Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.263888

ARTIFICIAL INTELLIGENCE AND INFORMATICS

LETTER TO THE EDITOR

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

Olga Bayar Kapıcı[1]
Erman Altunışık[2]
Feyza Musabeyoğlu[2]
Şeyda Dev[2]
Ömer Kaya[3]

[1]Seyhan State Hospital, Clinic of Radiology, Adana, Türkiye

[2]University of Health Sciences Türkiye, Gaziantep City Hospital, Clinic of Neurology, Gaziantep, Türkiye

[3]Çukurova University Faculty of Medicine, Department of Radiology, Adana, Türkiye

Dear Editor,

We would like to thank the correspondent for their thoughtful and constructive comments on our article evaluating the diagnostic sensitivity of a multimodal large language model (MLLM) (ChatGPT-4V) for detecting intracranial hemorrhage in non-contrast cranial computed tomography.[1] We appreciate the opportunity to clarify the rationale behind our experimental design and to outline directions for future work.

First, we agree that subtle and borderline hemorrhagic findings represent a well-known diagnostic gray zone, even for human readers, and that discordance may partly reflect intrinsic interpretive uncertainty rather than a purely model-specific limitation.[1] In our dataset, ChatGPT-4V's performance was clearly influenced by lesion conspicuity; larger hemorrhage diameters were associated with higher correct classification rates, particularly for epidural and subdural hematomas.[2] This finding is consistent with the broader literature showing that MLLM performance using direct image inputs remains variable across tasks and settings and may be limited in the context of real-world radiologic interpretation.[3,4]

Second, we fully concur that clinical contexts can materially shape diagnostic reasoning.[1] Our study intentionally adopted an image-only framework to quantify baseline behavior under controlled conditions and to isolate the effect of the prompt structure. Specifically, after an initial open-ended prompt (Q2), we introduced a minimal, targeted clue ("There is bleeding…") (Q3) to test whether structured guidance influences performance.[2] The substantial improvement observed with this guided prompt supports the correspondent's emphasis on input conditions and prompt engineering.[2] It also aligns with published radiology-focused research indicating that prompt optimization (including structured prompting and few-shot approaches) can meaningfully influence LLM outputs and utility.[5]

Third, regarding the reliance on one or two preselected slices and the absence of dynamic window/level adjustments, we agree this differs from the routine radiologic workflow, in which multi-slice review and interactive windowing are integral, especially for subtle hemorrhage and artifact discrimination.[1] In our Methods section, we provided representative two-dimensional slices to approximate a best-case static-input scenario.[2] We acknowledge that a workflow-faithful evaluation would ideally allow multi-slice correlation (or a full-series review) and window/level control. These priorities are also reflected in broader multimodal GPT-4V radiology evaluations that highlight sensitivity to input presentation and context handling.[6]

Finally, we strongly support the safety considerations highlighted by the correspondent.[1] In our conclusion, we emphasized that the model is not suitable for autonomous radiologic interpretation and should be considered, at most, as a supervised adjunct within human-in-the-loop paradigms.[2] This caution is consistent with the emerging radiology-related literature emphasizing that MLLMs that use direct image input have not yet reached a level appropriate for unsupervised clinical deployment.[3,4,6]

Corresponding author: Olga Bayar Kapıcı

E-mail: olgasahbayar@gmail.com

Received 20 January 2026; accepted 31 January 2026.

Epub: 16.02.2026

Publication date: xx.xx.2026

DOI: 10.4274/dir.2026.263888

You may cite this article as: Kapıcı OB, Altunışık E, Musabeyoğlu F, Dev Ş, Kaya Ö. Reply: 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.2026.263888 [Epub Ahead of Print].

We thank the correspondent again for their insightful remarks, which closely align with the key implications of our findings and help frame a clear agenda for clinically meaningful and safe evaluation of multimodal language–vision models in radiology.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Kartal YC. Letter to the Editor: comment on the diagnostic sensitivity of ChatGPT for detecting hemorrhages in cranial computed tomography scans**.** *Diagn Interv Radiol.* [Crossref]
2. 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]
3. Zhou Y, Ong H, Kennedy P, et al. Evaluating GPT-V4 (GPT-4 with Vision) on detection of radiologic findings on chest radiographs. *Radiology*. 2024;311(2):e233270. Erratum in: *Radiology*. 2024;311(2):e249016. [Crossref]
4. Suh PS, Shim WH, Suh CH, et al. Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini Pro Vision using image inputs from diagnosis please cases. Radiology. 2024;312(1):e240273. [Crossref]
5. Russe MF, Reisert M, Bamberg F, Rau A. Improving the use of LLMs in radiology through prompt engineering: from precision prompts to zero-shot learning. Rofo. 2024;196(11):1166-1170. [Crossref]
6. Busch F, Han T, Makowski MR, Truhn D, Bressem KK, Adams L. Integrating text and image analysis: exploring GPT-4V's capabilities in advanced radiological applications across subspecialties. J Med Internet Res. 2024;26:e54948. Erratum in: J Med Internet Res. 2024;26:e64411. [Crossref]