

Letter to the Editor

In Reply to Daungsupawong and Wiwanitkit

Fabio Dennstädt, MD,^{a,b,*} Janna Hastings, PhD,^{c,d}
Paul Martin Putora, MD, PhD,^{a,b} Erwin Vu,^a Galina Fischer, MD, PhD,^a
Krisztian Süveg, MD,^a Markus Glatzer, MD,^a Elena Riggenschach, MD,^b
Hông-Linh Hà,^a and Nikola Cihoric, MD^b



^aDepartment of Radiation Oncology, Kantonsspital St. Gallen, St. Gallen, Switzerland; ^bDepartment of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland; ^cSchool of Medicine, University of St. Gallen, St. Gallen, Switzerland; and ^dInstitute for Implementation Science in Health Care, University of Zurich, Zurich, Switzerland

We thank the colleagues from Laos and India for their valuable comment on our work. Due to the considerable advantages in the field and the emerging new capabilities of large language models (LLMs), there is increasing interest in developing and applying such models in medicine.¹ Powerful LLMs optimized for application in the clinical context may support clinicians in daily life in the future. Radiation oncology, as a data-driven and technical medical discipline in the first place, may be particularly suitable for using these new technologies. To properly contextualize the role of this technology in the clinic and to recognize potential weaknesses and dangers of evaluation studies are of great importance. Our work is one of several new studies evaluating modern LLMs (in particular, ChatGPT) for different questions in radiation oncology. Other examples would include the works of Huang et al,² Floyd et al,³ or Holmes et al.⁴ For doing such evaluations, the answers of an LLM need to be compared with some sort of ground truth. Many studies done so far (including our own in part one of the evaluation) have used frameworks such as multiple-choice tests in which a clear correct or incorrect answer can be identified. However, if LLMs are indeed to be used for supporting clinicians in daily life, they have to give useful answers not only in standardized medical tests, but also to open-ended and real-life questions occurring in clinical practice. Evaluating the performance in these situations is more challenging because

there is usually not a single correct answer to each question. In part 2 of our evaluation, the quality of the LLM's answer was evaluated by independent radiation oncologists. Although this approach can give some impression of an answer's quality, the factual correctness and usefulness of an answer are not necessarily determined this way, as it rather resembles a democratic consensus among domain experts. The limitations of our study are discussed in more detail in our article. As you mentioned, the subjectivity of the evaluation is a notable limitation. Assessing the quality of answers to relevant medical questions about topics of limited knowledge remains an unresolved issue. We also agree on your point that the sample size of questions used in our exploratory study is not sufficient to provide a comprehensive picture on the subject. Further studies investigating and evaluating LLMs in medicine and radiation oncology are surely needed, and we are participating in ongoing trials with the International Society for Radiation Oncology Informatics (ISROI). More research on the subject is urgently needed as these technologies are increasingly being used by clinicians and patients.^{1,5}

Disclosures

Dr Cihoric is a technical lead for the SmartOncology project and medical advisor for Wemedoo AG, Steinhau-sen AG, Switzerland. The authors declare no other conflicts of interest.

Sources of support: This work had no specific funding.

*Corresponding author: Fabio Dennstädt, MD; Email: fabio.dennstaedt@insel.ch

<https://doi.org/10.1016/j.adro.2024.101511>

2452-1094/© 2024 The Author(s). Published by Elsevier Inc. on behalf of American Society for Radiation Oncology. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

References

1. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med.* 2023;3:141.
2. Huang Y, Goma A, Semrau S, et al. Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: potentials and challenges for ai-assisted medical education and decision making in radiation oncology. *Front Oncol.* 2023;13:1265024.
3. Floyd W, Kleber T, Pasli M, et al. Evaluating the reliability of ChatGPT model responses for radiation oncology patient inquiries. *Int J Radiat Oncol Biol Phys.* 2023;117:e383.
4. Holmes J, Liu Z, Zhang L, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Front Oncol.* 2023;13: 1219326.
5. Ravi A, Neinstein A, Murray SG. Large language models and medical education: Preparing for a rapid transformation in how trainees will learn to be doctors. *ATS Scholar.* 2023;4:282-292.