Received: 17 August 2023 | Revised: 20 March 2024 | Accepted: 23 March 2024

DOI: 10.1111/medu.15402

REVIEW ARTICLE

Medical Education In Review

A systematic review of large language models and their implications in medical education

Harrison C. Lucas[1] | Jeffrey S. Upperman[2] | Jamie R. Robinson[2,3]

[1]Brandeis University, Waltham, Massachusetts, USA

[2]Department of Pediatric Surgery, Vanderbilt University Medical Center, Nashville, Tennessee, USA

[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA

Correspondence
Jamie R. Robinson, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2200 Children's Way, 7th Floor DOT, Suite 7100, Nashville, TN 37232, USA.
Email: jamie.robinson@vumc.org

Funding information
Not applicable.

Abstract

Introduction: In the past year, the use of large language models (LLMs) has generated significant interest and excitement because of their potential to revolutionise various fields, including medical education for aspiring physicians. Although medical students undergo a demanding educational process to become competent health care professionals, the emergence of LLMs presents a promising solution to challenges like information overload, time constraints and pressure on clinical educators. However, integrating LLMs into medical education raises critical concerns and challenges for educators, professionals and students. This systematic review aims to explore LLM applications in medical education, specifically their impact on medical students' learning experiences.

Methods: A systematic search was performed in PubMed, Web of Science and Embase for articles discussing the applications of LLMs in medical education using selected keywords related to LLMs and medical education, from the time of ChatGPT's debut until February 2024. Only articles available in full text or English were reviewed. The credibility of each study was critically appraised by two independent reviewers.

Results: The systematic review identified 166 studies, of which 40 were found by review to be relevant to the study. Among the 40 relevant studies, key themes included LLM capabilities, benefits such as personalised learning and challenges regarding content accuracy. Importantly, 42.5% of these studies specifically evaluated LLMs in a novel way, including ChatGPT, in contexts such as medical exams and clinical/biomedical information, highlighting their potential in replicating human-level performance in medical knowledge. The remaining studies broadly discussed the prospective role of LLMs in medical education, reflecting a keen interest in their future potential despite current constraints.

Conclusions: The responsible implementation of LLMs in medical education offers a promising opportunity to enhance learning experiences. However, ensuring information accuracy, emphasising skill-building and maintaining ethical safeguards are crucial. Continuous critical evaluation and interdisciplinary collaboration are essential for the appropriate integration of LLMs in medical education.

1 | INTRODUCTION

In recent years, the advent of large language models (LLMs) has piqued interest in an array of fields, specifically in medical education. Although the underlying technologies have undergone decades of evolution, the widespread availability to the public has expanded utilisation across a variety of applications. LLMs such as OpenAI's GPT-4 and Google's Bard models are the representations of years of research of natural language processing and machine learning.[1,2] These complex models trained on copious amounts of data can generate, interpret and respond to human language with unprecedented accuracy. Their ubiquity offers a myriad of potential clinical uses, being capable of serving as personalised learning tools and aiding in decision-making. These rapidly changing technologies also come at a time of rapid evolution in medical education.[3] However, integrating these technologies into medical education poses challenges and raises concerns about accuracy, ethical implications and detriments to critical thinking. Despite the surge in interest, there remains a notable gap in the comprehensive evaluation of LLMs within medical educational settings. In this review, we provide a comprehensive investigation of current knowledge on LLM applications in medical education. We aim to analyse the integration of LLMs with the intention of promoting responsible use and guiding informed adoption of these tools.

2 | METHODS

The aim of this systematic review was to review and analyse the existing literature, focusing on its impact on medical education. This study adhered to the PRISMA 2020 guidelines, ensuring a rigorous evaluation of LLMs in medical education (Figure 1). We conducted a comprehensive search utilising keywords related to LLMs and medication education. This involved database searches in PubMed, Web of Science and Embase. Additional studies were identified through online news resources. The search was restricted to studies published post-ChatGPT's debut, reflecting the field's recent developments. Inclusion criteria were studies on LLMs in educational and clinical settings. Articles that were not available in full text or the English language were excluded. Final article selection was performed by two independent reviewers (HL and JU) who initially screened titles and abstracts, followed by a detailed full-text evaluation based on relevance to LLMs in medical education. The review involved evaluation of studies for credibility in the methods and results, followed by a synthesis of themes around LLM utilisation in medical education. Discrepancies were resolved through discussion to ensure a balanced selection. A qualitative analysis was performed based upon common themes within the studies. This approach provided a comprehensive analysis, informing future policies and practices in AI technology within academic medicine.

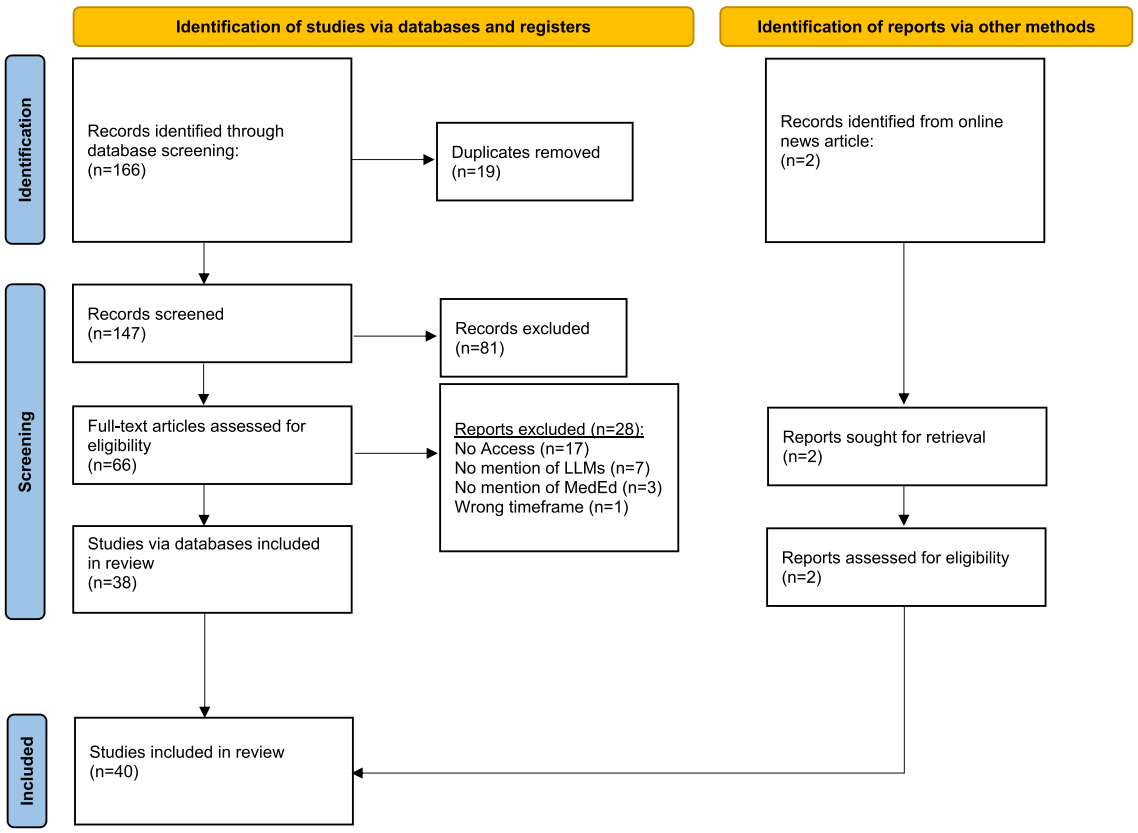

FIGURE 1 Systematic review flow chart. The flow chart illustrates the studies and news articles included and excluded from the review as well as the reasons for exclusion. [Color figure can be viewed at wileyonlinelibrary.com]

3 | RESULTS

3.1 | Study selection

Through an initial database keyword search, 166 records were identified and screened. Duplicates were removed (19 abstracts), and the remaining studies were screened by their abstract, resulting in a total of 66 potentially eligible articles. After exclusion of articles not meeting inclusion criteria, 38 studies were identified through a database search. An additional two studies were identified via online news sources, resulting in a total of 40 studies included in the systematic review (Table 1).

3.2 | Applications of LLMs

Recent studies have shown that LLMs like ChatGPT can be highly beneficial in medical education. These models have demonstrated competence in multiple standardised clinical exams, potentially making them a valuable reference for medical students and educators. For instance, in a study conducted by Kung et al., ChatGPT was able to pass all written steps of the United States Medical Licensing Exam (USMLE) (consisting of Step 1, Step 2CK and Step 3) without any input from human trainers, reflecting knowledge typically gained in the entirety of a medical school education.[4] For medical students who are required to learn an incredible amount of information in a short time, LLMs provide immediate access to medical knowledge, research studies and clinical guidelines. From a high-level perspective, with ready access to answers on how to manage traditional or even complex clinical scenarios at the fingertips of students and clinicians, it raises the question of what truly is the fundamental medical knowledge and comprehension required of medical students prior to advancing to higher studies. Although it remains unclear the full potential of how the utilisation of this resource will assist medical educators, the utilisation of AI may assist educators in creation of tests with improved ability to measure students' knowledge in realistic settings.

After evaluating ChatGPT for performance in a simulated exam for the specialised field of ophthalmology, Antaki et al. found that it showed promising performance but suggested that further specialisation of LLMs could enhance their capabilities.[5] In a study on the Polish Medical Final Examination, Rosol et al. discovered that GPT-4 had a higher passing rate compared with other versions, yet the scores were still lower than those achieved by medical students themselves.[6] This indicates that there certainly are limitations in comparison with human knowledge, and although LLMs show remarkable potential, they typically fall short of human performance in certain aspects of medical knowledge and application.

In other medical disciplines, LLMs have demonstrated greater promise and potential. For instance, ChatGPT attained a passing score on the Canadian Otolaryngology-Head and Neck Surgery Board exams.[7] Another study showed that ChatGPT exhibited 76.5% accuracy in the Korean general surgery board exam.[8] Furthermore, ChatGPT displayed impressive performance on cardiology questions derived from the European Exam in Core Cardiology.[9] In plastic surgery, ChatGPT has been found to perform similarly to a first-year plastic surgery resident on the Plastic Surgery In-Service Examination.[10] Abdel-Messih and Boulos assessed ChatGPT's performance using a clinical toxicology case and found that it responded skilfully to questions and provided relevant information.[11] Additionally, GPT-4 showcased success in responding to advanced neurosurgery case scenarios for oral board preparation, outperforming ChatGPT and Google's Bard.[12] What is notable about these examination successes is the varied but specialised fields with which the LLMs demonstrated aptitude. However, exam performance alone may not always be indicative of the clinical knowledge of a student or clinician. Along similar lines, good performance of the LLMs, albeit impressive, does not solely provide practical guidance or use of these technologies in the education of future physicians.

Medical education and training have the potential to use LLMs like ChatGPT as a tool to enhance traditional coursework and teaching (Figure 2). These models can provide physician trainees with realistic clinical scenarios and feedback, thus enhancing clinical education.[13] Further, GPT-4 has also been shown to generate high-quality dermatology case reports and demonstrate proficiency in radiation oncology exam questions, clinical care paths and paediatrics.[14,15] From a literacy perspective, LLMs like GPT-4 can assist medical students and educators by ensuring writing accuracy, improving style and formatting for clarity and coherence and providing appropriate language and terminology.[16,17] LLMs can also aid medical students in ensuring that comprehensive evaluation of literature is performed and that the information is synthesised in a readily comprehendible form.[18,19] When confirmed to be accurate, certain LLM functions such as gathering and organising information can save time for busy clinicians and support the education of their learners. Additionally, these functions can offer student-learners nearly real-time feedback on current clinical experiences. Regarding educators, LLMs facilitate teaching by providing curriculum and assessment planning, grading rubrics and support educators and administration by allowing them to effectively allocate resources.[20] In general, LLMs offer instant feedback and support, interactive learning experiences and scalability, for both students and educators. This provides an opportunity for artificial intelligences to have economical and widespread impact by enhancing education that is not bounded by socioeconomic or physical constraints. Further, by improving education efficiency, LLMs can help reduce the shortage of medical educators, especially in resource-constrained environments and provide more time for educators to focus on individual student mentorship.

LLMs offer an exciting opportunity for students to explore subjects in greater detail and gain insights by analysing larger amounts of data to solve current problems (Figure 2). LLM systems can improve clinical reasoning and problem-solving by allowing students to ask questions, receive explanations and generate hypotheses about diseases and biological processes.[21] These technologies can become utilised early in medical education even in non-clinical environments, integrating with current resources such as flashcards and practice

TABLE 1 Summary of included sources.

Authors	Study title	Key takeaway	Ref
Kung et al.	Performance of ChatGPT on USMLE	Demonstrates ChatGPT's potential for AI-assisted medical education, achieving near-passing scores on the USMLE without specialised training.	[4]
Antaki et al.	Evaluating the Performance of ChatGPT in Ophthalmology	Evaluates ChatGPT's performance in ophthalmology medical education. ChatGPT showed improved accuracy on multiple-choice questions. Domain-specific training may enhance performance.	[5]
Rosol et al.	Evaluation of the Performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination	Evaluated ChatGPT (GPT-3.5) and GPT-4 in medical education exams, finding GPT-4 outperformed GPT-3.5, showing potential for medical education support.	[6]
Long et al.	Evaluating ChatGPT-4 in Otolaryngology–Head and Neck Surgery Board Examination	ChatGPT-4 performs well on open-ended medical board exams, showing potential for clinical use, but raises safety concerns due to occasional hallucinations.	[7]
Oh et al.	ChatGPT Goes to the Operating Room	LLMs, like GPT-4, show promise in understanding complex surgical information, achieving 76.4% accuracy in the Korean general surgery board exam.	[8]
Skalidis et al.	ChatGPT Takes on the European Exam in Core Cardiology	ChatGPT performs well on post-graduate medical exams.	[9]
Humar et al.	ChatGPT Is Equivalent to First-Year Plastic Surgery Residents	ChatGPT performs at the level of a first-year resident in Plastic Surgery In-Service Examination but is outperformed by more advanced residents.	[10]
Abdel-Messih et al.	ChatGPT in Clinical Toxicology	ChatGPT correctly answered questions relating to clinical toxicology, highlighting the potential for AI-assisted medical education and diagnosis.	[11]
Ali et al.	Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards	Compared the performance of three LLMs on a neurosurgery exam, with GPT-4 outperforming GPT-3.5 and Google's Bard.	[12]
Webb	Proof of Concept: Using ChatGPT to Teach Emergency Physicians	ChatGPT was used as a tool for physicians to learn how to break bad news effectively to patients but requires further research.	[13]
Huang et al.	Benchmarking ChatGPT-4 on ACR Radiation Oncology In-Training (TXIT) Exam	With potential for personalised treatment suggestions, LLMs show promise in radiation oncology.	[14]
Dunn et al.	Artificial Intelligence–derived Dermatology Case Reports are Indistinguishable from Those Written by Humans	LLMs generate medical case reports that are indistinguishable from human-written content in dermatology.	[15]
Abd-Alrazaq et al.	Large Language Models in Medical Education	The potential benefits and challenges of LLM integration in medical education are discussed.	[16]
Gandhi et al.	ChatGPT: Roles and Boundaries of the New Artificial Intelligence Tool in Medical Education	Though concerns are raised about academic integrity and authorship in research, LLMs can aid medical education.	[17]
Liu et al.	Large Language Models are Few-Shot Health Learners	As a result of fine-tuning, LLMs analyse medical data, such as vital signs and activity levels.	[18]
Parsa et al.	ChatGPT in Medicine; a Disruptive Innovation or Just One Step Forward?	New opportunities and challenges of LLMs in medical education and knowledge assessment are presented.	[19]
Shorey et al.	A Scoping Review of ChatGPT's Role in Healthcare Education and Research	Discusses ChatGPT's emergence in health care education and research, being able to aid in assessments and teaching, though suffers from hallucinations and biases.	[20]
Ahn	The Impending Impacts of Large Language Models on Medical Education	LLMs can simulate patient interactions, give oral evaluations and create a dynamic learning environment.	[21]
Singh et al.	Implications and Future directions of ChatGPT Utilization in Neurosurgery	LLMs can provide personalised learning resources and interactive clinical simulations.	[22]

(Continues)

TABLE 1 (Continued)

Authors	Study title	Key takeaway	Ref
Sng et al.	Potential and Pitfalls of ChatGPT and Natural-Language Artificial Intelligence Models for Diabetes Education	Assessed ChatGPT's ability to provide diabetes self-management advice. ChatGPT demonstrated consistency and clear instructions, but lacked specificity in certain areas.	[23]
Wang et al.	Accelerating the Integration of ChatGPT and other Large-scale AI models into Biomedical Research and Healthcare	LLM benefits and applications in medical education are highlighted.	[24]
Qui et al.	Large AI Models in Health Informatics: Applications, Challenges, and the Future	Implementation of LLMs in medical education and health informatics is explored.	[25]
Han et al.	MedAlpaca -- An Open-Source Collection of Medical Conversational AI Models and Training Data	Creates a large dataset designed to train LLMs to support medical education. Author suggests open-source models to protect patient privacy.	[26]
Tian et al.	Opportunities and Challenges for ChatGPT and Large Language Models in Biomedicine and Health	Explores the applications of LLMs in biomedicine and health. Shows promising results in text generation, but problems in other areas.	[27]
Karabacak et al.	Embracing Large Language Models for Medical Applications: Opportunities and Challenges	Through collaboration, education and validation, LLMs can revolutionise medicine. Challenges are discussed.	[28]
Wang et al.	ChatCAD: Interactive Computer-Aided Diagnosis on Medical Image using Large Language Models	Discusses the integration LLMs with medical-image CAD networks to enhance clinical decision-making, improving diagnostics.	[29]
Desaire et al.	ChatGPT or academic scientist? Distinguishing Authorship with over 99% Accuracy using Off-the-shelf Machine Learning Tools	Highlights a method to distinguish AI-generated text from human text in academic contexts with high accuracy.	[30]
Komorowski et al.	How could ChatGPT Impact My Practice as an Intensivist?	LLMs have the potential to support clinical decision-making and be a robust educational tool but have limitations and risks.	[31]
Lee	The Rise of ChatGPT: Exploring its Potential in Medical Education	LLMs can aid in curriculum design, offering promising applications in medical education. Ethical concerns do arise.	[32]
Arif et al.	The Future of Medical Education and Research: Is ChatGPT a Blessing or Blight in Disguise?	Addresses the potential benefits and ethical concerns of LLM integration.	[33]
Huh	Can We Trust AI Chatbots' Answers about Disease Diagnosis and Patient Care?	Evaluates ChatGPT for medical diagnosis and treatment suggestions. Results show its limitations and discuss need for expert judgement.	[34]
Gunawardene et al.	Teaching the Limitations of Large Language Models in Medical School	Acknowledges beneficial use and highlights notable limitations of LLMs in medical school training.	[35]
Loh	ChatGPT and Generative AI Chatbots: Challenges and Opportunities for Science, Medicine and Medical Leaders	LLMs present opportunities in medical education and health care leadership, but challenges are present.	[36]
Gravel et al.	Learning to Fake It: Limited Responses and Fabricated References Provided by ChatGPT for Medical Questions	Evaluates ChatGPT's responses to medical questions, revealing limited quality answers and fabricated references. Caution is advised when using it for medical transcripts.	[37]
Arachchige	Large Language Models (LLM) and ChatGPT: A Medical Student Perspective	Discusses the impact of LLMs on medical education. Highlights concerns about plagiarism and potential benefits.	[38]
De Angelis et al.	ChatGPT and the Rise of Large Language Models: The New AI-driven Infodemic Threat in Public Health	ChatGPT can be utilised as a tool for research support and impact academic research, though there are concerns about misinformation.	[39]
Bair et al.	Large Language Models and Their Implications on Medical Education	While LLMs show high accuracy in medical tasks, study highlights the need for competency guidelines in their use in medical education and practice.	[41]
Hashimoto et al.	The Use of Artificial Intelligence Tools to Prepare Medical School Applications	Discusses role of LLMs in improving medical school applications and the need for ethical guidelines in their use.	[42]

TABLE 1 (Continued)

Authors	Study title	Key takeaway	Ref
Munaf et al.	ChatGPT: A Helpful Tool for Resident Physicians?	Although requiring quality data and supervision, ChatGPT can aid resident physicians through data analysis and simulations.	[43]
Temsah et al.	Overview of Early ChatGPT's Presence in Medical Literature	Showcases benefits, ethical concerns and the need for human oversight in various health care applications of LLMs.	[44]

Note: Table presenting the key findings of the papers included in the study.
Abbreviations: AI, artificial intelligence; LLM, large language model.

FIGURE 2 Opportunities and challenges of large language models (LLMs) in medical education. The opportunities have a blue icon, and the challenges have a red icon. LLMs can serve as writing and personalised learning tools, synthesise and summarise large quantities of information, improve clinical decision support, provide instant access to medical knowledge for students and assist with course and assessment creation. However, challenges of LLMs remain, including issues with incorrect responses or fabrications, assurance of academic integrity, overreliance on technology and safeguarding of ethical, legal and privacy concerns, especially for patients. [Color figure can be viewed at wileyonlinelibrary.com]

question banks.[22] According to a study by Sng et al., ChatGPT demonstrated a systematic and concise ability to answer medical questions.[23] However, Wang et al. noted that although LLM-based medical dialogue systems show promise in tasks such as diagnosis, treatment recommendations and providing medical information, they have not yet achieved conversational capabilities equivalent to that of human interaction. LLMs do not have the function of replacing the in-person medical education provided by classes and clinical rotations.[24] Further, these technologies do not provide support for the goal of not simply training competent medical professionals but also compassionate and approachable physicians and educators.

LLMs like ChatGPT and GPT-4, nevertheless, are valuable resources for medical students, providing instant access to medical knowledge and research studies. Although their performance varies, further specialisation can enhance their capabilities. LLMs can aid in exam preparation, generate case reports, support clinical care evaluation and improve clinical reasoning by answering medical questions and generating hypotheses. However, in its current form, certain challenges remain in achieving human-level dialogue and interaction. It is important to evaluate their limitations and effectively integrate LLMs moving forward (Figure 2).

3.3 | Cases of biomedical LLMs

The creation of domain-specific biomedical language models (BioLLMs) has proven to be highly effective in enhancing medical natural language processing (NLP) tasks (Table 2). One notable BioLLM, BioBERT, has shown greater success in biomedical recognition when compared with general language models.[24] Taking inspiration from BioBERT's achievements, researchers have utilised medical text sources such as electronic health records (EHRs) and PubMed to

TABLE 2 Overview of LLMs.

LLMs	Developer	Summary/key applications
ChatGPT	Open AI	NLP and text generation
GPT 3, 3.5. 4	Open AI	NLP and text generation
Bard	Google	NLP and text generation
BioGPT	Luo et al.	Biomedical text generation and mining
BioBERT	Lee et al.	Biomedical text generation and mining
BioMedLM	Stanford (Bolton et al.)	Biomedical text generation and mining
ChatCAD	Wang et al.	Computer-aided diagnosis system by analysing medical images
ClinicalBERT	Huang et al.	Multibillion parameter language model built on clinical electronic health record data
BlueBERT	Peng et al.	Fine-tuned BERT model trained on PubMed abstracts and clinical notes
Clinical Camel	Toma et al.	Fine-tuned from Llama and trained on biomedical research
BioMegatron	Shin et al.	Trained on PubMed abstracts and research
BioMedRoBERTa	Gururangan et al.	Fine-tuned from RoBERTA and trained on biomedical research
MedAlpaca	Han et al.	Fine-tuned from Alpaca and trained on biomedical research
Med-PaLM	Google (Singhal et al.)	Trained on biomedical research

Note: This table showcases all of the LLMs included in this paper, their developer and their key applications. Note this is not an exhaustive list due to their myriad and rapid development.

develop tailored BioLLMs like ClinicalBERT, BioMegatron and BioMedRoBERTa. By fine-tuning these models with domain-specific data, they have been successful in processing medical language with exceptional results.[25] MedAlpaca by Han et al. is a specialised language model designed for further fine-tuning of LLMs specifically for medical applications.[26] MedAlpaca represents a significant step forward as a model purposefully built to optimise other LLMs for the medical domain with potential to be a great resource for medical education.

BioLLMs have shown promise, specifically, in medical education when properly fine-tuned. For instance, BioBERT improves research accessibility and summarisation, whereas models like BioMedLM, BioGPT, MedPaLM and ClinicalCamel, which are trained on medical question-answering datasets and clinical articles, aid in question answering.[27] Many of these BioLLMs can complete novel downstream medical tasks utilising few-shot prompting, use of a few explicit examples (or shots) to guide the AI to respond in a specific way.[27] Furthermore, ClinicalBERT and BlueBERT, fine-tuned on EHRs and biomedical text, exhibit enhanced clinical NLP performance, highlighting the value of in-domain tuning.[28] LLMs also hold promise for computer-aided diagnosis (CAD). Wang et al. developed ChatCAD, an LLM-based model that improves CAD outputs for diagnosis, reporting and information summarisation or reorganisation.[29] As a decision support tool, ChatCAD enables students to discuss cases and receive guidance on differential diagnoses and diagnostic tests, thus enhancing their clinical skills. Despite this progress, it remains important for educators on the front line to confirm accuracy prior to integration of these tools within the existing competency-based system.

To ensure the responsible use of LLMs, Desaire et al. have created a reliable method for distinguishing between text generated by humans and that generated by AI.[30] By utilising classification models, this method can help to detect inappropriate use of AI in academic and medical writing. When implemented thoughtfully, these models hold great promise as tools for improving medical education by providing research support, aiding clinical decision-making, offering personalised tutoring and more. However, it is equally crucial to develop responsible use models to ensure that medical education and clinical care are not inversely impacted as technology continues to evolve.

3.4 | Limitations

Although LLM may prove to be a robust educational tool, the use of LLM in medical education has limitations. These include incorrect responses, overreliance on technology, impact on critical thinking and academic integrity concerns (Figure 2).[31] Therefore, it is essential to verify the information provided by LLM and continue emphasising practical skills. Frequent updates to LLMs require monitoring to ensure accuracy. There are also concerns about the negative impact of overreliance on ChatGPT on clinical reasoning development and the lack of contextual understanding in its responses.[16] Additionally, experts have raised questions about redundancy and lack of original thought, which may affect students' critical thinking and reasoning skills.[32–35] Organisational guidelines have been revised because of ChatGPT's ability to generate papers.[36] Other risks include fabricated references and perpetuation of false information.[37–39] Some experts suggest halting advanced LLM training because of societal concerns.[40] More tempered recommendations include thorough testing, supervised trainee exploration and open discussions around responsible use.[41] It is crucial for medical students and resident physicians to gain a thorough understanding of ChatGPT's appropriate and effective utilisation, allowing application as a resource to overcome traditional challenges, even as early as medical school application preparation.[42,43] Although LLMs offer many benefits, responsible implementation through verification, emphasis on practical skills, monitoring for accuracy and mitigating risks remain vital in medical education.

3.5 | Ethical, legal and privacy concerns

The use of LLMs in medical education raises important ethical, legal and privacy concerns. One major concern is the potential for biases in

AI models. If the training data or algorithms themselves are biased, there is a risk that AI models may perpetuate biases and discrimination. This can lead to inaccurate or discriminatory information being provided to students and a lack of diversity and representation in medical education.[44] In February 2023, the CEO of Open AI, which created ChatGPT, acknowledged the presence of biases in LLMs.[45] These biases can have negative consequences for learning and patient care in the context of medical education.

Privacy and confidentiality are also important considerations when using LLMs, as sensitive information, including patient data, is shared in medical education settings. Patient clinical and research data are protected by the Health Insurance Portability and Accountability Act (HIPAA), and implementation of LLMs into medical education and clinical care must be certain that patient information and privacy cannot be breached. Institutions such as the University of California San Diego have thus aimed to stay ahead of the curve with creation of EHR-integrated LLMs, such as Dr. Chatbot, which drafts responses to patient EHR messages, to complement clinical care and ensure privacy is maintained.[46] Although addressing these concerns improves the safety and usability of LLMs for education or care involving patient information, addressing resistance from physicians to the technology climate change is an additional challenge.

The use of LLMs in projects or paper preparation, both in medical education and research, raises ethical questions and the need for clarity regarding appropriate methods of disclosure. Advances in the knowledge around these technologies will be improved with transparency regarding the use of LLMs. From a medical literature perspective, if the LLM significantly influenced the study, it should be reported. Further, minor tools such as spell checker use in paper preparation for full clarity should be reported but typically are not. Researchers should indicate the extent to which LLMs were involved in their projects or manuscript preparation as transparency is the best method to combat resistance to its utilisation. For project and paper preparation in classroom instruction, institutions and their educators must develop guidelines to assist students in how to appropriately utilise these resources. A lack of clarity regarding the restrictions of its use upfront could lead to habits that are counterproductive to learning and lack ethical grounds.

4 | DISCUSSION

The adoption of LLMs in medical education evokes a spectrum of reactions among educators, clinicians and learners, influenced by a variety of factors. This review critically addresses a significant gap in current research on LLMs in medical education by expanding beyond the predominant focus regarding potential misuses of LLMs in education to the exploration of their transformative potential. Although 57.5% of the studies primarily discuss the theoretical potential of LLMs, the rest provide quantitative insights and underscore a consistent theme that although LLMs show promise, they still fall short in certain aspects of medical knowledge and application compared with human performance. It is clear that LLMs and technology are not a replacement for the knowledge gained through education and experience.

Individuals who welcome LLMs often recognise their potential to address issues in medical education, such as the overload of medical curricula and everchanging knowledge provided through research. The wealth of information prohibits the expectation for students or clinicians to be experts in all fields of medicine. LLMs offer a way to enhance the learning process, focusing on the most crucial aspects necessary at each particular level of training. In addition, they could make medical knowledge more accessible providing widespread access to the latest medical information, thus reducing barriers and disparities in medical education across the world. This is especially beneficial to individuals with limited access to costly printed or online up-to-date resources, as they would be able to acquire medical information readily and at minimal cost.

LLMs have the potential to provide tailored explanations, practice questions and feedback that offers more personalised learning approaches tailored to each student's needs. With medical students grasping complex concepts more quickly in a way suited to their individual learning styles, the overall efficiency of the educational process will be enhanced. Furthermore, LLMs can have a positive impact on the development of clinical reasoning by engaging students in formulating, analysing and discussing real patient cases, providing a safe and interactive environment for students to develop critical thinking and decision-making skills.

However, resistance or fear towards LLMs stems from concerns about their potential to propagate misinformation, the ethical implications of their use and overall reliability. There is also apprehension about the use of LLMs leading to a reduced emphasis on critical thinking and decision-making skills in students. Further, there are concerns about data privacy and the depersonalization of education as well. It is quite likely that the degree of familiarity with these tools impacts comfort levels and understanding. Those with greater exposure to LLMs and AI methods are more understanding of the technology's current capabilities and limitations. LLMs' tendency to 'hallucinate' or produce factually incorrect information adds another element of anxiety among the public. Hallucinations are notably problematic in medical education, where accuracy is paramount, necessitating validation mechanisms and critical evaluation of their outputs. Integration of these technologies in a way that enhances education requires addressing these concerns in a way that is forward-thinking and collaborative across disciplines to transform educational strategies and ethical guidelines.

Although educators and learners recognise their potential to revolutionise medical education, there remains apprehension about reliability and impact on critical thinking skills of students. To mitigate these concerns, it is vital to develop mechanisms for verifying information accuracy and integrate LLMs in a way that complements traditional teaching methods, rather than replacing them. To truly harness benefits of LLMs in medical education, it is crucial that we implement them responsibly, which includes rigorous oversight to verify the accuracy of information provided by LLMs. Accuracy is paramount in clinical care; thus, the future generations of clinicians must be

provided with reliable tools to enhance their learning. We would contend that the integration of LLMs into medical education provides opportunities to enhance curricula rather than detract from it. LLMs and the utilisation of available technologies provide opportunities for students to focus less on memorisation and building of mnemonics for test-taking, but rather on understanding, logical thinking, problem-solving and the application of concepts into clinical settings.

Importantly, integration of LLMs must continue to ensure patient privacy and confidentiality within the health care setting. These priorities must be emphasised with learners first and foremost before any usage of these technologies is undertaken. Collaboration between technologists and medical experts to align priorities and ensure that the development of these tools reflects the actual needs and challenges of medical education is vital. The medical education community can enhance learning experiences while mitigating risks and challenges of these technologies by understanding the factors that influence individual's attitudes towards LLMs. This discussion has aimed to address these nuanced perspectives, challenge negative perceptions and advocate for responsible and informed integration of the technologies.

5 | CONCLUSION

LLMs hold immense potential to transform medical education by offering innovative solutions for personalised learning, intelligent tutoring, content generation and clinical decision support. The applications of these models in medical education can enhance access to knowledge and support self-directed learning. However, important aspects such as ethical considerations, content quality, privacy concerns and long-term impact of LLMs on learning outcomes and clinical practice must be carefully addressed. These technologies are expected to only continue to gain acceptance, grow in applicability and improve in accuracy. Embracing the opportunities presented by LLMs while being mindful of their challenges will allow the medical education system to become more efficient, effective and learner-centred. Understanding the perspectives and concerns of all stakeholders involved, including educators, professionals, students and patients, will be crucial in shaping the future of medical education and harnessing the full potential of LLMs.

OTHER DISCLOSURES

Although authorship was fully performed by the listed individuals with the creation of the original draft, revisions and final copy, ChatGPT was utilised for outline preparation and ensuring accuracy in grammar.

AUTHOR CONTRIBUTIONS

Harrison C. Lucas: Conceptualization; investigation; validation; methodology; data curation; writing - original draft; visualization. **Jeffrey S. Upperman:** Supervision; methodology; conceptualization; validation; writing - review and editing. **Jamie R. Robinson:** Writing - review and editing; methodology; supervision; project administration; visualization; validation.

ACKNOWLEDGEMENTS

[Correction added on 02 MAY 2024, after first online publication: Reference citations 23 onwards were corrected].

CONFLICT OF INTEREST STATEMENT

None.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ETHICS STATEMENT

Not applicable.

ORCID

Jamie R. Robinson https://orcid.org/0000-0003-0888-0156

REFERENCES

1. GPT-4 [Internet]. [cited 2024 Mar 19]. Available from: https://openai.com/research/gpt-4
2. Gemini - chat to supercharge your ideas [Internet]. Gemini [cited 2024 Mar 19]. Available from: https://gemini.google.com
3. Buja LM. Medical education today: all that glitters is not gold. *BMC Med Educ*. 2019;19(1):110. doi:10.1186/s12909-019-1535-9
4. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
5. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmol Sci*. 2023;3(4):100324. doi:10.1016/j.xops.2023.100324
6. 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*. 2023;13(1):20512. doi:10.1038/s41598-023-46995-z
7. Long C, Lowe K, dos Santos A, et al. Evaluating ChatGPT-4 in otolaryngology–head and neck surgery board examination using the CVSA model [internet]. *medRxiv*. 2023 [cited 2024 Mar 19]. p. 2023.05.30.23290758. doi:10.1101/2023.05.30.23290758v1
8. 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*. 2023;104(5):269-273. doi:10.4174/astr.2023.104.5.269
9. Skalidis I, Cagnina A, Luangphiphat W, et al. ChatGPT takes on the European exam in core cardiology: an artificial intelligence success story? *Eur Heart J Digit Health*. 2023;4(3):279-281. doi:10.1093/ehjdh/ztad029
10. Humar P, Asaad M, Bengur FB, Nguyen V. ChatGPT is equivalent to first-year plastic surgery residents: evaluation of ChatGPT on the plastic surgery in-service examination. *Aesthet Surg J*. 2023;43(12):NP1085-NP1089. doi:10.1093/asj/sjad130
11. Sabry Abdel-Messih M, Kamel Boulos MN. ChatGPT in clinical toxicology. *JMIR Med Educ*. 2023;8(9):e46876. doi:10.2196/46876
12. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *Neurosurgery*. 2023;93(5):1090-1098. doi:10.1227/neu.0000000000002551
13. Webb JJ. Proof of concept: using ChatGPT to teach emergency physicians how to break bad news. *Cureus*. 2023;15(5):e38755. doi:10.7759/cureus.38755

14. Huang Y, Gomaa A, Semrau S, et al. Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: potentials and challenges for ai-assisted medical education and decision making in radiation oncology. *Front Oncol*. 2023;13:1265024. doi:10.3389/fonc.2023.1265024
15. 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
16. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med Educ*. 2023 Jun;1(9):e48291. doi:10.2196/48291
17. Gandhi Periaysamy A, Satapathy P, Neyazi A, Padhi BK. ChatGPT: roles and boundaries of the new artificial intelligence tool in medical education and health research - correspondence. *Ann Med Surg (Lond)*. 2023;85(4):1317-1318. doi:10.1097/MS9.0000000000000371
18. Liu X, McDuff D, Kovacs G et al. Large Language Models are Few-Shot Health Learners [Internet]. arXiv; 2023 [cited 2024 Mar 19]. Available from: http://arxiv.org/abs/2305.15525
19. Parsa A, Ebrahimzadeh MH. ChatGPT in medicine; a disruptive innovation or just one step forward? *Arch Bone Jt Surg*. 2023;11(4): 225-226. doi:10.22038/abjs.2023.22042
20. Shorey S, Mattar C, Pereira TLB, Choolani M. A scoping review of ChatGPT's role in healthcare education and research. *Nurse Educ Today*. 2024;135:106121. doi:10.1016/j.nedt.2024.106121
21. Ahn S. The impending impacts of large language models on medical education. *Korean J Med Educ*. 2023;35(1):103-107. doi:10.3946/kjme.2023.253
22. Singh R, Reardon T, Srinivasan VM, Gottfried O, Bydon M, Lawton MT. Implications and future directions of ChatGPT utilization in neurosurgery. *J Neurosurg*. 2023;139(5):1487-1489. doi:10.3171/2023.3.JNS23555
23. Sng GGR, Tung JYM, Lim DYZ, Bee YM. Potential and pitfalls of ChatGPT and natural-language artificial intelligence models for diabetes education. *Diabetes Care*. 2023;46(5):e103-e105. doi:10.2337/dc23-0197
24. Wang DQ, Feng LY, Ye JG, Zou JG, Zheng YF. Accelerating the integration of ChatGPT and other large-scale AI models into biomedical research and healthcare. *MedComm – Future Medicine*. 2023;2(2):e43. doi:10.1002/mef2.43
25. Qiu J, Li L, Sun J, et al. Large AI models in health informatics: applications, challenges, and the future. *IEEE J Biomed Health Inform*. 2023; 27(12):6074-6087. doi:10.1109/JBHI.2023.3316750
26. Han T, Adams LC, Papaioannou JM et al. MedAlpaca -- An Open-Source Collection of Medical Conversational AI Models and Training Data [Internet]. arXiv; 2023 [cited 2024 Mar 19]. Available from: http://arxiv.org/abs/2304.08247
27. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Brief Bioinform* 2023;25(1):bbad493.
28. Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. *Cureus*. 2023 May; 15(5):e39305. doi:10.7759/cureus.39305
29. Wang S, Zhao Z, Ouyang X, Wang Q, Shen D. ChatCAD: Interactive Computer-Aided Diagnosis on Medical Image using Large Language Models [Internet]. arXiv; 2023 [cited 2024 Mar 19]. Available from: http://arxiv.org/abs/2302.07257
30. Desaire H, Chua AE, Isom M, Jarosova R, Hua D. ChatGPT or academic scientist? Distinguishing authorship with over 99% accuracy using off-the-shelf machine learning tools [Internet]. arXiv; 2023 [cited 2024 Mar 19]. Available from: http://arxiv.org/abs/2303.16352
31. Komorowski M, Del Pilar Arias López M, Chang AC. How could ChatGPT impact my practice as an intensivist? An overview of potential applications, risks and limitations. *Intensive Care Med*. 2023;49(7): 844-847. doi:10.1007/s00134-023-07096-7
32. Lee H. The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ*. 2023. doi:10.1002/ase.2270
33. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: is ChatGPT a blessing or blight in disguise? *Med Educ Online*. 2023;28(1):2181052. doi:10.1080/10872981.2023.2181052
34. Huh S. Can we trust AI chatbots' answers about disease diagnosis and patient care? *J Korean Med Assoc*. 2023;10(66):218-222.
35. Gunawardene AN, Schmuter G. Teaching the limitations of large language models in medical school. *J Surg Educ*. 2024;S1931-7204(24): 00049-7. doi:10.1016/j.jsurg.2024.01.008
36. Loh E. ChatGPT and generative AI chatbots: challenges and opportunities for science, medicine and medical leaders. *BMJ Lead*. 2023. leader-2023-000797
37. Gravel J, D'Amours-Gravel M, Osmanlliu E. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. *Mayo Clinic Proc Digital Health*. 2023;1(3):226-234. doi: 10.1016/j.mcpdig.2023.05.004
38. Arachchige PM, AS. Large language models (LLM) and ChatGPT: a medical student perspective. *Eur J Nucl Med Mol Imaging*. 2023;50(8): 2248-2249. doi:10.1007/s00259-023-06227-y
39. De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health*. 2023;11:1166120. doi:10.3389/fpubh.2023.1166120
40. Pause Giant AI Experiments: An Open Letter [Internet]. Future of Life Institute. [cited 2024 Mar 20]. Available from: https://futureoflife.org/open-letter/pause-giant-ai-experiments/
41. Bair H, Norden J. Large language models and their implications on medical education. *Acad Med*. 2023;98(8):869-870. doi:10.1097/ACM.0000000000005265
42. Hashimoto DA, Johnson KB. The use of artificial intelligence tools to prepare medical school applications. *Acad Med*. 2023;98(9):978-982. doi:10.1097/ACM.0000000000005309
43. Munaf U, Ul-Haque I, Arif TB. ChatGPT: a helpful tool for resident physicians? *Acad Med*. 2023;98(8):868-869. doi:10.1097/ACM.0000000000005250
44. 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*. 2023;15(4):e37281. doi:10.7759/cureus.37281
45. Chowdhury H. Sam Altman has one big problem to solve before ChatGPT can generate big cash — making it "woke" [Internet]. Business Insider. [cited 2024 Mar 19]. Available from: https://www.businessinsider.com/sam-altmans-chatgpt-has-a-bias-problem-that-could-get-it-canceled-2023-2
46. Introducing Dr. Chatbot [Internet]. [cited 2023 Jul 25]. Available from: https://today.ucsd.edu/story/introducing-dr-chatbot

How to cite this article: Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Med Educ*. 2024;58(11): 1276-1285. doi:10.1111/medu.15402