PRiMER
peer-reviewed reports in medical education research
letter to the editor

The Hazards of Using ChatGPT: A Call to Action for Medical Education Researchers

Winston Liaw, MD, MPH | Summer Chavez, DO, MPH, MPM | Cecilia Pham | Salik Tehami | Romi Govender

PRiMER. 2023;7:27.

Published: 8/21/2023 | DOI: 10.22454/PRiMER.2023.295710

To the Editor:

We share Dr Hanna's enthusiasm for ChatGPT and believe it will revolutionize teaching, but we also want to highlight concerns standing in the way of greater impact.[1]

1. **Overreliance.** An overreliance on ChatGPT has the potential to stunt the development of critical thinking and medical knowledge, similar to how excessive dependence on imaging studies can erode clinical skills. Because the output is only as good as the input, ChatGPT may provide inaccurate information if details are missing. For example, if a learner uses ChatGPT to create a differential diagnosis but fails to recognize a critical physical exam finding, the generated list will be incomplete.
2. **Accuracy.** Regardless of the user's level of training, the output may not be accurate, as the tool may struggle to distinguish between reliable sources and those propagating misinformation. Furthermore, its knowledge is currently limited to data up to September 2021 and thus does not include the millions of papers published since then. When training data are missing, ChatGPT can hallucinate, or generate new information in the absence of real-world data. During one task, researchers were unable to locate 16% (28) of the references cited by ChatGPT.[2] Accuracy is important because some are exploring whether the tool can replicate the work of physicians. Another recent study evaluated ChatGPT's performance on a licensing exam and found that it achieved a passing score.[3] However, it still missed over one-third of the questions, despite having access to vast medical references.
3. **Bias.** ChatGPT may teach learners the wrong lessons by perpetuating biases. For example, when asked to suggest professions for people from different racial backgrounds, genders, and sexual orientations, its predecessor (GPT-2) offered responses that reinforced stereotypes.[4] While more recent versions (GPT-3.5) showed increased awareness, these concerns will remain as long as bias is endemic in the training data.
4. **Communication Breakdown.** Because ChatGPT does not incorporate auditory and visual information, its assessment of learners is limited. For example, a learner may report comprehension, though their tone and facial expressions indicate otherwise.

Given the risks, efforts to integrate ChatGPT into teaching should be accompanied by evaluation, and we recommend leveraging previously-published AI competencies.[5] Using this framework, teachers and learners should be able to explain the tool, appraise the evidence behind it, identify appropriate indications for its use, operate the tool effectively, communicate its output, and recognize adverse effects. Through this process, they can identify the gaps in knowledge that require further investigation. Because ChatGPT can be used in a range of scenarios, the level of scrutiny required for implementation needs to be tailored, with a focus on more rigorous evidence for use cases that directly affect student learning and patient well-being.

ChatGPT has garnered attention because of its ability to mimic humans. Unfortunately, because humans are flawed and ChatGPT learns from humans, ChatGPT is flawed too. As a result, medical education researchers are desperately needed to quantify and mitigate the risks and ensure that the tool's adoption leads to better training and health.

Author Affiliations

Winston Liaw, MD, MPH - Department of Health Systems and Population Health Sciences, University of Houston Tilman J. Fertitta Family College of Medicine, Houston, TX

Summer Chavez, DO, MPH, MPM - Department of Health Systems and Population Health Sciences, University of Houston Tilman J. Fertitta Family College of Medicine, Houston, TX

Cecilia Pham - University of Houston Tilman J. Fertitta Family College of Medicine, Houston, TX

Salik Tehami - University of Houston Tilman J. Fertitta Family College of Medicine, Houston, TX

Romi Govender - University of Houston Tilman J. Fertitta Family College of Medicine, Houston, TX

References

1. Hanna K. Exploring the Applications of ChatGPT in Family Medicine Education: Five Innovative Ways for Faculty Integration. PRiMER. 2023;7:26. doi:10.22454/PRiMER.2023.985351
2. Athaluri SA, Manthena SV, Kesapragada VSRKM, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*. 2023;15(4):e37432. doi:10.7759/cureus.37432
3. 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
4. Sheng E, Chang KW, Natarajan P, Peng N. The woman worked as a babysitter: on biases in language generation. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics; 2019:3405-3410. doi:10.18653/v1/D19-1339
5. Liaw W, Kueper JK, Lin S, Bazemore A, Kakadiaris I. Competencies for the use of artificial intelligence in primary care. *Ann Fam Med*. 2022;20(6):559-563. doi:10.1370/afm.2887

Copyright © 2023 by the Society of Teachers of Family Medicine