Canadian Medical Education Journal

Commentary and Opinions

Large language models in medical education: new tools for experimentation and discovery

Les grands modèles de langage dans l'enseignement médical : de nouveaux outils pour l'expérimentation et la découverte

Akshay Rajaram[1]

[1]Departments of Family and Emergency Medicine, Queen's University, Ontario, Canada

Correspondence to: Akshay Rajaram, MMI MD CCFP(EM), 809-121 Queen Street, Kingston, ON K7K 0G6; email: akshay.rajaram@queensu.ca

Published ahead of issue: Apr 29, 2024; published: Jul 12, 2024. CMEJ 2024, 15(3) Available at https://doi.org/10.36834/cmej.78879

Introduction

More than 20 years ago, the latest technology was making waves around the world – the Internet. As Le and Stein wrote in their viewpoint, it was "an exciting time for experimentation and discovery."[1] Today, we are in another period of experimentation and discovery with generative artificial intelligence (AI). As the Internet improved connectivity and access to information, generative AI – specifically large language models (LLMs) – are unlocking new opportunities for how we learn, teach, and assess in medicine. However, to take advantage of LLMs appropriately, trainees and faculty physicians supervising them must have a reasonable sense of how they work, their capabilities, and their limitations. This commentary addresses these areas with a focus on medical education, ChatGPT, and considerations for adoption.

How do they work?

Fundamentally, LLMs predict the probability of the next word in a given sequence of words.[2] To facilitate these predictions, words are first broken down into tokens.[2] Tokens are then mapped onto a multi-dimensional vector plane to indicate their relationship to other tokens. Words with similar meanings are grouped together creating embeddings.[2] Context is added to these embeddings, which are then fed into a neural network. Neural networks comprise multiple layers where initial layers extract simple features from input data and subsequent layers build on those features to generate predictions.[3] This neural network-based architecture is called a transformer.[2] Transformers are pre-trained on large amounts of unlabelled text and then fine-tuned on smaller corpi of texts for specific tasks.[2]

What are some medical education uses?

While there are several base LLMs and hundreds of derivative models, we have not observed specific LLMs for medical education (i.e., that have been trained on resources used to teach medical students or residents).[4] Despite this observation, trainees and faculty are already using LLMs, most commonly OpenAI's ChatGPT, for learning, teaching, and assessment.[5] The model has been used to assist with the creation of dynamic learning materials (e.g., practice cases), explain concepts using analogies, create questions and answer keys, and grade responses relative to an answer key.[5]

What are the limitations?

Despite these use cases, there are several limitations to ChatGPT that must be considered before its widespread adoption in medical education.[6] It is generally accepted that ChatGPT was trained on proprietary data that is not available for interrogation.[6] It is unclear whether these data are representative of the sources used to historically train medical students and residents (e.g., physical textbooks whose online versions are often paywalled, lecture slides and recordings protected within learning management systems, etc.). Even if these sources were incorporated, we know that many traditional sources of medical education contain bias, with textbooks and peer-reviewed papers focusing on the experiences and observations of those who have the socioeconomic means to access or receive care. Accordingly, the outputs generated by LLMs may not appropriately represent individuals from structurally vulnerable groups. ChatGPT can also fabricate information ("hallucinations"), is not deterministic in its responses, and be confidently wrong.[6] These issues, as well as privacy and concerns around the generation of copyrighted material warrant caution before widespread use.[6]

A way forward

Although these limitations exist, we know we cannot restrict access to ChatGPT and other LLMs. We know trainees are using them, and we must equip them with the knowledge, skills, and opportunities to experiment and discover safely. Such knowledge includes concepts covered in this viewpoint, an understanding of how to critically appraise machine learning research, and key questions to consider (Box 1). Following acquisition of these basics, students and residents would benefit from opportunities to practice using LLMs facilitated by interdisciplinary instructors, including clinicians and data scientists, before independent use.

Is completion of this task by an LLM allowed by my governing body/University?
Does this task require sharing of personal health information or identifying information?
Has the LLM I plan to use been trained on data I think are required to reasonably complete the task?
Am I able to judge the quality of the output of the LLM?
Am I able to replicate the output of the task using the same LLM?
Will the LLM be able to explain its response(s)?

Box 1. Considerations for use of LLM for medical education tasks

Conclusions

Much like the Internet, LLMs are here to stay. Safe and appropriate use requires updates to undergraduate and postgraduate medical curricula as well as continuing education opportunities for teaching faculty. With these changes, we can begin to truly discover and experiment with new ways of teaching, learning, and assessing in medicine.

Conflicts of Interest: Akshay co-founded and serves as the Chief Product Officer of 12676362 Canada Inc (dba Caddie Health), an early-stage start-up developing AI-powered medical billing software. While a version of the product has leveraged a large language model, our product is not available for sale and the company has no revenue.
Funding: Akshay owns equity in the company, but does not receive any other funding or salary support. He did not receive any other grant funding or support for this article.
Edited by: Marcel D'Eon (editor-in-chief)

References

1. Le T, Stein ML. Medical education and the internet: this changes everything. *JAMA*. 2001;285(6):809- https://doi.org/10.1001/jama.285.6.809-JMS0214-6-1
2. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *JAMA*. 2023;330(9):866-9. https://doi.org/10.1001/jama.2023.14217
3. Liu Y, Chen P-HC, Krause J, et al. How to read articles that use machine learning: users' guides to the medical literature. *JAMA.* 2019;322(18):1806-16. https://doi.org/10.1001/jama.2019.16489
4. Zhou H, Gu B, Zou X, et al. A survey of large language models in medicine: Progress, application, and challenge. arXiv preprint arXiv:231105112. 2023.
5. Safranek CW, Sidamon-Eristoff AE, Gilson A, et al. The role of large language models in medical education: applications and implications. *JMIR Med Educ*. 2023;9:e50945. https://doi.org/10.2196/50945
6. GPT-4 Technical Report: OpenAI. 2023. Available from: https://cdn.openai.com/papers/gpt-4.pdf. [Accessed Jan 4, 2024].