):1362–1368. doi:10.1080/0142159X.2024.2308782
30. Hu T, Kyrychenko Y, Rathje S, et al. Generative language models exhibit social identity biases. *Nat Comput Sci*. 2024;5:65–75. doi:10.1038/s43588-024-00741-1
31. Al-Jamal DA. The role of linguistic clues in medical students' reading comprehension. *Psychol Res Behav Manag*. 2018;11:395–401. doi:10.2147/PRBM.S174087
32. Benedict KM, Rivera MC, Antia SD. Instruction in metacognitive strategies to increase deaf and hard-of-hearing students' reading comprehension. *J Deaf Stud Deaf Educ*. 2015;20(1):1–15. doi:10.1093/deafed/enu026
33. Buil-Legaz L, Aguilar-Mediavilla E, Rodríguez-Ferreiro J. Oral morphosyntactic competence as a predictor of reading comprehension in children with specific language impairment. *Int J Lang Commun Disord*. 2016;51(4):473–477. doi:10.1111/1460-6984.12217

34. Kresevic S, Giuffrè M, Ajcevic M, et al. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *NPJ Digit Med*. 2024;7(1):102. doi:10.1038/s41746-024-01091-y
35. Brügge E, Ricchizzi S, Arenbeck M, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Med Educ*. 2024;24(1):1391. doi:10.1186/s12909-024-06399-7
36. Greenberg A, Olvet DM, Brenner J, et al. Strategies to support self-regulated learning in integrated, student-centered curricula. *Med Teach*. 2023;45(12):1387–1394. doi:10.1080/0142159X.2023.2218538
37. Gan W, Ouyang J, Li H, et al. Integrating ChatGPT in orthopedic education for medical undergraduates: randomized controlled trial. *J Med Internet Res*. 2024;26:e57037. doi:10.2196/57037
38. Wang J, Gao F, Li J, et al. The usability of WeChat as a mobile and interactive medium in student-centered medical teaching. *Biochem mol Biol Educ*. 2017;45(5):421–425. doi:10.1002/bmb.21065
39. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Med Educ*. 2024. doi:10.1111/medu.15402
40. Dahlke S, Davidson S, Kalogirou MR, et al. Nursing faculty and students' perspectives of how students learn to work with older people. *Nurse Educ Today*. 2020;93:104537. doi:10.1016/j.nedt.2020.104537
41. Baglivo F, De Angelis L, Casigliani V, et al. Exploring the possible use of AI Chatbots in public health education: feasibility study. *JMIR Med Educ*. 2023;9:e51421. doi:10.2196/51421
42. Park J. Medical students' patterns of using ChatGPT as a feedback tool and perceptions of ChatGPT in a Leadership and Communication course in Korea: a cross-sectional study. *J Educ Eval Health Prof*. 2023;20:29. doi:10.3352/jeehp.2023.20.29
43. Shang L, Li R, Xue M, et al. Evaluating the application of ChatGPT in China's residency training education: an exploratory study. *Med Teach*;2024:1–7. doi:10.1080/0142159X.2024.2377808
44. 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*. 2024;10:e50545. doi:10.2196/50545
45. Dennstädt F, Zink J, Putora PM, et al. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. *Syst Rev*. 2024;13(1):158. doi:10.1186/s13643-024-02575-4
46. Mathis WS, Zhao S, Pratt N, et al. Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: how does it compare to traditional methods? *Comput Methods Programs Biomed*. 2024;255:108356. doi:10.1016/j.cmpb.2024.108356
47. Leaman R, Wei C-H, Allot A, et al. Ten tips for a text-mining-ready article: how to improve automated discoverability and interpretability. *PLoS Biol*. 2020;18(6):e3000716. doi:10.1371/journal.pbio.3000716
48. Yang J, Liu Y, Shang J, et al. BioVisReport: a Markdown-based lightweight website builder for reproducible and interactive visualization of results from peer-reviewed publications. *Comput Struct Biotechnol J*. 2022;20:3133–3139. doi:10.1016/j.csbj.2022.06.009
49. Burgess RR. A brief review of common grammatical and scientific errors seen in reviewing protein purification manuscripts for 25 years. *Protein Expr Purif*. 2016;120:106–109. doi:10.1016/j.pep.2015.12.002
50. Hsiao TK, Torvik VI. OpCitance: citation contexts identified from the PubMed Central open access articles. *Sci Data*. 2023;10(1):243. doi:10.1038/s41597-023-02134-x
51. Jiang S, Evans-Yamamoto D, Bersenev D, et al. ProtoCode: leveraging large language models (LLMs) for automated generation of machine-readable PCR protocols from scientific publications. *SLAS Technol*. 2024;29(3):100134. doi:10.1016/j.slast.2024.100134
52. Fijačko N, Gosak L, Štiglic G, et al. Can ChatGPT pass the life support exams without entering the American heart association course? *Resuscitation*. 2023;185:109732. doi:10.1016/j.resuscitation.2023.109732
53. Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the National Nurse Examinations in Japan: evaluation study. *JMIR Nurs*. 2023;6:e47305. doi:10.2196/47305
54. Piao Y, Chen H, Wu S, et al. Assessing the performance of large language models (LLMs) in answering medical questions regarding breast cancer in the Chinese context. *Digit Health*. 2024;10:20552076241284771. doi:10.1177/20552076241284771
55. Shang L, Xue M, Hou Y, et al. Can ChatGPT pass China's national medical licensing examination? *Asian J Surg*. 2023;46(12):6112–6113. doi:10.1016/j.asjsur.2023.09.089
56. Funer F, Liedtke W, Tinnemeyer S, et al. Responsibility and decision-making authority in using clinical decision support systems: an empirical-ethical exploration of German prospective professionals' preferences and concerns. *J Med Ethics*. 2023;50(1):6–11. doi:10.1136/jme-2022-108814
57. Chang C-Y, Kuo S-Y, Hwang G-H. Chatbot-facilitated nursing education: incorporating a knowledge-based Chatbot system into a nursing training program. *Educational Technology & Society*. 2022.
58. Mehta S, Mehta N. Embracing the illusion of explanatory depth: a strategic framework for using iterative prompting for integrating large language models in healthcare education. *Med Teach*. 2024;1–4. doi:10.1080/0142159X.2024.2418937
59. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep*. 2023;13(1):14045. doi:10.1038/s41598-023-41032-5
60. Knee CJ, Campbell RJ, Graham DJ, et al. Examining the role of ChatGPT in the management of distal radius fractures: insights into its accuracy and consistency. *ANZ J Surg*. 2024;94(7–8):1391–1396. doi:10.1111/ans.19143
61. Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. *J Med Internet Res*. 2024;26:e53164. doi:10.2196/53164
62. Masters K. Medical Teacher's first ChatGPT's referencing hallucinations: lessons for editors, reviewers, and teachers. *Med Teach*. 2023;45(7):673–675. doi:10.1080/0142159X.2023.2208731
63. Hagger MS, Hamilton K. Grit and self-discipline as predictors of effort and academic attainment. *Br J Educ Psychol*. 2019;89(2):324–342. doi:10.1111/bjep.12241
64. Cannaerts N, Gastmans C, Dierckx de Casterlé B. Contribution of ethics education to the ethical competence of nursing students: educators' and students' perceptions. *Nurs Ethics*. 2014;21(8):861–878. doi:10.1177/0969733014523166
65. Zhui L, Fenghe L, Xuehu W, et al. Ethical considerations and fundamental principles of large language models in medical education: viewpoint. *J Med Internet Res*. 2024;26:e60083. doi:10.2196/60083
66. Meo SA, Talha M. Turnitin: is it a text matching or plagiarism detection tool? *Saudi J Anaesth*. 2019;13(Suppl 1):S48–s51. doi:10.4103/sja.SJA_772_18

67. Kar SK, Bansal T, Modi S, et al. How sensitive are the free AI-detector tools in detecting AI-generated texts? A comparison of popular AI-detector tools. *Indian J Psychol Med*;2024:02537176241247934. doi:10.1177/02537176241247934
68. Hassanipour S, Nayak S, Bozorgi A, et al. The ability of chatgpt in paraphrasing texts and reducing plagiarism: a descriptive analysis. *JMIR Med Educ*. 2024;10:e53308. doi:10.2196/53308
69. Peh W, Saw A. Artificial intelligence: impact and challenges to authors, journals and medical publishing. *Malays Orthop J*. 2023;17(3):1–4. doi:10.5704/MOJ.2311.001
70. Burchett H, Umoquit M, Dobrow M. How do we know when research from one setting can be useful in another? A review of external validity, applicability and transferability frameworks. *J Health Serv Res Policy*. 2011;16(4):238–244. doi:10.1258/jhsrp.2011.010124
71. Jones EP. Use of large language model (LLM) to enhance content and structure of a school of dentistry LibGuide. *J Med Libr Assoc*. 2025;113 (1):96–97. doi:10.5195/jmla.2025.2084
72. Morath B, Chiriac U, Jaszkowski E, et al. Performance and risks of ChatGPT used in drug information: an exploratory real-world analysis. *Eur J Hosp Pharm*. 2024;31(6):491–497. doi:10.1136/ejhpharm-2023-003750
73. Han JY. Usefulness and limitations of ChatGPT in getting information on teratogenic drugs exposed in pregnancy. *Obstet Gynecol Sci*. 2024.
74. Gilbert S, Harvey H, Melvin T, et al. Large language model AI chatbots require approval as medical devices. *Nat Med*. 2023;29(10):2396–2398. doi:10.1038/s41591-023-02412-6
75. Daws R. Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. AINews; 2020. Available from: https://www.artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/. Accessed April 15, 2025.
76. Yogarajan V, Dobbie G, Keegan TT. Debiasing large language models: research opportunities. *J R Soc N Z*. 2025;55(2):372–395. doi:10.1080/03036758.2024.2398567
77. Omiye JA, Lester JC, Spichak S, et al. Large language models propagate race-based medicine. *NPJ Digit Med*. 2023;6(1):195. doi:10.1038/s41746-023-00939-z
78. Fang X, Che S, Mao M, et al. Bias of AI-generated content: an examination of news produced by large language models. *Sci Rep*. 2024;14(1):5224. doi:10.1038/s41598-024-55686-2
79. Pfohl SR, Cole-Lewis H, Sayres R, et al. A toolbox for surfacing health equity harms and biases in large language models. *Nat Med*. 2024;30 (12):3590–3600. doi:10.1038/s41591-024-03258-2
80. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *NPJ Digit Med*. 2024;7(1):183. doi:10.1038/s41746-024-01157-x
81. Bozeman JF, Hollauer C, Ramshankar AT, et al. Embed systemic equity throughout industrial ecology applications: how to address machine learning unfairness and bias. *J Ind Ecol*. 2024;28(6):1362–1376. doi:10.1111/jiec.13509
82. van Kolfschooten H. The AI cycle of health inequity and digital ageism: mitigating biases through the EU regulatory framework on medical devices. *J Law Biosci*. 2023;10(2):lsad031. doi:10.1093/jlb/lsad031
83. Lang M, Bernier A, Knoppers BM. Artificial intelligence in cardiovascular imaging: "unexplainable" legal and ethical challenges? *Can J Cardiol*. 2022;38(2):225–233. doi:10.1016/j.cjca.2021.10.009
84. Kırboğa KK, Abbasi S, Küçüksille EU. Explainability and white box in drug discovery. *Chem Biol Drug Des*. 2023;102(1):217–233. doi:10.1111/cbdd.14262
85. Joyce DW, Kormilitzin A, Smith KA, et al. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *NPJ Digit Med*. 2023;6(1):6. doi:10.1038/s41746-023-00751-9
86. Ferrario A. Design publicity of black box algorithms: a support to the epistemic and ethical justifications of medical AI systems. *J Med Ethics*. 2022;48(7):492–494. doi:10.1136/medethics-2021-107482
87. Tang YD, Dong ED, Gao W. LLMs in medicine: the need for advanced evaluation systems for disruptive technologies. *Innovation*. 2024;5 (3):100622. doi:10.1016/j.xinn.2024.100622

Advances in Medical Education and Practice

Dovepress
Taylor & Francis Group

Publish your work in this journal

Advances in Medical Education and Practice is an international, peer-reviewed, open access journal that aims to present and publish research on Medical Education covering medical, dental, nursing and allied health care professional education. The journal covers undergraduate education, postgraduate training and continuing medical education including emerging trends and innovative models linking education, research, and health care services. The manuscript management system is completely online and includes a very quick and fair peer-review system. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

Submit your manuscript here: http://www.dovepress.com/advances-in-medical-education-and-practice-journal