Original Article | Artificial Intelligence

eISSN 2005-8330
https://doi.org/10.3348/kjr.2025.1045
Korean J Radiol 2026;27(3):214-226

Korean Journal of Radiology
KJR

Check for updates

Evaluating the Accuracy and Diagnostic Reasoning of Multimodal Large Language Models in Interpreting Neuroradiology Cases From *RadioGraphics*

Pae Sun Suh[1]*, Ji Su Ko[2]*, Woo Hyun Shim[3,4], Hwon Heo[4], Chang-Yun Woo[5], Hyungjun Park[6], Chong Hyun Suh[3]

[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, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[3]Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[4]Department of Medical Science, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[5]Department of Internal Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[6]Department of Pulmonology, Shihwa Medical Center, Siheung, Republic of Korea

Objective: To evaluate the accuracy and reasoning capabilities of large multimodal language models compared with those of neuroradiology subspecialty-trained radiologists in neuroradiology case interpretation.
Materials and Methods: This experimental study used custom-made 401 radiologic quizzes derived from articles published in *RadioGraphics* covering neuroradiology and head and neck topics (October 2020 to February 2024). We prompted the GPT-4 Turbo with Vision (GPT-4V), GPT-4 Omni, Gemini Flash, and Claude models to provide the top three differential diagnoses with a rationale and describe examination characteristics such as imaging modality, sequence, use of contrast, image plane, and body part. The temperature was adjusted to 0 and 1 (T1). Two neuroradiologists answered the same questions. The accuracies of the large language models (LLMs) and the neuroradiologists were compared using generalized estimating equations. Three neuroradiologists assessed the rationale provided by the LLMs for their differential diagnoses using four-point scales, separately for specific lesion locations and imaging findings, and evaluated the presence of hallucinations and the overall acceptability of the responses.
Results: Top-3 accuracy (i.e., correct answers present among top-3 differential diagnoses) of LLMs ranged from 29.9% (120 of 401) to 49.4% (198 of 401, obtained with GPT-4V in the T1 setting), while radiologists achieved 80.3% (322 of 401) and 68.3% (274 of 401), respectively ($P < 0.001$). Regarding the rationale for differential diagnoses, GPT-4V (T1) accurately identified both the specific lesion location and imaging findings in 30.7% (123 of 401) and 12.9% (16 of 124) of cases without textual clinical history. Hallucinations occurred in 4.5% (18 of 401), and only 29.4% (118 of 401) of the LLM-generated analyses were deemed acceptable. GPT-4V (T1) demonstrated high accuracy in identifying the imaging modality (97.4% [800 of 821]) and scanned body parts (92.2% [756 of 820]).
Conclusion: LLMs remarkably underperformed compared with neuroradiologists and showed unsatisfactory reasoning for their differential diagnoses, with performance declining further in cases without textual input of clinical history. These findings highlight the limitations of current multimodal LLMs in neuroradiological interpretation and their reliance on text input.
Keywords: Large language model; Vision capability; Image interpretation; Rationale evaluation

Received: May 21, 2025 **Revised:** December 18, 2025 **Accepted:** December 18, 2025
*These authors contributed equally to this work.
Corresponding author: Chong Hyun Suh, MD, PhD, Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea
• E-mail: chonghyunsuh@amc.seoul.kr

INTRODUCTION

The rapid development of large language models (LLMs) has led to significant advancements in natural language processing, with applications extending beyond general tasks to specialized fields such as medicine and radiology. LLMs, such as OpenAI's Generative Pretrained Transformer 4 (GPT-4), have demonstrated the ability to understand complex languages and contexts, showing substantial results in diagnostic performance and generating radiology reports [1-3]. Multimodal LLMs with vision capabilities, such as the GPT-4 Turbo with Vision (GPT-4V), which can process both text and image inputs, have been proposed. They have demonstrated the ability to solve medical quizzes with radiologic image inputs, and have shown the potential to transform radiologic practice by streamlining workflows, providing clinical decision support, and offering second opinions on complex cases [3-6].

Despite their notable performance, the effectiveness of LLMs in interpreting radiological images remains uncertain. Although LLMs can accurately recognize imaging modalities and anatomical regions from radiologic images [6-8], several studies have consistently reported that their diagnostic accuracy is higher when using text-based inputs than when using image inputs [9-12]. Moreover, no significant advantage has been demonstrated with multimodal inputs over text alone [8,10,13]. Their performance varies significantly depending on the context and details of the input, raising concerns regarding their ability to replicate the nuanced reasoning of human radiologists [6,14]. This challenge highlights the need for a more rigorous evaluation of LLMs' diagnostic capabilities of LLMs, focusing on their ability to closely simulate radiologists in clinical practice, and their rationale for diagnosis. Few studies have investigated the rationale behind interpreting LLMs' images. These studies demonstrated that LLMs often fail to identify or describe key abnormalities in images when using publicly available open-source or real-world clinical data [8,15,16].

This study assessed the diagnostic accuracy and interpretative ability of multimodal LLMs in neuroradiology. We created radiological quizzes derived from *RadioGraphics*, a highly respected and educational journal in the field of radiology, to ensure independence from potential training data and reduce the risk of overestimating the model performance [17]. Using these as test datasets, we aimed to provide a comprehensive assessment of performance of LLMs by evaluating their accuracy and reasoning underlying their differential diagnoses and to compare their accuracy with that of neuroradiology subspecialty-trained radiologists.

MATERIALS AND METHODS

Approval from the Institutional Review Board was waived for this experimental study because it did not involve human subjects. The methodology of our study followed the Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare (MI-CLEAR-LLM) [17-19].

Creating Radiologic Quizzes From *RadioGraphics*

Review articles from the journal *RadioGraphics* covering neuroradiology and head & neck topics, published between October 2020 and February 2024 in *RadioGraphics*, were searched on the journal's website (https://pubs.rsna.org/journal/radiographics). After applying this filter to the search for 'neuroradiology', 55 consecutive articles were identified. Of these, 21 articles were excluded based on the following predefined criteria: 1) Tutorial articles on the technical aspects of neuroradiology (n = 2), 2) Articles focusing on facial bone imaging (n = 4), and 3) Topics unrelated to neuroradiology (n = 15). From the remaining 34 articles, 401 radiological quizzes were created for the independent test dataset. The details of the included and excluded articles are provided in Supplementary Table 1.

Among the 401 quizzes, 47.6% (191 of 401) were brain imaging cases derived from 16 articles, 2.7% (11 of 401) were spine imaging cases from two articles, and 49.6% (199 of 401) were head and neck cases from 16 articles. The quizzes contained 821 images, with magnetic resonance imaging (MRI) being the most common (66.3%, 544 of 821). Clinical history was not provided in 30.9% (124 of 401) of cases after removal of imaging findings and diagnoses from the original figure legends (Table 1).

An experienced radiologist (C.H.S., 13 years of experience in diagnostic radiology, subspecialty training in neuroradiology, and 3 years of research experience in evaluating radiology applications of LLMs), who did not participate as a human reader, created radiologic quizzes for the independent test dataset based on figures and figure legends. The text input for the quizzes consisted of patients' clinical information (age, sex, and clinical history) (Supplementary Table 2). Clinical information was extracted from the figure legends and any description of imaging findings or diagnoses was removed. If no usable clinical

 https://doi.org/10.3348/kjr.2025.1045

Table 1. Characteristics of the included articles and created radiologic quizzes

	No. (%)
Articles (n = 34)	
Brain	16 (47.1)
Spine	2 (5.9)
Head & neck	16 (47.1)
Cases for differential diagnosis (n = 401)	
Age, mean ± SD, yrs	48.2 ± 20.7
Sex	
Male	195 (48.6)
Female	142 (35.4)
Not available	64 (16.0)
Topics	
Brain	191 (47.6)
Spine	11 (2.7)
Head & neck	199 (49.6)
Textual clinical history	
Available	277 (69.1)
Unavailable	124 (30.9)
Total images for imaging information description (n = 821)	
MRI	544 (66.3)
CT	243 (29.6)
Ultrasound	2 (0.2)
X-ray	9 (1.1)
Angiography	8 (1.0)
Nuclear medicine	15 (1.8)

SD = standard deviation

history remained after the removal, no clinical history was provided. The image inputs were extracted from these articles. A flowchart of the study design is depicted in Figure 1.

Using LLM for Answering Radiologic Quizzes

LLMs with vision capabilities were used in our study: GPT-4 Turbo with Vision (GPT-4V; gpt-4-vision-preview) and GPT-4 Omni (GPT-4o; gpt-4o-2024-08-06) by OpenAI, Gemini Flash (gemini-1.5-flash-002, released September 24, 2024) by Google DeepMind, and Claude 3.0 (claude-3-opus-20240229) and Claude 3.5 (claude-3-5-sonnet-20240620) by Anthropic. T0 was selected to produce more deterministic responses, ensuring consistency in the generated differential diagnoses, whereas T1 was used to allow for greater variability and potentially broader diagnostic considerations. GPT-4V, Gemini Flash, and Claude 3.0 were accessed between March 30, 2024, and April 3, 2024. GPT-4o and Claude 3.5 were accessed between July 22, 2024, and July 24, 2024.

The LLMs were tasked with using a single integrated prompt that requested both examination characteristic extraction and differential diagnosis. Specifically, the LLMs were tasked with providing the top three differential diagnoses using the created text and image inputs. Prompts were crafted to provide differential diagnoses by considering both common and rare diseases, along with a detailed rationale for each diagnosis. Additionally, prompts were designed to ensure that the answers provided were not used in actual clinical practice, minimizing the likelihood of response rejection. An experienced radiologist (C.H.S.) checked the differential diagnoses from a single query and calculated their accuracy. The accuracy was calculated as both top-3 accuracy and top-1 accuracy. Top-3 accuracy was defined as correct if the true diagnosis was included among the three differential diagnoses provided, whereas top-1 accuracy required the true diagnosis to match only the first-listed differential diagnoses [3].

To evaluate the ability of the LLM to recognize fundamental imaging information from radiological images, we performed this assessment using a model that demonstrated the highest top-3 diagnostic accuracy in our primary differential diagnosis analysis. This model was tasked with describing the specific examination characteristics of each input image, including the imaging modality, specific sequence, image plane, use of contrast, and scanned body parts. The imaging modalities were categorized as MRI, CT, ultrasound, radiography, angiography, and nuclear medicine. An experienced radiologist (C.H.S.) reviewed all the images. If specific examination characteristics could not be confidently determined from the images, they were excluded from the corresponding accuracy calculations. Additionally, the model was prompted to describe the specific lesion location and relevant imaging findings to further assess its interpretative capability beyond basic imaging information.

All LLM queries were performed using the Application Programming Interface (API) access provided by each vendor. Each quiz question was processed independently using stateless API calls, without any sequential interactions or shared session memories between questions. For prompt development and optimization, we evaluated the full case set, identified aberrant responses, analyzed the error patterns, and iteratively refined the prompts. Given the stateless nature of API calls, this process is distinct from parameter training or information retention, which ensures that it does not influence the model's parameters

Korean Journal of Radiology
KJR

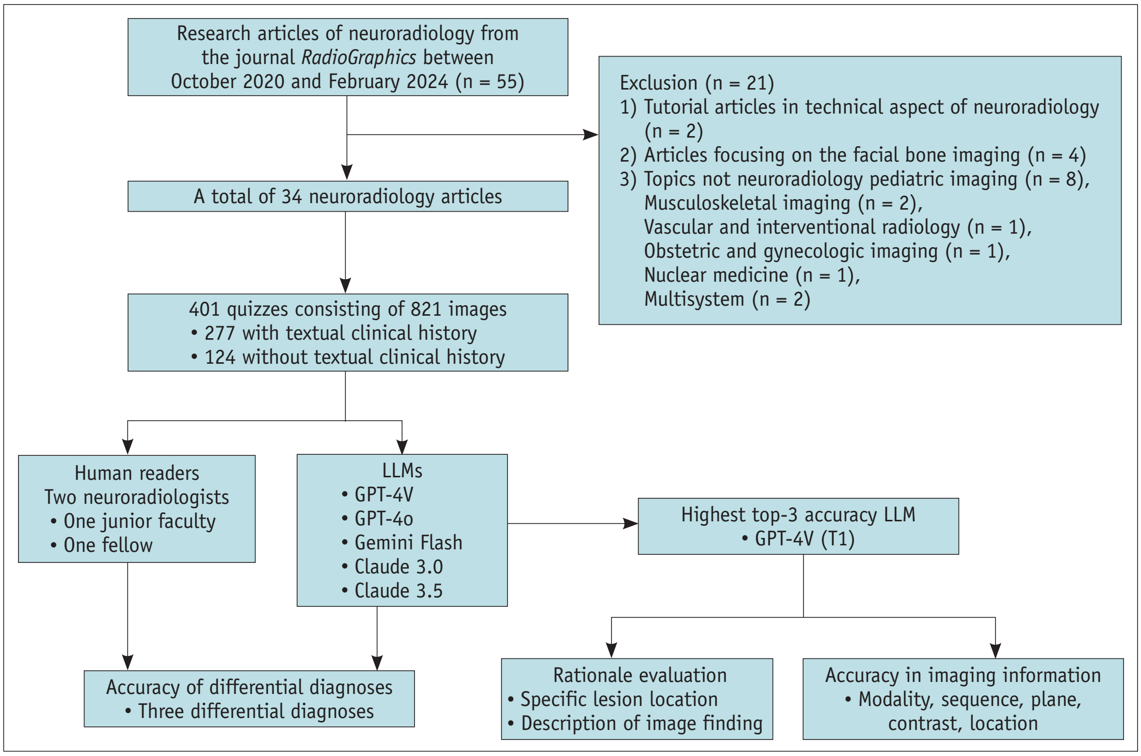

Fig. 1. Flowchart of the study. LLMs = large language models, GPT-4V = GPT-4 Turbo with Vision, GPT-4o = GPT-4 Omni

or outputs. Details of the prompts are presented in the Supplement.

Evaluation by Radiologists

To compare the diagnostic accuracy of LLMs and radiologists, two board-certified neuroradiology subspecialty-trained radiologists (one junior faculty member: P.S.S., 3 years of experience in neuroradiology; one fellow: J.S.K., 2 years of experience in neuroradiology) provided three differential diagnoses based on the same texts and images given to LLMs. The radiologists were blinded to the sources of the cases, which originated from *RadioGraphics*. Additionally, the readers were thoroughly instructed on the guidelines for answering quizzes: 1) Complete one session with breaks every hour; 2) Prohibited from searching for textbooks or the Internet during the session. An experienced radiologist (C.H.S.) checked the answers and calculated the top-3 and top-1 accuracy.

After completing the answering session, two neuroradiologists (P.S.S. and J.S.K.) scored the LLM-generated descriptions of the lesion location, imaging findings, and presence or absence of hallucinations, and each radiologist evaluated half of the cases. An experienced neuroradiologist (C.H.S.) supervised the evaluation. The goal was to evaluate the ability of the LLM to interpret radiological images in a manner comparable to that of a radiologist. Evaluation was conducted on two aspects using four-point scales: 1) "Did the LLM correctly identify the specific location of the lesion?" and 2) "Did the LLM accurately describe the imaging findings?" Specific lesion locations were categorized as follows: 4 = correct, 3 = incorrect but identifiable, 2 = incorrect and unidentifiable, and 1 = not described. Imaging findings were categorized as follows: 4 = correct, 3 = partially correct, 2 = incorrect, and 1 = not described. The details of the four-point scale used to evaluate the ability of the LLM to interpret radiological images are shown in Table 2. If the LLM correctly described the lesion location and correctly or partially correctly described the imaging findings, a precise interpretation was considered. Conversely, if the LLM was incorrect or failed to describe both lesion location and imaging findings, it was considered an inaccurate interpretation. Additionally, the presence of hallucinations unrelated to the lesion location or imaging findings (e.g., errors in clinical information, knowledge of the disease, or examination characteristics) was also assessed. Finally, LLMs' responses that demonstrated precise interpretation without hallucinations

Table 2. Four-point scale for evaluating the LLM's reasoning in interpreting radiologic images

Point	Interpretation	Explanation
Specific lesion location		
4	Correct	Consistent with radiologist's interpretation • The response accurately identifies the lesion location and matches the radiologist's interpretation
3	Incorrect but identifiable	Location present in images • The response identifies an incorrect lesion location, but the specified location is visible and can be identified in the images
2	Incorrect and unidentifiable	Location absent in images • The response identifies a lesion location that is incorrect and completely unidentifiable, with no corresponding structure on the images
1	Not described	No lesion location provided • The response does not provide any lesion location information
Description of imaging findings		
4	Correct	Consistent with radiologist's interpretation • The response accurately identifies all significant imaging findings and aligns closely with the radiologist's interpretation
3	Partially correct	Incomplete or partially consistent with radiologist's interpretation • The response correctly identifies some, but not all, imaging findings. It may miss key imaging findings or provide an incomplete interpretation that only partially consistent with the radiologist's interpretation
2	Incorrect	Misinterpretation or major discrepancies with radiologist's interpretation • The response contains significant errors or reaches a conclusion that significantly diverges from the radiologist's interpretation
1	Not described	Text-based result, no imaging interpretation provided • The response fails to provide any relevant imaging interpretation, instead focusing on text information. No attempt is made to address the imaging findings

LLM = large language model

were classified as acceptable and defined as composite metrics (lesion location = 4; imaging findings = 3–4; no hallucination) [20].

Statistical Analysis

The accuracy of the five LLMs (GPT-4V, GPT-4o, Gemini Flash, Claude 3.0, and Claude 3.5) with two temperature settings and two human readers (junior faculty and fellow) was evaluated and compared using generalized estimating equations to account for repeated measures within the same cases. The LLM performance was assessed for the entire dataset and for predefined subgroups, namely, cases with and without an accompanying textual clinical history. Statistical analyses were performed using SPSS statistical software (version 27.0 for Windows; IBM Corp., Armonk, NY, USA) and MedCalc (version 22.021.; MedCalc Software Ltd, Ostend, Belgium).

RESULTS

Diagnostic Accuracy of LLMs and Human Readers

The accuracies of the LLMs and human readers for differential diagnoses are listed in Table 3. Top-3 accuracy of the LLMs ranged from 29.9% (120 of 401) to 49.4% (198 of 401), with GPT-4V (T1) achieving the highest top-3 accuracy. In the subgroup of cases with textual clinical history, the accuracy of LLM top-3 increased to 36.1%–61.0% (100–169 of 277), whereas in those without a clinical history, it dropped to 2.4%–25.0% (3–31 of 124) (Table 4). The accuracy of the junior faculty was 80.3% (322 of 401), and that of the fellow faculty was 68.3% (274 of 401), both of which significantly outperformed all LLMs (all $P < 0.001$). Additionally, junior faculty showed significantly higher accuracy than fellow faculty ($P < 0.001$).

When considering the accuracy limited to the first differential diagnosis, the top-1 accuracy of the LLMs decreased, ranging from 15.7% (63 of 401) to 29.4% (118 of

Table 3. Accuracy of LLMs and human readers in making differential diagnoses

LLM or reader	Top-3				Top-1			
	Accuracy	95% CI	*P**	*P*†	Accuracy	95% CI	*P**	*P*†
GPT-4V (T0)	45.9 (184/401)	41.1–50.8	<0.001	<0.001	25.4 (102/401)	21.4–29.9	<0.001	<0.001
(T1)	49.4 (198/401)	44.5–54.3	<0.001	<0.001	27.2 (109/401)	23.1–31.7	<0.001	<0.001
GPT-4o (T0)	49.1 (197/401)	44.3–54.0	<0.001	<0.001	29.4 (118/401)	25.2–34.1	<0.001	<0.001
(T1)	46.4 (186/401)	41.6–51.3	<0.001	<0.001	26.2 (105/401)	22.1–30.7	<0.001	<0.001
Gemini Flash (T0)	29.9 (120/401)	25.6–34.6	<0.001	<0.001	15.7 (63/401)	12.5–19.6	<0.001	<0.001
(T1)	31.2 (125/401)	26.8–35.9	<0.001	<0.001	16.7 (67/401)	13.4–20.7	<0.001	<0.001
Claude 3.0 (T0)	31.2 (125/401)	26.8–35.9	<0.001	<0.001	19.5 (78/401)	15.9–23.6	<0.001	<0.001
(T1)	31.9 (128/401)	27.5–36.6	<0.001	<0.001	19.2 (77/401)	15.6–23.4	<0.001	<0.001
Claude 3.5 (T0)	40.6 (163/401)	35.9–45.5	<0.001	<0.001	25.4 (102/401)	21.4–29.9	<0.001	<0.001
(T1)	41.9 (168/401)	37.2–46.8	<0.001	<0.001	26.7 (107/401)	22.6–31.2	<0.001	<0.001
Junior faculty	80.3 (322/401)	76.1–83.9		<0.001	59.6 (239/401)	54.7–64.3		0.08
Fellow	68.3 (274/401)	63.6–72.7	<0.001		53.6 (215/401)	48.7–58.4	0.08	

Data are percentage values with corresponding numerators and denominators in parentheses.
**P*-value for comparison with the junior faculty, †*P*-value for comparison with the fellow.
LLM = large language model, CI = confidence interval, GPT-4V = GPT-4 Turbo with Vision, T0 = temperature 0, T1 = temperature 1, GPT-4o = GPT-4 Omni

Table 4. LLM accuracy based on the presence or absence of textual clinical history

LLM	With textual clinical history		Without textual clinical history	
	Top-3	Top-1	Top-3	Top-1
GPT-4V (T0)	58.5 (162/277)	34.3 (95/277)	17.7 (22/124)	5.6 (7/124)
(T1)	61.0 (169/277)	36.5 (101/277)	23.4 (29/124)	6.4 (8/124)
GPT-4o (T0)	59.9 (166/277)	37.2 (103/277)	25.0 (31/124)	12.1 (15/124)
(T1)	58.8 (163/277)	33.9 (94/277)	18.5 (23/124)	8.9 (11/124)
Gemini Flash (T0)	36.1 (100/277)	19.5 (54/277)	16.1 (20/124)	7.3 (9/124)
(T1)	37.2 (103/277)	19.9 (55/277)	17.7 (22/124)	9.7 (12/124)
Claude 3.0 (T0)	44.0 (122/277)	27.4 (76/277)	2.4 (3/124)	1.6 (2/124)
(T1)	44.0 (122/277)	26.7 (74/277)	4.8 (6/124)	2.4 (3/124)
Claude 3.5 (T0)	52.7 (146/277)	33.2 (92/277)	13.7 (17/124)	8.1 (10/124)
(T1)	53.8 (149/277)	34.7 (96/277)	15.3 (19/124)	8.9 (11/124)

Data are percentage values with corresponding numerators and denominators in parentheses.
LLM = large language model, GPT-4V = GPT-4 Turbo with Vision, T0 = temperature 0, T1 = temperature 1, GPT-4o = GPT-4 Omni

401), with GPT-4o (T0) achieving the highest accuracy. In cases with a textual clinical history, the top-1 accuracy ranged from 19.5% to 37.2% (54–103 of 277), whereas in those without a clinical history, it dropped to 1.6%–12.1% (2–15 of 124). The junior faculty (59.6%, 239 of 401) and the fellow (53.6%, 215 of 401) significantly outperformed all LLMs (all $P < 0.001$). The accuracy of the fellow faculty was not significantly different from that of the junior faculty ($P = 0.08$).

Evaluating LLM's Rationale for Differential Diagnoses

The rationale for the differential diagnoses was assessed using GPT-4V (T1), which achieved the highest diagnostic accuracy. GPT-4V (T1) correctly identified the specific lesion location in 35.2% (141 of 401) of cases, while providing an unidentifiable location from the images in 10.7% (43 of 401) and no location description in 24.7% (99 of 401). GPT-4V (T1) correctly described the imaging findings in 30.4% (122 of 401) and partially correctly described them in 29.2% (117 of 401), imaging findings were omitted entirely, relying solely on text information in 6.2% (25 of 401).

To evaluate the radiological image interpretation in a manner similar to that of radiologists, we assessed a matrix combining lesion location and imaging findings (Fig. 2). GPT-4V (T1) showed precise interpretation in 30.7% of the cases (correct location and imaging findings, 17.5% [70 of 401]; correct location and partially correct imaging findings, 13.2% [53 of 401]). In contrast, inaccurate interpretation

A

Location	Finding: Correct	Partially correct	Incorrect	Not described
Correct	17.5	13.2	3.0	1.5
Incorrect but identifiable	2.2	6.0	19.5	1.7
Incorrect and unidentifiable	0.7	0.7	8.7	0.5
Not described	10.0	9.2	3.0	2.5

B

Location	Finding: Correct	Partially correct	Incorrect	Not described
Correct	33.3	17.7	1.0	2.0
Incorrect but identifiable	2.0	4.0	3.0	0.0
Incorrect and unidentifiable	1.0	0.5	1.0	0.0
Not described	16.7	13.6	1.0	3.0

C

Location	Finding: Correct	Partially correct	Incorrect	Not described
Correct	2.0	8.9	4.9	1.0
Incorrect but identifiable	2.5	7.9	35.5	3.4
Incorrect and unidentifiable	0.5	1.0	16.3	1.0
Not described	3.4	4.9	4.9	2.0

D

Location	Finding: Correct	Partially correct	Incorrect	Not described
Correct	23.1	15.5	2.5	2.2
Incorrect but identifiable	1.8	4.0	11.2	2.5
Incorrect and unidentifiable	0.4	0.0	5.1	0.7
Not described	12.6	11.9	2.9	3.6

E

Location	Finding: Correct	Partially correct	Incorrect	Not described
Correct	4.8	8.1	4.0	0.0
Incorrect but identifiable	3.2	10.5	37.9	0.0
Incorrect and unidentifiable	1.6	2.4	16.9	0.0
Not described	4.0	3.2	3.2	0.0

Fig. 2. Matrix evaluation using a four-point scale for specific lesion location and imaging findings to assess GPT-4V's (temperature 1) reasoning in image interpretation. Yellow squares indicate precise interpretation (describing the correct location and correct/partially correct imaging findings), and red squares indicate inaccurate interpretation (incorrect or not describing both location and imaging findings). **A:** The matrix of overall image interpretation ability shows 30.7% precision and 35.9% inaccuracy. **B, C:** Matrix for cases with correct answers shows 51.0% precise interpretation **(B)** and 63.1% inaccurate interpretation in cases with incorrect answers **(C)**. **D:** The matrix in cases using text inputs with clinical history showed a precise interpretation of 38.6%; however, lesion location and imaging findings were not described in 31.0% and 9.0% of the cases, respectively. **E:** In cases using text inputs without clinical history, 58.1% showed inaccurate interpretations.

was observed in 35.9% (144 of 401). Examples of precise and inaccurate interpretations are shown in Figure 3 and Figure 4, respectively.

The GPT-4V (T1) showed precise interpretation in 51.0% (101 of 198) and inaccurate interpretation in 8.1% (16 of 198) of the cases in which the diagnosis was correct. In cases with incorrect answers, precise interpretation occurred in 10.8% (22 of 203) and inaccurate interpretation occurred in 63.1% (128 of 203). With clinical history, the precise and inaccurate interpretation rates were 38.6% (107 of 277) and 26.0% (72 of 277), whereas without clinical history, these rates were 12.9% (16 of 124) and 58.1% (72 of 124), respectively. GPT-4V (T1) omitted lesion location in 31.0% (86 of 277) and imaging findings in 9.0% (25 of 277) of cases when clinical history was available, whereas omission rates were 10.5% (13 of 124) for lesion location and 0% for imaging findings when no clinical history was provided.

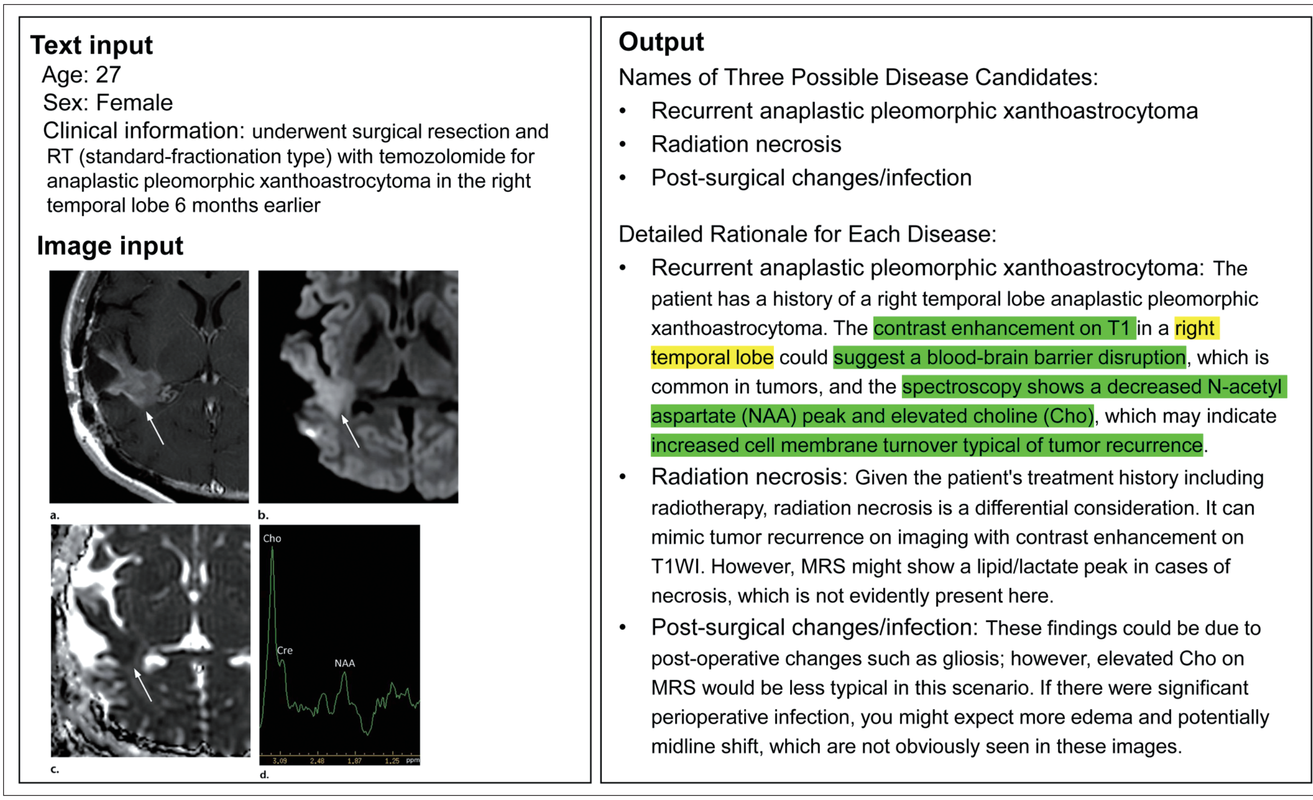

Fig. 3. An example of accurate interpretation by GPT-4V with T1. This radiologic quiz was created based on a study by Katsura et al. [34], featuring a 27-year-old female patient with a history of surgical resection and radiation therapy with temozolomide for anaplastic pleomorphic xanthoastrocytoma in the right temporal lobe. The correct diagnosis is "tumor recurrence." GPT-4V (T1) listed tumor recurrence as the first differential diagnosis, with correct identification of the specific lesion location and imaging findings. Yellow highlights indicate the described lesion location, and green highlights indicate imaging findings. GPT-4V = GPT-4 Turbo with Vision, T1 = temperature 1

Hallucinations were observed in 4.5% of the cases (18 of 401), of which 14 (3.5%) were related to disease knowledge and four (1.0%) were associated with radiologic knowledge. The acceptability of the LLM-generated analyses, defined as cases with precise image interpretation without accompanying hallucinations, was 29.4% (118 of 401) (Supplementary Table 3).

Accuracy for Describing Imaging Information

The accuracy of the imaging information was assessed using the GPT-4V (T1), which achieved the highest diagnostic accuracy. GPT-4V (T1) showed 97.4% (800 of 821) accuracy in identifying the imaging modality and 92.2% (756 of 820) accuracy in identifying the scanned body part. The lowest accuracy was observed in determining the use of contrast medium (64.8%, 529 of 816). The accuracy of the imaging information is shown in Figure 5.

DISCUSSION

In our study, we evaluated the diagnostic accuracy of LLMs and their ability to interpret radiologic images similarly to those of radiologists using radiologic quizzes created by the journal *RadioGraphics*. LLMs underperformed compared to neuroradiology subspecialty-trained radiologists in terms of both top-3 and top-1 accuracies. Although LLMs accurately identify radiologic imaging information from image inputs, their rationale for providing differential diagnoses in terms of interpreting specific lesion locations and imaging findings from radiologic images is unsatisfactory. Their performance decreased further when text inputs without a patient's clinical history were used, and they occasionally failed to describe imaging findings.

Previous studies have evaluated LLMs using open-source cases from medical journals [3,6,21-23], potentially leading to data leakage, limiting the test dataset's independence from the training data of potential models, and raising

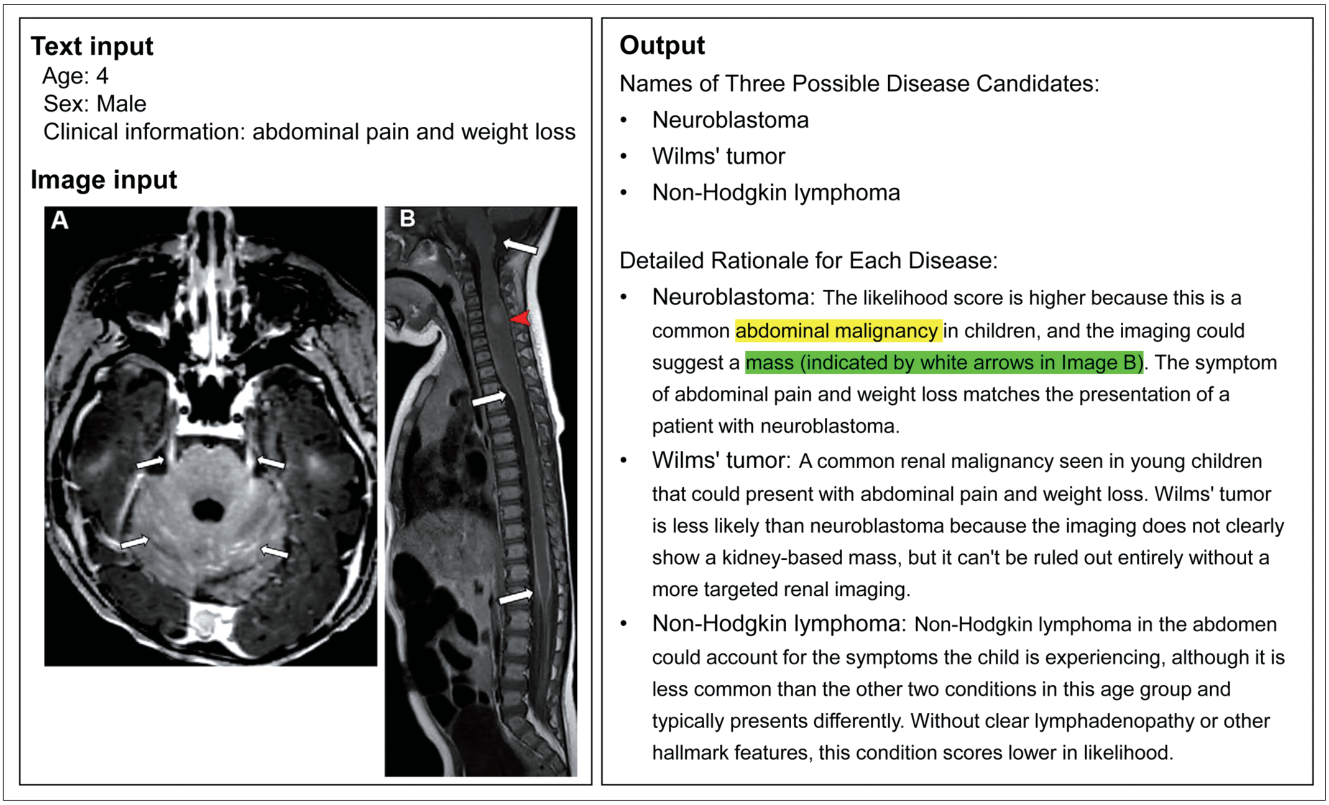

Fig. 4. An example of inaccurate interpretation by GPT-4V with T1. This radiologic quiz was created based on an article by Kurokawa et al. [35]. This case features a 4-year-old male patient presenting with symptoms of abdominal pain and weight loss. The correct diagnosis is "ALK-positive histiocytosis." GPT-4V (T1) provided abdominal malignancies as differential diagnoses based on the patients' symptoms, with incorrect and unidentifiable lesion locations and incorrect imaging findings. Yellow highlights indicate the described lesion location, and green highlights indicate imaging findings. GPT-4V = GPT-4 Turbo with Vision, T1 = temperature 1

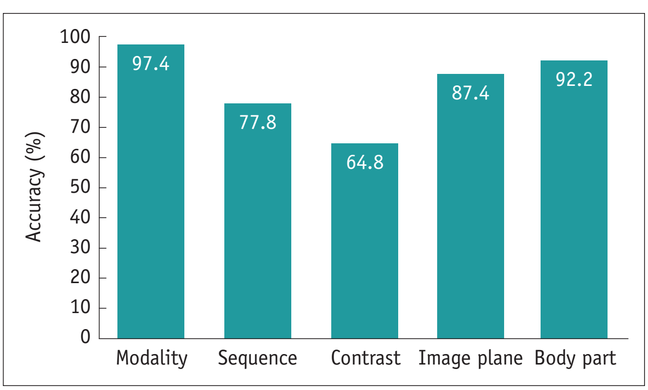

Fig. 5. Accuracy for describing imaging information using GPT-4V with temperature 1. GPT-4V = GPT-4 Turbo with Vision

concerns about overestimating performance of LLMs [17]. A notable strength of our study was that we created radiological quizzes based on cases from *RadioGraphics*, a reputable review-focused journal in the field of radiology. This approach adds value by testing LLMs with cases that are definitely independent of the model's training data, using complex, peer-reviewed cases specifically designed for education and professional development in radiology. These quizzes involved diverse cases that mimicked real clinical scenarios by including patient age, sex, and relevant clinical history, while carefully excluding direct imaging findings or diagnoses to allow LLMs to rely solely on their interpretative abilities. Additionally, LLMs were asked to provide three differential diagnoses in an open-ended format aligned with practical clinical decision making, where multiple potential diagnoses must be considered. By requiring LLMs to generate these differential diagnoses rather than selecting them from multiple choices, our study provides a more realistic assessment of their diagnostic capabilities compared to prior studies that primarily focused on simpler question-answering tasks.

Previous studies have reported that LLMs often struggle to detect abnormalities and perform multistep reasoning in radiology, leading to inaccurate or incomplete interpretations [24-27]. Considering this unreliable vision capability, we went beyond simply evaluating the accuracy of the LLMs' diagnoses by analyzing their reasoning processes in depth. We assessed how accurately the LLMs could identify specific lesion locations and describe imaging findings from the provided radiologic image inputs using a four-point scale, allowing us to determine whether the LLMs were performing like human radiologists or were merely guessing based on textual input. In our evaluation, GPT-4V (T1), which achieved the highest diagnostic accuracy among the LLMs, still struggled with core image-interpretation tasks, frequently misidentifying lesion locations or imaging findings, and producing hallucinations. Consequently, fewer than one-third of responses were considered acceptable. Similarly, GPT-4o and Gemini 1.5 Pro failed to describe imaging abnormalities in more than 80% when solving Diagnosis Please cases from the *Radiology* journal [8] and fabricated imaging findings in 62.6% of real-world cases [16]. These findings emphasize that although LLMs are becoming increasingly proficient in understanding radiologic imaging information, they still lack the nuanced interpretative skills required for precise diagnosis and decision-making by human radiologists.

In addition, we assessed the impact of clinical history on the interpretive performance of the LLMs. This study included cases with incomplete information, specifically those lacking clinical history after the removal of imaging findings or diagnoses from the figure legends, to evaluate the impact of textual input. We found that when clinical history was included in the text input, GPT-4V (T1) showed a higher rate of top-3 accuracy and more precise image interpretation than when clinical history was absent. However, in some instances, LLMs neglect to describe lesion location or imaging findings altogether, instead relying solely on the clinical history provided. This observation suggests that LLMs tend to overrely on textual input rather than performing a detailed analysis of the images themselves. This overdependence on text input aligns with previous reports indicating that the performance of LLMs, such as GPT-4V, can vary significantly depending on the nature and amount of textual data provided [6,28]. A recent meta-analysis reported that the GPT performed significantly better with text-based inputs (67.1%) than with text-and-image inputs (46.5%) [12]. So far, radiologic images alone, without textual descriptions of imaging findings or medical history, have yielded markedly low diagnostic performance [8,29]. These findings highlight the current limitations of LLMs in real-world clinical practice, where the integration of both clinical and imaging data is essential for an accurate diagnosis. Moreover, LLMs are unlikely to replace radiologists.

Our study has several limitations. First, although we generated 401 quiz questions based on 34 *RadioGraphics* articles, the dataset was limited in size and scope. In particular, key real-world imaging modalities, such as angiography, perfusion imaging, and three-dimensional reconstruction, were not represented because these imaging types were largely absent from the *RadioGraphics* figures used to construct the dataset. This structural limitation prevented the evaluation of LLM performance on modalities that are essential in routine clinical decision-making and may render the study less reflective of real-world radiological practice. Furthermore, because all the cases were derived from an educational journal, they may not fully capture the complexity and variability of routine radiological practices. Second, we did not apply or assess diverse prompt strategies, such as chain-of-thought or few-shot learning, to evaluate their potential impact on the rationale behind the generated differential diagnoses. While this absence represents a methodological limitation that may have underestimated LLM performance, recent studies suggest that such advanced prompting techniques may not meaningfully enhance the performance of current multimodal LLMs in radiological interpretation because of their limited visual reasoning capabilities and continued reliance on textual information [30-33]. Third, our study focused on neuroradiology subspecialty-trained radiologists, which limits the generalizability of our findings to radiologists with different levels of expertise or those from other subspecialties. Moreover, the diagnostic evaluations were performed in a controlled, retrospective setting rather than in real-time clinical workflows. Therefore, the performance and utility of LLMs in actual clinical environments, where time constraints, incomplete information, and complex decision-making processes play critical roles, remain to be validated. Fourth, although we minimized the risk of data leakage by directly creating quizzes from *RadioGraphics* cases, the figure legends and some text are accessible online, meaning that the possibility of leakage could not be completely eliminated. Given that figures may be searchable via common internet tools and that subscription journal content can sometimes be accessed through unofficial routes or publicly posted copies, the risk of leakage may be greater than assumed. If such leakage occurs, the relatively low diagnostic performance we observed suggests that the limitations of current LLMs in radiologic interpretation may be even more substantial than those demonstrated in this study. Fifth, one critical consideration is that the

LLMs evaluated in our study were not explicitly built or fine-tuned with a radiology-specific knowledge base or designed as agents for this task. Therefore, the diagnostic accuracy and interpretative reasoning observed in this study should be interpreted with caution because the evaluated LLMs without radiology-specific fine-tuning may not have fully demonstrated their true potential for image interpretation and differential diagnosis. Additionally, our study did not include a direct comparison with medical imaging–specialized AI models or radiology-specific LLMs, which could have provided a more balanced context for interpreting the performance of general-purpose LLMs. Further studies should explore the diagnostic capabilities of LLMs when adapted or fine-tuned specifically for radiology-related tasks, by incorporating structured radiology datasets and domain-specific training to enhance reasoning and interpretative performance. Sixth, the assessment of repeatability in the LLM responses was not planned a priori and could not be performed retrospectively because most proprietary models underwent updates during the study period. This is a clear methodological shortcoming of this study. Finally, because this study involved a large number of comparisons, no multiple comparison correction was applied in keeping with its exploratory nature, which may have increased the risk of type I errors and should be considered when interpreting the results.

In conclusion, LLMs markedly underperformed compared with neuroradiologists and demonstrated unsatisfactory reasoning for their differential diagnoses, with performance declining further in cases without a textual clinical history. These findings emphasize the limitations of the current multimodal LLMs in neuroradiological interpretation and their reliance on textual inputs.

Supplement

The Supplement is available with this article at https://doi.org/10.3348/kjr.2025.1045.

Availability of Data and Material

The datasets generated or analyzed during the study are included in this published article and its supplement.

Conflicts of Interest

Chong Hyun Suh, an 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: Chong Hyun Suh, Woo Hyun Shim. Data curation: Woo Hyun Shim, Hwon Heo. Formal analysis: Woo Hyun Shim, Hwon Heo. Funding acquisition: Chong Hyun Suh. Investigation: Pae Sun Suh, Ji Su Ko. Methodology: Pae Sun Suh, Woo Hyun Shim, Chong Hyun Suh. Project administration: Chang-Yun Woo, Hyungjun Park, Chong Hyun Suh. Resources: Woo Hyun Shim. Software: Woo Hyun Shim, Hwon Heo. Supervision: Chong Hyun Suh. Validation: Pae Sun Suh, Ji Su Ko. Visualization: Pae Sun Suh, Ji Su Ko. Writing—original draft: Pae Sun Suh, Ji Su Ko, Hwon Heo, Chong Hyun Suh. Writing—review & editing: Pae Sun Suh, Ji Su Ko, Chong Hyun Suh.

ORCID IDs

Pae Sun Suh
https://orcid.org/0000-0002-8618-9558
Ji Su Ko
https://orcid.org/0000-0001-6589-2431
Woo Hyun Shim
https://orcid.org/0000-0002-7251-2916
Hwon Heo
https://orcid.org/0000-0002-6103-4680
Chang-Yun Woo
https://orcid.org/0000-0002-8286-8481
Hyungjun Park
https://orcid.org/0000-0002-5523-5144
Chong Hyun Suh
https://orcid.org/0000-0002-4737-0530

Funding Statement

This research was supported by a grant (2025IT0007) from the Asan Institute for Life Sciences, Asan Medical Center, Seoul, Korea.

REFERENCES

1. Akinci D'Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol* 2024;30:80-90
2. Gertz RJ, Bunck AC, Lennartz S, Dratsch T, Iuga AI, Maintz D, et al. GPT-4 for automated determination of radiologic study and protocol based on radiology request forms: a feasibility study. *Radiology* 2023;307:e230877
3. 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
4. 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
5. Horiuchi D, Tatekawa H, Oura T, Shimono T, Walston SL, Takita H, et al. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. *Eur Radiol* 2025;35:506-516
6. 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
7. 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
8. 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
9. 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
10. 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
11. Sun SH, Chen K, Anavim S, Phillipi M, Yeh L, Huynh K, et al. Large language models with vision on diagnostic radiology board exam style questions. *Acad Radiol* 2025;32:3096-3102
12. 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
13. Agbareia R, Omar M, Soffer S, Glicksberg BS, Nadkarni GN, Klang E. Visual-textual integration in LLMs for medical diagnosis: a preliminary quantitative analysis. *Comput Struct Biotechnol J* 2024;27:184-189
14. Wu SH, Tong WJ, Li MD, Hu HT, Lu XZ, Huang ZR, et al. Collaborative enhancement of consistency and accuracy in US diagnosis of thyroid nodules using large language models. *Radiology* 2024;310:e232255
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. 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
17. Park SH, Suh CH, Lee JH, Kahn CE, Moy L. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). *Korean J Radiol* 2024;25:865-868
18. Park SH, Suh CH. Reporting guidelines for artificial intelligence studies in healthcare (for both conventional and large language models): what's new in 2024. *Korean J Radiol* 2024;25:687-690
19. 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
20. Hong EK, Ham J, Roh B, Gu J, Park B, Kang S, et al. Diagnostic accuracy and clinical value of a domain-specific multimodal generative AI model for chest radiograph report generation. *Radiology* 2025;314:e241476
21. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. *Radiology* 2023;308:e231040
22. Li D, Gupta K, Bhaduri M, Sathiadoss P, Bhatnagar S, Chong J. Comparing GPT-3.5 and GPT-4 accuracy and drift in Radiology diagnosis please cases. *Radiology* 2024;310:e232411
23. Han T, Adams LC, Bressem KK, Busch F, Nebelung S, Truhn D. Comparative analysis of multimodal large language model performance on clinical vignette questions. *JAMA* 2024;331:1320-1321
24. Yan Z, Zhang K, Zhou R, He L, Li X, Sun L. Multimodal ChatGPT for medical applications: an experimental study of GPT-4V. arXiv [Preprint]. 2023 [accessed on May 10, 2025]. Available at: https://doi.org/10.48550/arXiv.2310.19061
25. Wu C, Lei J, Zheng Q, Zhao W, Lin W, Zhang X, et al. Can GPT-4V(ision) serve medical applications? Case studies on GPT-4V for multimodal medical diagnosis. arXiv [Preprint]. 2023 [accessed on May 10, 2025]. Available at: https://doi.org/10.48550/arXiv.2310.09909
26. Liu Y, Li Y, Wang Z, Liang X, Liu L, Wang L, et al. A systematic evaluation of GPT-4V's multimodal capability for chest X-ray image analysis. *Meta Radiol* 2024;2:100099
27. Strotzer QD, Nieberle F, Kupke LS, Napodano G, Muertz AK, Meiler S, et al. Toward foundation models in radiology? Quantitative assessment of GPT-4V's multimodal and multianatomic region capabilities. *Radiology* 2024;313:e240955
28. Waisberg E, Ong J, Masalkhi M, Zaman N, Sarker P, Lee AG, et al. GPT-4 and medical image analysis: strengths, weaknesses and future directions. *J Med Artif Intell* 2023;6:29
29. 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
30. Hong J, Chen T, Luo L, Zheng C, Xu T, Yu H, et al. Benchmarking the thinking mode of multimodal large language models in clinical tasks. arXiv [Preprint]. 2025 [accessed on November 26, 2025]. Available at: https://doi.org/10.48550/arXiv.2511.03328
31. Takita H, Walston SL, Mitsuyama Y, Watanabe K, Ishimaru S,

Ueda D. Comparative performance of large language models in structuring head CT radiology reports: multi-institutional validation study in Japan. *Jpn J Radiol* 2025;43:1445-1455
32. 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
33. 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
34. Katsura M, Sato J, Akahane M, Furuta T, Mori H, Abe O. Recognizing radiation-induced changes in the central nervous system: where to look and what to look for. *Radiographics* 2021;41:224-248
35. Kurokawa R, Kurokawa M, Isshiki S, Harada T, Nakaya M, Baba A, et al. Dural and leptomeningeal diseases: anatomy, causes, and neuroimaging findings. *Radiographics* 2023;43:e230039