Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.263854

NEURORADIOLOGY

LETTER TO THE EDITOR

Letter to the Editor: Comment on the diagnostic sensitivity of ChatGPT for detecting hemorrhages in cranial computed tomography scans

Yiğit Can Kartal

University of Health Sciences, Başakşehir Çam and Sakura City Hospital, Department of Radiology, İstanbul, Türkiye

Dear Editor,

I read with great interest the article by Bayar-Kapıcı et al.[1] evaluating the diagnostic sensitivity of a multimodal large language model (MLLM) in detecting intracranial hemorrhage in non-contrast cranial computed tomography. The authors are to be commended for addressing a timely and clinically relevant topic and for providing a systematic evaluation of an MLLM in the context of acute neuroimaging.

The reported findings offer valuable insight into the current capabilities and limitations of MLLMs in image-based diagnostic tasks. In particular, the observed performance of MLLMs in subtle and borderline hemorrhagic findings may be interpreted in light of the inherent diagnostic complexity of these cases. In routine clinical practice, such findings often fall into a diagnostic gray zone, in which a degree of interpretive variability is unavoidable. From this perspective, discordance between an MLLM's output and the reference interpretation may, in part, reflect intrinsic diagnostic uncertainty rather than the true inadequacy of the model.

The image-only evaluation framework employed in the study represents a deliberately controlled and methodologically sound approach. However, given the reasoning-based architecture of MLLMs, diagnostic behavior may reasonably be influenced by the availability of a minimal, clinically relevant context. The inclusion of limited clinical cues—such as trauma history, anticoagulant use, or the patient's overall clinical condition—may provide a more representative assessment of how these models could function in real-world decision-support scenarios. In this regard, recent evidence suggests that prompt engineering and input conditions play a decisive role in how MLLMs integrate the clinical context with imaging data, with measurable effects on diagnostic performance.[2] Collectively, these observations underscore the synergistic role of clinical contexts in diagnostic reasoning.

Another aspect worth considering is the reliance on one or two preselected image slices for evaluation. Radiologic interpretation in clinical practice often benefits from reviewing adjacent slices across an image series together with dynamic window and level adjustments, particularly for subtle hemorrhagic findings or for distinguishing a true pathology from artifacts. Evaluation based on isolated images without the ability to adjust window settings, although practical for experimental design, may therefore differ from routine diagnostic workflows and influence measured performance metrics.

Finally, although sensitivity remains a critical metric, the potential for incorrect affirmative or negative outputs in the absence of sufficient contextual and visual grounding represents an important safety consideration, particularly in high-stakes neuroimaging scenarios, in which confidently presented but incorrect model outputs may have clinical consequences. Future investigations incorporating clinically realistic prompts, slice-to-slice correlation with adjustable windowing, and human-in-the-loop frameworks may help clarify how MLLMs—designed to jointly process visual and textual information—can best be integrated into clinical practice as assistive tools rather than standalone diagnostic systems. Within such frameworks, human–artificial intelligence collaboration, in which MLLMs support radiologists by highlighting potential abnormalities while final interpretation and decision-making remain

Corresponding author: Yiğit Can Kartal

E-mail: yckartal@hotmail.com

Received 05 January 2026; accepted 16 January 2026.

Epub: 16.02.2026

Publication date: xx.xx.2026

DOI: 10.4274/dir.2026.263854

You may cite this article as: Kartal YC. Letter to the Editor: comment on the diagnostic sensitivity of ChatGPT for detecting hemorrhages in cranial computed tomography scans. *Diagn Interv Radiol.* 16 February 2026 DOI: 10.4274/dir.2026.263854 [Epub Ahead of Print].

clinician driven, may represent the most appropriate and safest paradigm for clinical deployment.

I congratulate the authors on this valuable contribution and hope that these considerations may support continued efforts to develop clinically meaningful and safe applications for MLLMs in radiology.

Conflict of interest disclosure

The author declared no conflicts of interest.

References

1. Bayar-Kapıcı O, Altunışık E, Musabeyoğlu F, Dev Ş, Kaya Ö. Artificial intelligence in radiology: diagnostic sensitivity of ChatGPT for detecting hemorrhages in cranial computed tomography scans. *Diagn Interv Radiol.* 2026;32(1):27-32. **[Crossref]**
2. Han T, Jeong WK, Shin J. Diagnostic performance of multimodal large language models in radiological quiz cases: the effects of prompt engineering and input conditions. *Ultrasonography*. 2025;44(3):220-31. **[Crossref]**