Editorial

eISSN 2005-8330
https://doi.org/10.3348/kjr.2025.1805
Korean J Radiol 2026;27(4):297-300

Korean Journal of Radiology
KJR

Do General-Purpose Multimodal Large Language Models Really See Radiologic Images or Rely on Text?

Pae Sun Suh[1], Chong Hyun Suh[2]

[1]Department of Radiology, Research Institute of Radiological Science and Center for Clinical Imaging Data Science, Yonsei University College of Medicine, Seoul, Republic of Korea
[2]Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

Keywords: Large language model; Vision capability; Radiologic image

INTRODUCTION

Large language models (LLMs) have evolved rapidly and have been applied across many medical fields. While LLMs are fundamentally pretrained on massive amounts of textual data, the integration of "vision" capabilities into LLMs, most commonly referred to as multimodal LLMs, has generated substantial interest in the medical field, particularly in radiology. Since the introduction of GPT-4 Turbo with Vision (GPT-4V) in late 2023, expectations have risen that LLMs could serve supportive roles in radiology, especially for image interpretation and differential diagnosis. Numerous studies have evaluated both proprietary and open-source multimodal LLMs capable of processing both images and text, such as GPT, Gemini, Claude, and LLaMA series, to assess their performance and to determine whether they can approach radiologist-level performance.

However, recent evidence commonly suggests a different reality, with overall unsatisfactory results regarding the understanding of radiologic images. While these models may perform well on standardized examination-style diagnostic questions, their performance appears to depend primarily on textual processing rather than genuine visual interpretation. This editorial aims to comprehensively understand current evidence on multimodal LLMs' performance in radiology, examine whether they truly interpret radiologic images, highlight their limitations in real-world clinical applications, and outline expectations for future development.

AT-A-GLANCE SNAPSHOT OF STUDIES REPORTING DIAGNOSTIC ACCURACY OF MULTIMODAL LLMs

Early studies investigated multimodal LLMs' diagnostic accuracy using medical quiz cases that included radiologic images sourced from journal websites or clinical vignettes. Using various quiz-style cases, LLMs have demonstrated noteworthy diagnostic performance compared to radiologists. In an experimental study that assessed the diagnostic accuracy of GPT-4V and Gemini Pro Vision on Diagnosis Please cases from *Radiology*, in which models generated three differential diagnoses after being provided both images and textual inputs, GPT-4V achieved a top-3 accuracy of 49%, without a statistically significant difference compared with board-certified subspecialty-trained radiologists [1]. Mukherjee et al. [2] reported that radiologists and residents failed to outperform GPT-4V on challenging diagnostic cases from the Case of the Day questions at the RSNA 2023 Annual Meeting. Similarly, OpenAI o1 and GPT-4o achieved accuracy comparable to senior radiologists when answering Case of the Day questions at the RSNA 2024 Annual Meeting [3]. Conversely, some studies have reported that LLMs did not reach the

Received: November 26, 2025 **Revised:** December 30, 2025
Accepted: December 31, 2025
Corresponding author: Pae Sun Suh, MD, PhD, Department of Radiology, Research Institute of Radiological Science and Center for Clinical Imaging Data Science, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea
• E-mail: blackmizz@yuhs.ac

diagnostic performance of radiologists. Their performance on challenging cases with radiologic images from the *New England Journal of Medicine* (*NEJM*) [4] and *Clinical Neuroradiology* [5] remained inferior to that of board-certified radiologists and residents.

Before digging deeper, it is worth noting that even when studies report LLMs' diagnostic performance on cases with radiologic images to be comparable to or close to that of radiologists, such results do not necessarily directly reflect the models' visual reasoning capabilities. For example, in the study by Suh et al. [1], LLMs' performance markedly decreased to 15% when considering top-1 accuracy, raising concerns that LLMs may simply list differential diagnoses probabilistically based on training data, rather than performing case-specific image interpretation. Therefore, to evaluate vision capabilities accurately, it is essential to examine the impact of image inputs and the model's image-based rationale for differential diagnosis.

TEXT DOMINANCE AND THE LIMITED IMPACT OF IMAGES ON DIAGNOSTIC PERFORMANCE

To evaluate the contribution of image input, many studies have compared LLMs' diagnostic accuracy across different input types composed of text and images. Several investigations have consistently reported that image inputs showed only a minor influence on LLMs' performance. On board-style examinations, LLMs demonstrated a significant drop in accuracy when answering image-based questions compared with text-only questions [6-8], and a recent meta-analysis further reinforced these findings [9]. Furthermore, adding images to textual inputs failed to improve accuracy in challenging medical cases [2,4,10]. While radiologists naturally improved their diagnostic accuracy when both text and images were available, LLMs demonstrated similar or even lower accuracy when images were additionally provided. This phenomenon persists when using real-world radiologic images. GPT-4V demonstrated poor diagnostic performance when only images were provided, with accuracy improving markedly when textual clinical information was added [11,12].

In contrast, textual inputs have shown a greater influence on performance. Text dependency was strongly highlighted in the analysis of *NEJM* Image Challenge cases [4]. This study compared accuracy between long and short text inputs, and found that longer texts significantly increased LLMs' accuracy, whereas human readers were unaffected by text length. Schramm et al. [13] assessed GPT-4V's diagnostic accuracy on challenging brain MRI cases under various input conditions, including images, annotations, medical history, and image descriptions. The authors found that annotated or unannotated images alone yielded very low diagnostic accuracy. In contrast, providing image descriptions had the strongest effect on diagnostic performance, with an odds ratio of 68. These findings suggest that current multimodal LLMs heavily rely on textual inputs, even when images are available.

RATIONALE ANALYSIS: DO LLMs TRULY "READ" RADIOLOGIC IMAGES?

Beyond diagnostic accuracy, analyzing the rationale for diagnoses based on image interpretation is crucial for assessing vision capabilities. Only a limited number of studies have systematically examined LLMs' response rationales, largely because such evaluation requires time-consuming and labor-intensive human assessment of each response. LLMs have shown notable limitations in image understanding, including difficulty in identifying key elements on medical images, particularly in incorrect answers [14,15]. On radiologic images, although LLMs can often correctly identify the imaging modality, imaging plane, sequence, and anatomical region [4], they frequently fabricate or overlook abnormalities even when they arrive at the correct final diagnosis. GPT-4o and Gemini 1.5 Pro frequently failed to identify key imaging findings in Diagnosis Please cases from *Radiology* [10], and similar hallucinations have been observed with real-world radiologic images [11]. Interestingly, providing textual information significantly reduced the number of hallucinations. This hallucination hazard poses a significant potential risk to patient safety if LLMs are used in clinical practice without rigorous oversight by radiologists.

RAPID EVOLUTION, BUT NOT YET READY FOR CLINICAL PRACTICE

Many studies have revealed unsatisfactory results regarding the vision capabilities of LLMs on radiologic images, indicating that these models are not currently capable of adequately interpreting radiologic images or achieving performance comparable to radiologists. Limited vision capabilities are obvious due to the dominance of language over visual data during pretraining. One major

https://doi.org/10.3348/kjr.2025.1805

Korean Journal of Radiology
KJR

barrier to using LLMs for radiologic image analysis is the limited availability of large-scale, high-quality annotated imaging datasets [16]. Even when image data are available, vision encoders typically capture visual features from images, which may restrict access to the full range of image information, such as exact lesion diameters in physical units or CT Hounsfield units. Additionally, training data are inherently biased across demographic groups, languages, and imaging modalities, which can reduce generalizability and increase the risk of hallucinations [17].

Nonetheless, LLMs are evolving rapidly, and their diagnostic performance continues to improve. Recently released models, such as OpenAI's o series and Google's Gemini 2.5 Pro, have demonstrated performance advancements on medical quizzes [3,18]. Furthermore, in contrast to earlier findings, some recent models have shown accuracy improvement when image inputs are added [19] and have exhibited no obvious hallucinations in specific experimental settings [18]. Future model iterations may further enhance vision capabilities on radiologic images. Furthermore, innovations in vision encoders, such as vision transformers [20], may improve the capacity to handle diverse radiologic imaging modalities. Moreover, the development and refinement of domain-specific multimodal LLMs tailored to radiologic image-text data are essential for accurate image interpretation [21]. Such domain-specific models may outperform general-purpose multimodal LLMs, which are primarily pretrained on web-based natural image datasets.

Nevertheless, existing studies evaluating vision capabilities have important limitations with respect to representing real-world clinical practice [22]. Most studies have used highly curated clinical vignettes or selected images as inputs and assessed performance via multiple-choice responses, which differ substantially from routine clinical workflows. In addition, systematic rationale analysis has been limited. Study transparency and reproducibility also need to be strengthened for reliable clinical application by adhering to dedicated reporting guidelines such as the Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare (MI-CLEAR-LLM) [23-25]. Therefore, further investigations are warranted to determine how multimodal LLMs can effectively assist radiologists' decision-making in complex real-world clinical settings.

CONCLUSION

At present, multimodal LLMs remain far from being comparable to radiologists for radiologic image interpretation. Although they may show promising diagnostic performance on standardized examination-style radiologic cases, their vision capabilities remain unreliable, are heavily dependent on textual inputs, and are prone to hallucinations. In the era of rapidly evolving LLMs, the next step is to move beyond simple accuracy assessments on standardized examination-style diagnostic questions and to incorporate rigorous rationale analysis within real-world studies, thereby enabling the safe and effective integration of these tools into clinical practice.

Conflicts of Interest

Chong Hyun Suh, Assistant to the Editor of the *Korean Journal of Radiology*, was not involved in the editorial evaluation or decision to publish this article. The remaining author has declared no conflicts of interest.

Author Contributions

Conceptualization: Pae Sun Suh. Supervision: Chong Hyun Suh. Writing—original draft: Pae Sun Suh. Writing—review & editing: Chong Hyun Suh.

ORCID IDs

Pae Sun Suh
https://orcid.org/0000-0002-8618-9558
Chong Hyun Suh
https://orcid.org/0000-0002-4737-0530

Funding Statement

None

REFERENCES

1. Suh PS, Shim WH, Suh CH, Heo H, Park CR, Eom HJ, et al. Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini Pro Vision using image inputs from diagnosis please cases. *Radiology* 2024;312:e240273
2. Mukherjee P, Hou B, Suri A, Zhuang Y, Parnell C, Lee N, et al. Evaluation of GPT large language model performance on RSNA 2023 case of the day questions. *Radiology* 2024;313:e240609
3. Hou B, Mukherjee P, Batheja V, Wang KC, Summers RM, Lu Z. One year on: assessing progress of multimodal large language model performance on RSNA 2024 case of the day questions. *Radiology* 2025;316:e250617

4. Suh PS, Shim WH, Suh CH, Heo H, Park KJ, Kim PH, et al. Comparing large language model and human reader accuracy with New England Journal of Medicine image challenge case image inputs. *Radiology* 2024;313:e241668
5. Horiuchi D, Tatekawa H, Oura T, Oue S, Walston SL, Takita H, et al. Comparing the diagnostic performance of GPT-4-based ChatGPT, GPT-4V-based ChatGPT, and radiologists in challenging neuroradiology cases. *Clin Neuroradiol* 2024;34:779-787
6. Hayden N, Gilbert S, Poisson LM, Griffith B, Klochko C. Performance of GPT-4 with vision on text- and image-based ACR diagnostic radiology in-training examination questions. *Radiology* 2024;312:e240153
7. Martini RS, Sang A, Saunders P, Bala W, Li H, Moon JT, et al. Artificial intelligence in radiology: performance of ChatGPT-4v and GPT-4o on diagnostic radiology in-training (DXIT) examination questions. *J Am Coll Radiol* 2025 Oct 30 [Epub]. http://doi.org/10.1016/j.jacr.2025.10.026
8. Choi A, Kim HG, Choi MH, Ramasamy SK, Kim Y, Jung SE. Performance of GPT-4 Turbo and GPT-4o in Korean Society of Radiology in-training examinations. *Korean J Radiol* 2025;26:524-531
9. Nguyen D, Kim GHJ, Bedayat A. Evaluating ChatGPT's performance across radiology subspecialties: a meta-analysis of board-style examination accuracy and variability. *Clin Imaging* 2025;125:110551
10. Le Guellec B, Bruge C, Chalhoub N, Chaton V, De Sousa E, Gaillandre Y, et al. Comparison between multimodal foundation models and radiologists for the diagnosis of challenging neuroradiology cases with text and images. *Diagn Interv Imaging* 2025;106:345-352
11. Huppertz MS, Siepmann R, Topp D, Nikoubashman O, Yüksel C, Kuhl CK, et al. Revolution or risk?—assessing the potential and challenges of GPT-4V in radiologic image interpretation. *Eur Radiol* 2025;35:1111-1121
12. Brin D, Sorin V, Barash Y, Konen E, Glicksberg BS, Nadkarni GN, et al. Assessing GPT-4 multimodal performance in radiological image analysis. *Eur Radiol* 2025;35:1959-1965
13. Schramm S, Preis S, Metz MC, Jung K, Schmitz-Koep B, Zimmer C, et al. Impact of multimodal prompt elements on diagnostic performance of GPT-4V in challenging brain MRI cases. *Radiology* 2025;314:e240689
14. Yang Z, Yao Z, Tasmin M, Vashisht P, Jang WS, Ouyang F, et al. Unveiling GPT-4V's hidden challenges behind high accuracy on USMLE questions: observational study. *J Med Internet Res* 2025;27:e65146
15. Jin Q, Chen F, Zhou Y, Xu Z, Cheung JM, Chen R, et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *NPJ Digit Med* 2024;7:190
16. Nam Y, Kim DY, Kyung S, Seo J, Song JM, Kwon J, et al. Multimodal large language models in medical imaging: current state and future directions. *Korean J Radiol* 2025;26:900-923
17. Sun Y, Wen X, Zhang Y, Jin L, Yang C, Zhang Q, et al. Visual-language foundation models in medical imaging: a systematic review and meta-analysis of diagnostic and analytical applications. *Comput Methods Programs Biomed* 2025;268:108870
18. Salbas A, Yogurtcu M. Performance of large language models on radiology residency in-training examination questions. *Acad Radiol* 2026;33:337-347
19. Hirano Y, Miki S, Yamagishi Y, Hanaoka S, Nakao T, Kikuchi T, et al. Assessing accuracy and legitimacy of multimodal large language models on Japan Diagnostic Radiology Board Examination. *Jpn J Radiol* 2026;44:209-217
20. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv [Preprint]. 2020 [accessed on December 24, 2025]. Available at: https://doi.org/10.48550/arXiv.2010.11929
21. Wu J, Wang Y, Zhong Z, Liao W, Trayanova N, Jiao Z, et al. Vision-language foundation model for 3D medical imaging. *Npj Artif Intell* 2025;1:17
22. Suh CH, Suh PS. Evolving multimodal large language models in radiology: a year of diagnostic progress. *Radiology* 2025;316:e252282
23. Park SH, Suh CH, Lee JH, Tejani AS, You SC, Kahn CE, et al. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM): 2025 updates. *Korean J Radiol* 2025;26:1123-1132
24. Suh CH, Yi J, Shim WH, Heo H. Insufficient transparency in stochasticity reporting in large language model studies for medical applications in leading medical journals. *Korean J Radiol* 2024;25:1029-1031
25. Ko JS, Heo H, Suh CH, Yi J, Shim WH. Adherence of studies on large language models for medical applications published in leading medical journals according to the MI-CLEAR-LLM checklist. *Korean J Radiol* 2025;26:304-312