Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.263940

ARTIFICIAL INTELLIGENCE AND INFORMATICS

LETTER TO THE EDITOR

Letter to the Editor: Artificial intelligence in radiology: diagnostic sensitivity of ChatGPT for detecting hemorrhages in cranial computed tomography scans

Emre Utkan Büyükceran
Ayça Seyfettin
Andelib Babatürk

Ankara Güven Hospital, Clinic of Radiology, Ankara, Türkiye

Corresponding author: Emre Utkan Büyükceran

E-mail: utkan.buyukceran91@gmail.com

Received 20 February 2026; accepted 20 February 2026.

Epub: 11.03.2026

Publication date:

DOI: 10.4274/dir.2026.263940

Dear Editor,

We read with great interest the article by Bayar-Kapıcı et al.[1] entitled "Artificial intelligence in radiology: Diagnostic sensitivity of ChatGPT for detecting hemorrhages in cranial computed tomography scans". The authors evaluated the performance of Generative Pre-trained Transformer 4 with Vision (GPT-4V) in detecting intracranial hemorrhage on non-contrast cranial computed tomography (CT) and demonstrated a marked increase in sensitivity following guided prompting. We commend the authors for addressing a timely and clinically relevant topic. Nevertheless, we believe several methodological considerations merit further discussion.

First, the study appears to assess model responses within a single session. Large language and vision models are inherently probabilistic systems, and identical inputs may yield variable outputs across independent runs. In diagnostic settings, such variability may introduce uncertainty, as the same clinical case could potentially generate inconsistent diagnostic suggestions upon repeated evaluation.[2,3] This issue reflects broader methodological concerns in artificial intelligence (AI) assessment. The U.S. Food and Drug Administration's draft guidance on AI-enabled medical software emphasizes the importance of quantifying model variability through repeatability and reproducibility analyses.[4] In radiology, diagnostic consistency is fundamental to clinical reliability. Accordingly, AI systems proposed for diagnostic support should also be examined for inter-session stability. A study design incorporating repeated measurements across independent sessions would have allowed an assessment of the robustness of the reported performance metrics. Without such analysis, it remains unclear whether the reported sensitivity reflects stable diagnostic behavior or a session-dependent fluctuation.

Second, the observed increase in sensitivity from Q2 to Q3 should be interpreted with caution. The Q3 prompt ("There is hemorrhage in this image…") implicitly assumes the presence of a pathology and narrows the diagnostic search space. Consequently, this approach evaluates conditional classification rather than primary detection performance. From a clinical safety perspective, particularly in acute intracranial hemorrhage for which early identification is critical, baseline detection sensitivity remains the principal parameter. Though structured prompting may enhance performance, it does not mitigate the relatively low initial sensitivity reported in the unguided setting (23.6%).

Third, similar to the aforementioned issue, the study design relies on isolated two-dimensional slices rather than full CT examinations. Radiologic interpretation inherently involves a volumetric assessment, including multi-slice review, window adjustments, and an evaluation of anatomical continuity. Assessment based on representative single images may not adequately reflect real-world clinical practice. Studies incorporating complete volumetric datasets and workflow-oriented evaluation designs have demonstrated that clinically applicable AI systems are typically validated at the scan level rather than on selected slices.[5] Future investigations employing volumetric imaging data and interactive reading conditions would likely provide more robust evidence regarding clinical applicability.

You may cite this article as: Büyükceran EU, Seyfettin A, Babatürk A, Letter to the Editor: Artificial intelligence in radiology: diagnostic sensitivity of ChatGPT for detecting hemorrhages in cranial computed tomography scans. *Diagn Interv Radiol.* 11 March 2026 DOI: 10.4274/dir.2026.263940 [Epub Ahead of Print].

Finally, GPT-4V is not a model specifically trained or optimized on radiologic datasets. Comparing its performance with convolutional neural networks developed explicitly for hemorrhage detection, or with radiologist performance benchmarks, would have provided clearer insight into its relative clinical utility.

Despite these limitations, the study offers valuable preliminary evidence regarding the potential role of large language models in radiology. We believe that future research should prioritize reproducibility analyses, the multimodal integration of clinical data, and prospective clinical validation to ensure safe and reliable implementation in diagnostic workflows.

We congratulate the authors for contributing empirical data to this rapidly evolving field and hope that an ongoing methodological dialogue will further refine the evaluation of AI in radiologic practice.

Conflict of interest disclosure

The authors 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. Büyükceran EU, Seyfettin A, Babatürk A, et al. Text-based prediction of ımmunohistochemical biomarkers in breast cancer using a generative large language model: a retrospective study. Health Inf Sci Syst. 2025;14(1):3. [Crossref]
3. Gu B, Desai RJ, Lin KJ, Yang J. Probabilistic medical predictions of large language models. NPJ Digit Med. 2024;7:367. [Crossref]
4. US Food and Drug Administration. Artificial intelligence-enabled device software functions: lifecycle management and marketing submission recommendations. Draft guidance for industry and Food and Drug Administration staff [Internet]. Silver Spring (MD): FDA; 2025 [cited 2026 Feb 13]. [Crossref]
5. Flanders AE, Prevedello LM, Shih G, et al. Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiol Artif Intell. 2020;2(3):e190211. Erratum in: Radiol Artif Intell. 2020;2(4):e209002. [Crossref]