Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.264022

ARTIFICIAL INTELLIGENCE AND INFORMATICS

LETTER TO THE EDITOR

Letter to the Editor: Rethinking reference accuracy in large language models for radiology

Hamza Eren Güzel

Ministry of Health, İzmir City Hospital, Department of Radiology, İzmir, Türkiye

Dear Editor,

I read with great interest the recent article by Güneş et al.[1] evaluating the reference accuracy of large language models (LLMs) in radiology. The authors should be congratulated for addressing a critical and timely issue—namely, the high rates of fabricated and inaccurate citations generated by contemporary LLMs. Their findings clearly demonstrate that, despite rapid advancements, substantial limitations remain in the reliability of LLM-generated academic references.

Although reference accuracy represents an important component of LLM evaluation, it may not fully capture the complexity of real-world model performance. One critical yet underexplored aspect is the temporal variability of LLM outputs. Due to stochastic decoding processes, model updates, and backend modifications, identical prompts may yield different responses across sessions or time points. This variability has direct implications for study design and interpretation. In the study by Güneş et al.,[1] each model was evaluated using a single response per query, which, although methodologically practical, may not fully reflect the range of possible outputs in real-world usage. Consequently, a model demonstrating acceptable reference accuracy in a single instance may still exhibit inconsistent or unreliable behavior across repeated interactions.

Another important consideration is that reference accuracy represents only one dimension of a broader construct encompassing clinical reasoning, contextual understanding, and decision relevance. In a recent study evaluating LLM performance in an examination-style radiology setting modeled after the European Diploma in Radiology, discrepancies were observed between diagnostic reasoning performance and the quality or validity of the supporting information.[2] Specifically, models were able to generate clinically plausible answers despite inconsistencies in explanations or evidentiary support, suggesting a disconnect between linguistic coherence and factual grounding.

This limitation is closely related to the phenomenon of hallucination, in which LLMs generate syntactically plausible but factually incorrect information. Multiple studies across different domains and model architectures have consistently demonstrated high rates of fabricated or inaccurate references in LLM-generated outputs.[3-5] Importantly, this issue is not confined to a single model family but appears to be a widely observed limitation of current generative artificial intelligence systems. As highlighted by Güneş et al.,[1] such inaccuracies may introduce misinformation into both clinical and academic contexts. Furthermore, erroneous citations may propagate through secondary referencing, ultimately distorting the scientific record and undermining evidence-based practice.[6]

These concerns extend beyond technical limitations and raise broader issues related to scientific integrity, reproducibility, and knowledge verification. As LLMs become increasingly integrated into academic writing and clinical decision support, the need for critical appraisal and human oversight becomes paramount. Without rigorous validation, reliance on generated content may inadvertently compromise both the quality of scientific output and patient safety.

Corresponding author: Hamza Eren Güzel

E-mail: hamzaerenguzel@gmail.com

Received 26 March 2026; accepted 30 March 2026.

Epub: 08.05.2026

Publication date: 01.07.2026

DOI: 10.4274/dir.2026.264022

You may cite this article as: Güzel HE. Letter to the editor: rethinking reference accuracy in large language models for radiology. *Diagn Interv Radiol*. 2026;32(4):395-396.

In this context, future evaluations of LLMs in radiology would benefit from approaches that account for response variability and integrate multiple performance dimensions, including clinical correctness, reasoning transparency, and hallucination risk. Such comprehensive frameworks may provide a more accurate representation of model capabilities and limitations, ultimately supporting safer and more effective implementation in both research and clinical practice.

In conclusion, the study by Güneş et al.[1] provides valuable insights into the current limitations of LLM-generated references. However, reference accuracy should be interpreted within a broader and temporally aware evaluation framework. Addressing these challenges will be essential for the responsible integration of LLMs into radiology.

Footnotes

Conflict of interest disclosure

The author 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. Güzel HE, Oleaga L, Koç AM, Junquero V, Merino C. Large language models solving the European Diploma in Radiology: a comparative evaluation. *Acad Radiol*. 2026;33(5):1871-1878. [Crossref]
3. 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]
4. Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. *J Med Internet Res.* 2024;26:e53164. [Crossref]
5. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep*. 2023;13:14045. [Crossref]
6. Dumas-Mallet E, Boraud T, Gonon F. Le mésusage des citations et ses conséquences en médecine [Citation misuse and its effects on public health]. *Med Sci (Paris).* 2021;37(11):1035-1041. [Crossref]