GPT-4: the future of artificial intelligence in medical school assessments

Lois Haruna-Cooper and Mohammed Ahmed Rashid
UCL Medical School, Faculty of Medical Sciences, University College London, London, WC1E 6BT, UK
Corresponding author: Lois Haruna-Cooper. Email: l.haruna@ucl.ac.uk

In recent years, artificial intelligence (AI) has been rapidly advancing in many fields, including medicine. One AI tool that has gained particular attention in the medical community is Chat-GPT, a large language model developed by OpenAI, which was first launched in November 2022, and its updated version, GPT-4, which was launched in March and which promises to be even more powerful and accurate in its capabilities. GPT-4 is a powerful tool that can generate human-like responses to a wide range of questions, prompts and images, and create academic articles, such as the one you are reading now. Using Chat-GPT to develop early iterations of this article and its title, some of which are shared in Appendix 1, highlights its potential to shift how we assess and prompt a careful reconsideration of the potential role of AI in medical education.

Medical school assessments are crucial to ensuring that newly graduated doctors are competent to provide safe and effective care and at an appropriate level to pursue postgraduate medical training. Traditionally, medical school examinations have been packaged into programmes and systems of assessment of written, practical and workplace-based assessments, overseen and graded by human examiners. However, the use of AI tools like Chat-GPT can provide new opportunities for medical school assessments.

The disruption caused by COVID-19 caused a shift in the delivery of medical education, with a review of how teaching is delivered, and many assessments being adapted to online versions. This led to increased innovation in assessments and sparked discussion about the potential impact of cheating and the ethics surrounding the increased use of technology in assessment.1 The idea of cheating forces greater interrogation of the purposes of assessment in a professional education programme like medicine and how students view their high-stakes exams.

Although workplace-based assessments have received great attention and priority in recent years because of their authenticity to ‘real-world’ clinical practice, students often focus on their performance on written and practical assessments. This may reflect that these occur less frequently than workplace-based assessments, and therefore there is less opportunity to view written and practical exams as assessments for learning, rather than assessments of learning.2 This view may influence how students behave in their assessments, as they perform to satisfy their assessor, rather than patient-centred care being the focus.3 With GPT-4’s ability to analyse diverse data, we must consider its potential to become an assessor, and how this could radically shift how students view and interact with assessment.

While AI may not have all the answers, it can be utilised as a tool to enhance current assessment models. If we consider the breadth of output that the current GPT-4 prototype offers, we can begin to imagine the potential it could bring to medical school assessments. Below are some examples of how it can be utilised in current assessment models:

- GPT-4 has the ability to generate questions on a wide range of topics and can adapt to different levels of difficulty; it can be utilised to write new multiple-choice questions and even scenarios for clinical exams.
- Another potential use of GPT-4 is as a tool for grading exams: it can be trained on a large dataset to identify key features that indicate a strong or weak response in written assessment.
- A recent study has shown that GPT-4 was capable of performing at or near the passing threshold for all three parts of the United States Medical Licensing Exam.4 This demonstrates first its potential to assist with clinical decision-making, but also forces us to consider the utility of
exams, such as USMLE, if it is possible to pass them without the need of human reasoning.

- There is also the potential to train GPT-4 to identify common mistakes in student responses and, integrated with deep fake technology, can create simulated patients and doctors to provide feedback to students on their performance.

- By using GPT-4 as the assessor, medical schools can reduce the workload on human examiners and ensure that feedback is consistent and of high quality, regardless of the workload of individual examiners.

It is easy to recognise the various ways GPT-4 can be used to enhance assessment, but there are also many potential risks. There is currently limited information about the accuracy of GPT-4 to assess practical skills and clinical knowledge, especially when applied to real-life situations. Another risk is that GPT-4 may not be able to fully understand the nuances of medical language, which could lead to errors in grading and feedback. One of the greatest challenges of using GPT-4, and AI more generally, is that it is only as unbiased as the data put into it. With the increased understanding we currently have of inequities in medical education and healthcare, GPT-4 could potentially perpetuate these biases leading to greater gaps in health and educational outcomes.

With Generation Z growing up in the social media age, where technology affords them instant access and rapid responses, GPT-4 has the potential to enhance current assessments by providing instant feedback, personalising learning experiences, improving objectivity and consistency of assessments, and reducing the workload of educators. However, in order to effectively address pre-existing concerns, and use assessment to truly enhance students’ learning, AI must be considered in the broader context of the constructive alignment of learning outcomes, teaching and assessment. Great attention must also be given to the possible unintended consequences that may only become apparent as AI finds its way into education practices. The use of AI in medical education is rapidly growing, and to enable us to take full advantage of all of its possible benefits, further research into how we can embrace and safely utilise to enhance the learning experience of students is needed.

Declarations
Competing Interests: None declared.

Funding: None declared.

Ethics approval: Not applicable as no personal data is used in this paper.

Guarantor: MAR, LHC.

Contributors: LHC conceived and wrote the article. MAR provided critical feedback.

Provenance: Not commissioned; editorial review.

ORCID iDs: Ahmed Rashid https://orcid.org/0000-0002-8443-1240
Lois Haruna-Cooper https://orcid.org/0000-0002-9729-0125

References