Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.264103

ARTIFICIAL INTELLIGENCE AND INFORMATICS

LETTER TO THE EDITOR

Reply: Evaluating the reference accuracy of large language models in radiology: a comparative study across subspecialties

Yasin Celal Güneş[1]
Turay Cesur[2]
Eren Çamur[3]

[1]Kırıkkale Yüksek İhtisas Hospital, Clinic of Radiology, Kırıkkale, Türkiye

[2]Mamak State Hospital, Clinic of Radiology, Ankara, Türkiye

[3]Ankara 29 Mayıs State Hospital, Clinic of Radiology, Ankara, Türkiye

Dear Editor,

We thank the authors for their thoughtful and well-articulated comments on our study.

We agree that large language model (LLM) performance is inherently multidimensional and that temporal variability and response stochasticity are important considerations. As outlined in our Discussion, these aspects—including the use of a single response per query and the absence of repeated sampling—were explicitly acknowledged as limitations of our study.[1]

Our research was intentionally designed to provide a standardized baseline comparison across models. The single-response-per-query approach was adopted to ensure methodological consistency and comparability while recognizing that it does not capture the full variability of LLM outputs. In this context, the points raised by the authors are valid and consistent with the methodological considerations outlined in our manuscript.

We also concur that reference accuracy represents only one component of overall LLM performance. However, we believe it remains a particularly critical domain in radiology, where clinical and academic practice depends on accurate and verifiable sources.[2,3] From this perspective, our focused evaluation addresses a fundamental aspect of LLM reliability.

The authors' emphasis on hallucination is particularly relevant. Our findings are consistent with prior studies demonstrating that fabricated or inaccurate references remain a persistent limitation across current LLMs, reinforcing the need for careful validation and human oversight.[4-6]

We agree that future research incorporating repeated sampling and broader performance metrics will further enhance the understanding of LLM behavior. Within this context, we believe our study provides a necessary and timely benchmark in this domain.

We thank the authors again for their valuable contribution to this discussion.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Güneş YC, Cesur T, Çamur E. Evaluating the reference accuracy of large language models in radiology: a comparative study across subspecialties. *Diagn Interv Radiol*. 2026;32(2):173-181 [Crossref]
2. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol*. 2024;30(2):80-90. [Crossref]
3. Wagner MW, Ertl-Wagner BB. Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. *Can Assoc Radiol J*. 2024;75(1):69-73. [Crossref]

Corresponding author: Yasin Celal Güneş

E-mail: gunesyasincelal@gmail.com

Received 01 May 2026; accepted 05 May 2026.

Epub: 08.05.2025

DOI: 10.4274/dir.2026.264103

You may cite this article as: Güneş YC, Cesur T, Çamur E. Reply: Evaluating the reference accuracy of large language models in radiology: a comparative study across subspecialties. *Diagn Interv Radiol*. 08 May 2026 DOI: 10.4274/dir.2026.264103 [Epub Ahead of Print].

4. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus*. 2023;15(5):e39238. [Crossref]

5. McGowan A, Gui Y, Dobbs M, et al. ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Res*. 2023;326:115334. [Crossref]

6. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep*. 2023;13(1):14045. [Crossref]