Canadian Medical Education Journal

Commentary and Opinions

Artificial Intelligence can transform formative assessment in medical education

L'intelligence artificielle peut transformer l'évaluation formative dans l'enseignement médical

Joshua Feldman,[1] Christopher Gilchrist,[1] Fok-Han Leung[1,2]

[1]MD Program, Temerty Faculty of Medicine, University of Toronto, Ontario, Canada; [2]Department of Family and Community Medicine, Unity Health Toronto, Ontario, Canada

Correspondence to: Joshua Feldman; email: joshua.feldman@mail.utoronto.ca

Published ahead of issue: Sep 18, 2024; published: Dec 22, 2025. CMEJ 2025, 16(6) Available at https://doi.org/10.36834/cmej.79759

In the University of Toronto's M.D. program, I (JF) was surprised to find that we are not told answers to questions we get wrong on tests. Assessments are the safest place for mistakes, yet the medical school sadly cannot take advantage of this chance to provide valuable feedback. By getting specific and timely feedback, it would help me identify and correct gaps in my knowledge so I can one day deliver better care to patients. Answers have not been released for educational purposes because generating new tests is too resource intensive and the worry that, by focusing on the minutiae of incorrect answers, students will miss the forest for the trees. With recent advances in generative artificial intelligence (AI), this situation has changed. Medical schools should iteratively implement and evaluate these technologies to alleviate the bottleneck of writing assessments, and even find new creative ways of providing feedback to students.

To offer a faculty perspective (CG, FHL), assessments must be written by physicians to ensure that they are accurate, fair, comprehensive, and at the appropriate level. A 40-question test quickly becomes a time-consuming affair, especially in an overwhelmed health system where time for medical education is scarce. Generating new tests each year is simply not possible. Thus, despite commitments to academic honesty, if answers were released and tests remained unchanged, solutions would likely find a way to students the following year. Assessment can and should serve as opportunities for feedback and growth. This current approach leaves room for improvement.

We all believe that generative AI can improve this situation. Generative AI algorithms are computer programs that create content that resembles content produced by people.[1] Generative AI can lessen the burden of writing medical exams by being instructed to write questions. There is preliminary evidence that this is possible.[2,3] By automating the process of writing assessments, clinician effort would be reduced to checking the questions and answers written by AI, allowing for new tests each year. Creative approaches such as automatically generating written summaries of where students should focus their learning could also be explored. Feedback following tests could then be given without fear of future academic dishonesty or excessive workload for faculty.

In practice, reality is more complicated. AI is vulnerable to hallucinating inaccuracies, behaving unexpectedly, and reinforcing injustice.[4] In the context of medical education assessment, this could result in tests that evaluate incorrect knowledge, include questions that do not meet academic standards, or exacerbate health inequities. Careful evaluation is necessary and clinician-computer collaboration will be needed, as indicated above.

From our experiences as a medical student who worked as a data scientist for several years and as medical education researchers, we believe this evaluation should take an iterative approach based in quality improvement. When introducing a new technology, it is a best practice to develop tools with short cycles of development and evaluation rather than executing a complex preconceived plan.[5] This is often called “agile” software development. Quality improvement in medicine has a similar ethos. Innovation is highly uncertain with many unknown unknowns, which makes stating well-defined questions and methodologies in advance challenging, as is required for traditional research. This evaluation of using AI to improve assessment feedback would still be based on the principles of scholarship and empiricism, but it should use methods from quality improvement rather than research to explore these questions.

Students are open to medical schools trying out new approaches to assessment without always having the rigor of experimental evidence. If medical education only implemented what was in the literature, learners would miss what could be learned from innovation at a local level. Students celebrate when programs try out new approaches, even when it doesn’t go perfectly at first, because it demonstrates that faculty are trying to make their education better. Ideally, these efforts include students in their design and evaluation. Students would be excited to collaborate with faculty to use technology to make assessment into a learning opportunity.

We believe generative AI can transform feedback in medical education. Medical schools should quickly work to responsibly adopt these technologies. Quality improvement scholarship represents an ideal modality to accomplish this aim. If this change is made, we can correct mistakes early, which will hopefully lead to better patient outcomes by better trained physicians.

Funding: None.
Conflicts of Interest: None.
Edited by: Marcel D’Eon (editor-in-chief)

References

1. Zewe A. MIT News. *Explained: generative AI. How do powerful generative AI systems like ChatGPT work, and what makes them different from other types of artificial intelligence?* Available from: https://news.mit.edu/2023/explained-generative-ai-1109. [Accessed Jul 11, 2024].
2. Xu Y, Jiang Z, Ting DSW, et al. Medical education and physician training in the era of artificial intelligence. *Singapore Med J* 2024 Mar 1;65(3):159–66. https://doi.org/10.4103/singaporemedj.SMJ-2023-203
3. Sauder M, Tritsch T, Rajput V, Schwartz G, Shoja MM. Exploring generative artificial intelligence-assisted medical education: assessing case-based learning for medical students*. Cureus*. 2024 Jan 9;16(1). https://doi.org/10.7759/cureus.51961
4. Kaddour J, Harris J, Mozes M, Bradley H, Raileanu R, McHardy R. Challenges and applications of large language models. arXiv.org. 2023. Available from: https://arxiv.org/abs/2307.10169;
5. Forsgren N, Humble J, Kim G. *Accelerate: the science of lean software and DevOps building and scaling high performing technology organizations.* 1st ed. IT Revolution Press; 2018.