Commentary

Integrating large language models into medical education: A commentary on opportunities, challenges, and future directions

Leo Morjaria[a], Levi Burns[a], Keyna Bracken[a,b], Anthony J Levinson[a], Quang N Ngo[a,b], Mark Lee[b], Matthew Sibbald[a,b]

a. Michael G. DeGroote School of Medicine, McMaster University, Hamilton, Ontario, Canada
b. McMaster Education Research, Innovation and Theory (MERIT) Program, McMaster University, Hamilton, Ontario, Canada

Abstract

Shortly following its release in November 2022, OpenAI's ChatGPT gained notoriety in the medical education community for its ability to perform at or near the passing threshold on the United States Medical Licensing Examination (USMLE). Although there is an overwhelming amount of excitement surrounding artificial intelligence (AI) and its potential to revolutionize medical training, this commentary seeks to explore both the opportunities and challenges posed by incorporating large language models (LLMs) such as ChatGPT into medical education. To evaluate ChatGPT's impact in the context of problem-based learning (PBL) medical education, we conducted two initial studies. The first study assessed ChatGPT's performance on concept application exercises (CAEs)—short-answer assessments used in our program to gauge student progress. After establishing ChatGPT's performance on CAEs, our second study aimed to evaluate ChatGPT's ability to effectively grade student-generated responses. Our results reveal that ChatGPT not only outperforms students who are marginally passing but also grades these assessments with promising alignment to human grading practices. Our team's future research plans include examining ChatGPT's ability to provide effective feedback, generate discerning assessment questions, create realistic training scenarios, and support continuous professional development. Although we are optimistic about future applications of LLMs, we emphasize the need for an AI-assisted approach that employs human oversight to mitigate the inherent risks associated with LLMs, such as bias perpetuation, inaccuracies, over-reliance, and potential misuse. Ultimately, through the thoughtful and evidence-based implementation of these new tools, we believe AI can be harnessed to augment rather than undermine the quality and effectiveness of medical education.

Keywords: ChatGPT, Medical Education, Artificial Intelligence, Large Language Models

Corresponding author: Dr. Matthew Sibbald, matthew.sibbald@medportal.ca

Introduction

The integration of artificial intelligence (AI) in medicine is rapidly expanding in scope from clinical practice and research to transforming medical education (1,2). In November 2022, OpenAI released its flagship large language model (LLM), ChatGPT, to the public. Within weeks of its release, ChatGPT gained notoriety in medical education circles for performing at or near the passing threshold for the United States Medical Licensing Examination (USMLE) (3,4). Since then, ChatGPT has continued to demonstrate a remarkable capacity to interpret complex prompts and respond to users in a human-like fashion. As a result, ChatGPT has drawn tremendous interest from the international medical education research community to explore the potential of these new LLMs as a transformative tool in medical education (5-7).

Incorporating AI into teaching and learning environments offers a wide variety of exciting opportunities to enhance educational practice (6-8). Example uses for AI include facilitating access to medical literature, enabling real-time analysis of student responses, and augmenting simulation exercises (5,9). Particularly of interest to our team is the potential for this new technology to enhance problem-based learning by providing students with the possibility for self-directed teaching and feedback (10). However, the rapid pace of implementation of these technologies makes it challenging to ensure that their use is both thoughtful and effective (1,2). There has also been discourse surrounding whether ChatGPT threatens the validity of traditional medical assessment, especially in the context of unproctored, online examinations which have gained popularity in the post-pandemic era (11).

Ultimately, as ChatGPT and similar LLMs continue to be developed, their integration into medical education prompts a reevaluation of pedagogical strategies, assessment methodologies, and the overall educational experience. The potential of LLMs to augment learning and assessment processes in medical education is promising but necessitates frank discussions and considerations with regards to risks, challenges and weaknesses associated with these tools. To evaluate if LLMs could be helpful in grading, we first decided to establish how well these new technologies performed on assessments.

Our research

For these new and innovative AI tools to be incorporated thoughtfully into our program, it is critical that their implementation be grounded in evidence. Currently, there is limited research on the impact of ChatGPT on problem-based learning medical education (10). One area of interest to our research team is the application of ChatGPT and other LLMs in assessing medical student performance. We have been particularly interested in the impact of LLMs on short-answer assessments, as these assessments are crucial for gauging a wide range of skills, including communication, reasoning, and analysis. At our institution, short-answer assessments in the form of concept application exercises (CAEs) are a foundational component of pre-clerkship student assessment (12). Unlike multiple choice exams, grading short answer assessments is resource-intensive, poses logistical challenges — such as delays between time of assessment and feedback

— and involves significant faculty time costs. Using LLMs to evaluate short-answer responses offers an interesting possible solution to streamline and accelerate the feedback process, allowing educators to allocate more time to student support and interaction.

Our first study evaluated the extent to which student access to ChatGPT could affect the validity of CAEs in measuring medical student knowledge (13). We compiled forty past CAE problems and compared the quality of ChatGPT-generated responses to past responses from pre-clerkship students which had earned scores of 3 out of 5 by their respective pre-clerkship tutors. A panel of experienced pre-clerkship tutors scored each response, blinded to whether the responses were produced by ChatGPT or past students. ChatGPT-generated responses scored a mean of 3.29 out of 5 compared to 2.38 for student-generated responses to the same prompts. Our study suggests that ChatGPT can outperform students who are marginally passing (i.e., receiving scores of 3 out of 5 in tutorial settings). However, ChatGPT-generated responses did not reach the performance level of the broader student population, as indicated by the historical class average of 3.67 to the CAE problems that were compiled.

After establishing ChatGPT's performance on the CAE, our second study aimed to validate the utility of ChatGPT in grading student-generated responses (14). We used sixty historical student responses to ten historical CAE prompts from the past five years. We then used ChatGPT to assign a score to these responses under four distinct conditions: with both a rubric and standard provided to ChatGPT, with only a standard, with only a rubric, and with neither supplemental resource. It is important to note that at our institution, rubrics are specific to each CAE question, while the standard is consistent across all CAE problems. These four conditions were used to reflect the grading materials that are provided to tutors to standardize CAE grading across human evaluators. Statistical analysis revealed correlations ranging from 0.6 to 0.7 ($p < 0.001$) between ChatGPT-assigned scores and human-assigned scores across the four conditions. Interestingly, the absence of a rubric resulted in systematically higher scores assigned by ChatGPT, highlighting the importance of clear grading criteria. Furthermore, we found that while ChatGPT's scoring may not always align perfectly with human assessors, deviations of more than 1 point between AI and human assigned scores were far less common. Ultimately, the results of our study suggest that there is a good degree of alignment between AI and human grading practices.

Together, our aim with these studies was to provide a comprehensive look at the potential and challenges of effectively and thoughtfully applying ChatGPT to short answer medical education assessments. While ChatGPT demonstrates considerable promise in helping students prepare for assessments, the possible use of this technology during unproctored assessments raises significant concerns regarding the validity of assessments as well as increased difficulty accurately identifying students needing additional support. Additionally, there are numerous pragmatic challenges with respect to more systematic integration of LLMs into grading and testing environments. These challenges include selecting the appropriate LLM for the specified task, managing both the initial setup costs as well as ongoing expenses, and addressing technical implementation difficulties of integrating the LLM within existing systems.

Ultimately, our findings underscore the importance of further research into optimizing AI-assisted assessment techniques and carefully considering their integration into competency-based education frameworks.

Future directions

Our research team is further exploring ChatGPT's capabilities in the context of the CAE, examining its ability to provide written feedback and generate new CAE questions. This next phase aims to critically examine the quality, relevance, consistency, and educational value of ChatGPT-generated feedback and assessment prompts and explore their alignment with the core competencies and learning objectives of our medical program. Al generated narrative feedback is particularly interesting as early work suggests that students often perceive some feedback as ineffective and hollow due to its brevity and lack of specificity (15,16). Moreover, our investigation will extend beyond the CAE to evaluate ChatGPT's utility within progress testing at our institution.

Additional future directions of research include the potential role of ChatGPT in student self-assessment, simulation scenarios, facilitating problem-based learning discussions, and supporting continuous professional development activities. Our exploratory research has been and will continue to be guided by the principle that conducting exploratory studies across diverse areas is essential to ensuring that the integration of a new technology into our problem-based learning program is both thoughtful and evidence-based.

Further considerations

It is important that enthusiasm for these technologies is tempered by an awareness of their varied limitations and potential risks (1,2). Firstly, despite their advanced capabilities, LLMs create output through complex probabilistic models (5). This lack of genuine contextual or semantic understanding can potentially lead to inaccuracies in complex medical discussions (5). Bias is another critical concern: as LLM training methods are often proprietary, these technologies may mislead students by perpetuating existing biases present in their training data (17). Thirdly, while there have been varied levels of concern regarding the impact of ChatGPT and similar models on academic integrity, potential misuse by students may undermine learning outcomes and hinder student progress. Finally, although often not discussed, there are significant greenhouse emission implications posed by LLMs as well as concerns of inequitable access (18).

Ensuring that the deployment of AI enhances rather than detracts from the educational experience requires careful evaluation and oversight to maintain the quality and integrity of medical training. It is for these reasons that we believe that deployment of these new technologies should follow an AI-assisted approach, meaning that AI should enhance rather than replace student and faculty contributions. In the case of assessment grading, we believe a combination of LLM grading with faculty supervision represents the most prudent approach.

This will ensure that human oversight is embedded into each of these applications by design and force accountability for model outputs to remain with human medical educators.

Conclusion

The integration of LLMs into medical education presents opportunities to enhance learning and assessment methodologies, provided that their limitations and risks are carefully considered. Through our research, we aim to underscore the importance of cautious optimism and rigorous evaluation in the deployment of these new technologies. Moving forward, we hope to encourage a collaborative approach involving educators, students, and technologists to navigate the ethical, academic, and practical challenges of integrating AI into education. In a technology-enabled world, it is our view that medical curriculum and assessment must be adapted to reflect the environment that future healthcare practitioners will be working in. As calls to integrate AI into medical education continue to build (19,20), and as an increasing number of Canadian medical programs incorporate this technology (21,22), it is crucial to take a balanced approach to ensure that AI augments, rather than undermines, the quality and effectiveness of medical education.

References

1. Wartman, Steven. Reimagining Medical Education in the Age of AI. AMA J Ethics. 2019;21(2):E146-152. doi:10.1001/amajethics.2019.146
2. Masters K. Artificial intelligence in medical education. Med Teach. 2019;41(9):976-980. doi:10.1080/0142159X.2019.1595557
3. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. Dagan A, ed. PLOS Digit Health. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
4. Gilson A, Safranek CW, Huang T, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312. doi:10.2196/45312
5. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. Pak J Med Sci. 2023;39(2). doi:10.12669/pjms.39.2.7653
6. Lee H. The rise of ChatGPT : Exploring its potential in medical education. Anat Sci Educ. Published online March 28, 2023:ase.2270. doi:10.1002/ase.2270
7. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare. 2023;11(6):887. doi:10.3390/healthcare11060887
8. Fischetti C, Bhatter P, Frisch E, et al. The Evolving Importance of Artificial Intelligence and Radiology in Medical Trainee Education. Acad Radiol. 2022;29:S70-S75. doi:10.1016/j.acra.2021.03.023

9. Winkler-Schwartz A, Bissonnette V, Mirchi N, et al. Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation. J Surg Educ. 2019;76(6):1681-1690. doi:10.1016/j.jsurg.2019.05.015

10. Divito CB, Katchikian BM, Gruenwald JE, Burgoon JM. The tools of the future are the challenges of today: The use of ChatGPT in problem-based learning medical education. Med Teach. 2023;46:320-322. doi:10.1080/0142159X.2023.2290997

11. Kumah-Crystal Y, Mankowitz S, Embi P, Lehmann CU. ChatGPT and the clinical informatics board examination: the end of unproctored maintenance of certification? J Am Med Inform Assoc. 2023;30(9):1558-1560. doi:10.1093/jamia/ocad104

12. Neville AJ, Cunnington J, Norman GR. Development of clinical reasoning exercises in a problem-based curriculum: Acad Med. 1996;71(1):S105-7. doi:10.1097/00001888-199601000-00058

13. Morjaria L, Burns L, Bracken K, et al. Examining the Threat of ChatGPT to the Validity of Short Answer Assessments in an Undergraduate Medical Program. J Med Educ Curric Dev. 2023;10:23821205231204178. doi:10.1177/23821205231204178

14. Morjaria L, Burns L, Bracken K, et al. Examining the Efficacy of ChatGPT in Marking Short-Answer Assessments in an Undergraduate Medical Program. Int Med Educ. 2024;3(1):32-43. doi:10.3390/ime3010004

15. Urquhart L, Rees C, Ker J. Making sense of feedback experiences: a multi-school study of medical students' narratives. Med Educ.:Feb 2014. doi:10.1111/medu.12304

16. Bowen L, Marshall M, Murdoch-Eaton D. Medical Student Perceptions of Feedback and Feedback Behaviors Within the Context of the "Educational Alliance." Acad Med J Assoc Am Med Coll. Published online September 1, 2017.

17. Gallegos IO, Rossi RA, Barrow J, et al. Bias and Fairness in Large Language Models: A Survey. Published online 2023. doi:10.48550/ARXIV.2309.00770

18. Li P, Yang J, Islam MA, Ren S. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. Published online 2023. doi:10.48550/ARXIV.2304.03271

19. Lee J, Wu AS, Li D, Kulasegaram K (Mahan). Artificial Intelligence in Undergraduate Medical Education: A Scoping Review. Acad Med. 2021;96(11S):S62-S70. doi:10.1097/ACM.0000000000004291

20. Pucchio A, Rathagirishnan R, Caton N, et al. Exploration of exposure to artificial intelligence in undergraduate medical education: a Canadian cross-sectional mixed-methods study. BMC Med Educ. 2022;22(1):815. doi:10.1186/s12909-022-03896-5

21. Hu R, Fan KY, Pandey P, et al. Insights from teaching artificial intelligence to medical students in Canada. Commun Med. 2022;2(1):63. doi:10.1038/s43856-022-00125-4

22. Walker D, Lee S, Kaka H, Campbell C, Sharma N. Preparing medical students for artificial intelligence: a survey of educational needs and impact of a lecture on perceptions. 2022. doi:10.36834/cmej.75002