



Diagnostic accuracy and consistency of ChatGPT-4o in radiology: influence of image, clinical data, and answer options on performance

Kadir Atakır
 Kaan Işın
 Abdullah Taş
 Hakan Önder

University of Health Sciences Türkiye, Prof. Dr. Cemil Taşcıoğlu City Hospital, Clinic of Radiology, İstanbul, Türkiye

PURPOSE

This study aimed to evaluate the diagnostic accuracy of Chat Generative Pre-trained Transformer (ChatGPT) version 4 Omni (ChatGPT-4o) in radiology across seven information input combinations (image, clinical data, and multiple-choice options) to assess the consistency of its outputs across repeated trials and to compare its performance with that of human radiologists.

METHODS

We tested 129 distinct radiology cases under seven input conditions (varying presence of imaging, clinical context, and answer options). Each case was processed by ChatGPT-4o for seven different input combinations on three separate accounts. Diagnostic accuracy was determined by comparison with ground-truth diagnoses, and interobserver consistency was measured using Fleiss' kappa. Pairwise comparisons were performed with the Wilcoxon signed-rank test. Additionally, the same set of cases was evaluated by nine radiology residents to benchmark ChatGPT-4o's performance against human diagnostic accuracy.

RESULTS

ChatGPT-4o's diagnostic accuracy was lowest for "image only" (19.90%) and "options only" (20.67%) conditions. The highest accuracy was observed in "image + clinical information + options" (80.88%) and "clinical information + options" (75.45%) conditions. The highest interobserver agreement was observed in the "image + clinical information + options" condition ($\kappa = 0.733$) and the lowest was in the "options only" condition ($\kappa = 0.023$), suggesting that more information improves consistency. However, there was no effective benefit of adding imaging data over already provided clinical data and options, as seen in post-hoc analysis. In human comparison, ChatGPT-4o outperformed radiology residents in text-based configurations (75.45% vs. 42.89%), whereas residents showed slightly better performance in image-based tasks (64.13% vs. 61.24%). Notably, when residents were allowed to use ChatGPT-4o as a support tool, their image-based diagnostic accuracy increased from 63.04% to 74.16%.

CONCLUSION

ChatGPT-4o performs well when provided with rich textual input but remains limited in purely image-based diagnoses. Its accuracy and consistency increase with multimodal input, yet adding imaging does not significantly improve performance beyond clinical context and diagnostic options alone. The model's superior performance to residents in text-based tasks underscores its potential as a diagnostic aid in structured scenarios. Furthermore, its integration as a support tool may enhance human diagnostic accuracy, particularly in image-based interpretation.

CLINICAL SIGNIFICANCE

Although ChatGPT-4o is not yet capable of reliably interpreting radiologic images on its own, it demonstrates strong performance in text-based diagnostic reasoning. Its integration into clinical workflows—particularly for triage, structured decision support, or educational purposes—may augment radiologists' diagnostic capacity and consistency.

KEYWORDS

Artificial intelligence, large language models, ChatGPT, diagnostic accuracy, observer variation, clinical decision support systems, medical imaging, multimodal input

Handling editor: Burak Kocak

Corresponding author: Abdullah Taş

E-mail: atas997@gmail.com

Received 14 May 2025; revision requested 08 June 2025;
last revision received 28 July 2025; accepted 04 August 2025.



Epub: 22.09.2025

Publication date: 01.09.2026

DOI: 10.4274/dir.2025.253460

You may cite this article as: Atakır K, Işın K, Taş A, Önder H. Diagnostic accuracy and consistency of ChatGPT-4o in radiology: influence of image, clinical data, and answer options on performance. *Diagn Interv Radiol.* 2026;32(5):555-564.

Artificial intelligence (AI) has emerged as a transformative tool in radiology. It offers the potential to organize diagnostic workflows, improve clinical decision-making, and reduce workload on radiologists. Among the different AI tools, large language models (LLMs), such as OpenAI's Chat Generative Pre-trained Transformer (ChatGPT), have shown marked capabilities in the medical field; for example, they can assist students in passing medical examinations and can support clinical decision-making.¹ In radiology, early applications of ChatGPT (ChatGPT-3.5 and 4) are variable and include activities such as data mining, structuring radiology reports, and answering text-based board-like examination questions.^{2,3} A recent study showed that ChatGPT-4 outperformed the previous version (ChatGPT-3.5) on written radiology questions.⁴ One of the latest updates of ChatGPT, ChatGPT-4 with vision (ChatGPT-4V), added the capability to evaluate images, which draws attention to applying this feature in radiological image interpretation. However, there have been few studies about ChatGPT-4V's accuracy in radiological image interpretation. Initial findings suggest that there is a considerable performance gap between the evaluation of images and the accomplishment of text-based tasks.^{2,5}

ChatGPT was originally developed as a text-based model. With the release of ChatGPT-4V, it gained the ability to interpret visual inputs. This became a turning point, as the model was able to process and analyze medical images along with clinical data. Multimodal functionality allows ChatGPT to open new opportunities for integration into radiology workflows and is prompting

researchers to evaluate its diagnostic performance.

Although numerous studies report high diagnostic accuracy (>80%) for LLMs in text-based tasks,^{1,6} only a limited number have demonstrated comparable performance in image-based scenarios; even then, this has been under highly controlled conditions, such as narrow diagnostic targets and optimized prompts.⁷ Conversely, many investigations have highlighted major limitations in "image only" scenarios, particularly in complex modalities, such as brain magnetic resonance imaging (MRI), where performance drops markedly in the absence of clinical context.^{4,8} Studies evaluating ChatGPT-4V against radiologists have revealed mixed outcomes, with some reporting comparable accuracy, whereas others found substantial underperformance by the model.^{9,10} This variability underscores the importance of input structure in multimodal tasks.

A growing body of literature emphasizes that prompt design, including differential diagnoses, clinical details, and structured questions, greatly influences LLM output.^{11,12} This variability raises concerns regarding reproducibility and reliability in diagnostic settings, where standardization is critical.^{3,13} Taken together, current evidence suggests that LLM performance in radiologic diagnosis is highly dependent on how and what information is provided. However, most prior studies have focused on isolated scenarios or specific modalities, without systematically comparing different input combinations under controlled conditions.^{14,15} Our study addresses this gap by evaluating the diagnostic accuracy and consistency of ChatGPT-4o across seven distinct input formats, ranging from image-only to full contextual prompts with multiple-choice options.

This study evaluates the diagnostic accuracy and consistency of ChatGPT-4o across varying structured input combinations and compares its performance with radiology residents under matched conditions. This benchmark is essential to contextualize the model's capabilities relative to human expertise and assess its potential clinical applicability.

Methods

Study designs and objective

This study aimed to evaluate ChatGPT-4o's diagnostic decision-making about radiological images and its potential role in real-life

diagnoses when provided with clinical information and differential diagnosis options. The objective was to investigate the radiological images' contribution in ChatGPT-4o's decision-making process and to understand the possible integration of ChatGPT for daily radiological tasks and workflows. An overview of the workflow is presented in Figure 1. Ethics committee approval was not required, as the study did not involve patient data or identifiable personal information. The participation of radiology residents was conducted in an educational setting and did not involve any intervention or collection of sensitive data. Informed consent was obtained from all participants prior to data collection. This study was conducted in compliance with the Minimum Reporting Items for Clear Evaluation of Accuracy Reports of LLMs in healthcare guidelines.¹⁶

Dataset

A total of 129 radiology cases were selected from two publicly available, peer-reviewed sources: the New England Journal of Medicine (NEJM) Image Challenge and the Radiological Society of North America (RSNA) Case Collection.^{17,18} The NEJM cases that matched the study format and included a radiological image were included in full. From the RSNA Case Collection, only cases published in 2023 that followed the same format were selected. The inclusion criteria required the presence of a diagnostic-quality radiological image (in single-slice static JPEG format) and a clearly defined radiologic diagnosis. Cases were excluded if they lacked radiological imaging (e.g., pathological or macroscopic photographs), if the diagnosis was uncertain, or if the case format did not conform to the five-option multiple-choice structure. Both sources were chosen for their educational value, structured content, and diagnostic clarity. The resulting dataset was designed to reflect a diverse range of real-world clinical scenarios rather than stylized board-examination questions. A detailed breakdown of cases by imaging modality and anatomical system is available in Supplementary Material 1.

Experimental setup and prompt design

To evaluate the impact of varying prompt components on the diagnostic performance of a large multimodal language model (ChatGPT-4o), a systematic prompt engineering framework was implemented inspired by recent studies on multimodal AI in radiology.^{2,11} A total of 129 radiological cases were compiled, each consisting of three main ele-

Main points

- Chat Generative Pre-trained Transformer version 4 Omni (ChatGPT-4o) achieved the highest diagnostic accuracy when clinical information and diagnostic options were provided and the lowest in the "image only" condition.
- Interobserver consistency improved with richer input combinations but was minimal in the "options only" setting, indicating a near-random output.
- Adding radiologic images to an already informative clinical prompt did not significantly enhance accuracy, highlighting current limitations in image interpretation.
- Compared with radiology residents, ChatGPT-4o outperformed humans in text-based tasks, whereas residents maintained an advantage in image-only diagnoses.

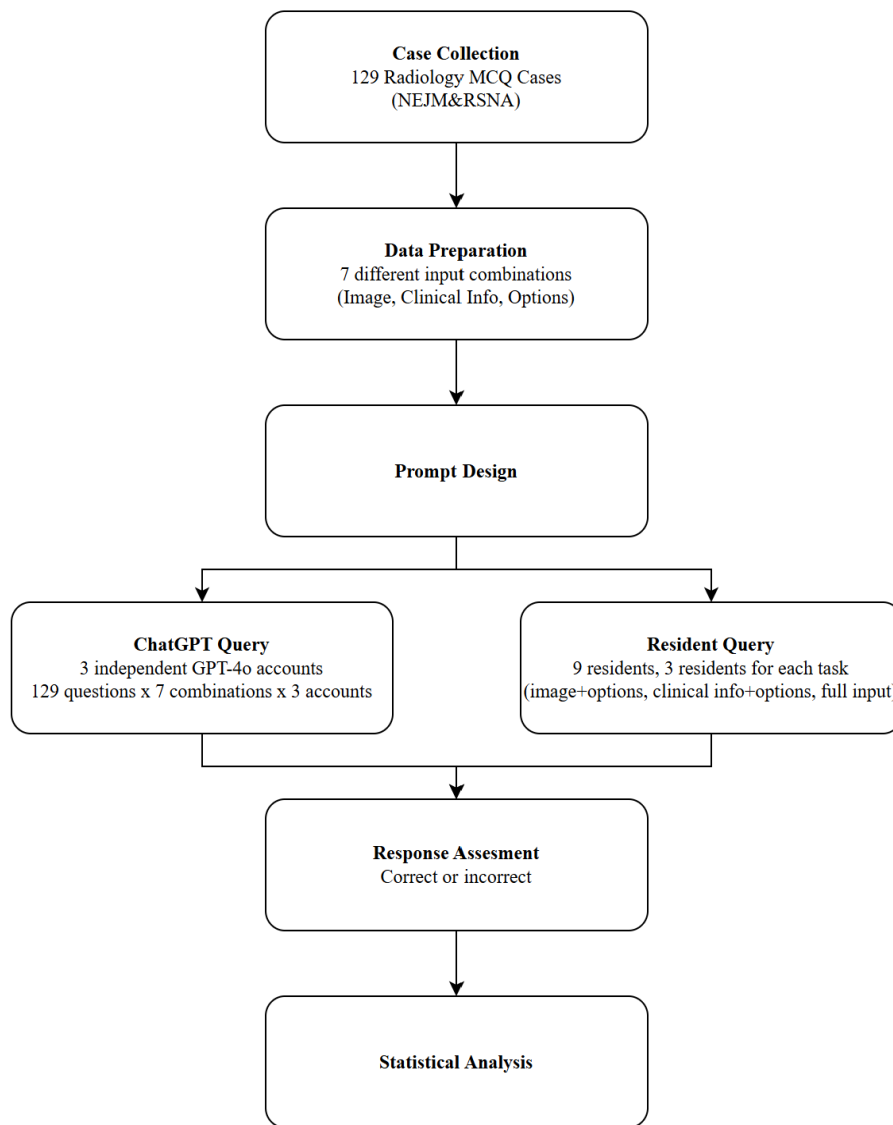


Figure 1. Flowchart of the study. MCQ, multiple choice question; NEJM, New England Journal of Medicine; RSNA, Radiological Society of North America; GPT-4o, Generative Pre-trained Transformer 4 Omni.

ments: a diagnostic-quality radiology image, structured clinical information, and five multiple-choice diagnostic options with a single correct answer. For each case, seven distinct prompt types were constructed by presenting the model with different combinations of these elements:

1. Radiological image only
2. Clinical information only
3. Differential diagnosis options only
4. Radiological image + clinical information
5. Radiological image + diagnostic options
6. Clinical information + diagnostic options

7. Radiological image + clinical information + diagnostic options

Each of these seven combinations was queried using ChatGPT-4o, accessed via the ChatGPT web interface (version released April 2024).

To explore how different types of input influence diagnostic performance, seven distinct prompt configurations were designed by selectively combining radiological images, clinical context, and diagnostic options. Each combination was chosen to isolate specific aspects of the model's reasoning process. For example, the "image only" prompt evaluates pure visual interpretation, whereas the "clinical only" prompt focuses on text-based decision-making. The "options only" condition was included to examine how the model performs without any

contextual information, relying solely on its internal knowledge and prior associations between disease names. This also helps reveal whether certain choices are favored due to frequency bias or random selection, offering a useful reference point for interpreting improvements seen in more informed conditions. Additionally, combinations of these inputs examine how multimodal integration affects performance. Together, these configurations allow for a structured comparison of how each input type contributes to both accuracy and consistency.

Prompts were automatically generated and submitted sequentially via a controlled Python-based interface. In all scenarios, the model was instructed to either choose the best-fitting diagnosis or provide a diagnostic decision based on the given input. The same base prompt format (e.g., "You will be provided with a radiologic image and associated clinical information. Based on both inputs, determine the most likely diagnosis...") was used to maintain consistency across conditions. No prompt engineering was performed beyond standardization for content structure and clarity. All interactions were conducted using the default ChatGPT interface settings, including default temperature and sampling parameters. All designed prompts are listed in Supplementary Material 2. The model was instructed to base its answers only on the information provided in each prompt. For each prompt, the model responded (Figures 2-5). The answers were then reviewed, and a record was made of whether the answer was correct. The model answered each question without knowledge of whether the previous answers were correct or not. All prior case context was then discarded before proceeding to the next prompt. Thus, the model was designed to approach each question independently. This design allowed for the isolated assessment of each input component's influence on diagnostic accuracy and ensured methodological reproducibility across all prompt combinations.

Repetition and observer simulation

First, to evaluate interobserver variability, all case combinations were submitted independently using three different ChatGPT-4o accounts, each accessed via a distinct computer and browser instance to eliminate any session overlap or caching artifacts. This resulted in a total of 2,709 model-generated responses (129 cases × 7 combinations × 3 accounts).

Answers were recorded as either correct (+) or incorrect (−), based on the alignment of

the model's final answer with the reference diagnosis provided in the source case material. In cases where multiple plausible answers were generated, responses were marked incorrect if the correct answer was not unambiguously prioritized.

Diagnostic comparison with radiology residents

To assess the performance of ChatGPT-4o, a supplementary analysis was conducted involving radiology residents to provide a benchmark for human diagnostic performance. Specifically, nine residents (post-graduate year 3 or 4 level) were recruited and randomly divided into three separate

groups, each assigned to one of the following input configurations: (1) image + options, (2) clinical information + options, and (3) image + clinical information + options. Among the seven configurations tested with ChatGPT-4o, only three were selected for human comparison, as these represented clinically meaningful scenarios; configurations lacking diagnostic options were excluded due to their anticipated low accuracy and limited comparative value. Each participant independently reviewed the same set of 129 cases in their assigned condition and submitted their diagnostic answers. In the second round (conducted 1 week later), they were allowed to use ChatGPT-4o as a supportive

tool and could revise their answers accordingly. This comparison is intended to explore the baseline performance of trained human readers and the potential additive value of LLM-generated input.

Data recording and analysis

All responses were manually recorded in structured Microsoft Excel spreadsheets (Microsoft Corporation, Redmond, WA, USA) by the research team. The accuracy rate (%) for each combination and account was calculated.

Statistical analysis

Descriptive statistics were used to calculate the mean diagnostic accuracy across all input combinations and ChatGPT-4o accounts. The distribution of variables was assessed with the Shapiro–Wilk test, revealing non-normal distributions for all combinations. For this reason, non-parametric tests were employed for further analysis. Fleiss' kappa was used to evaluate interobserver agreement across the three independent ChatGPT-4o accounts for each input combination. Pairwise comparisons between combinations were conducted using the Wilcoxon signed-rank test. All analyses were performed using the SPSS version 27.0 (IBM, USA) package, and statistical significance was set at $P < 0.05$.

Results

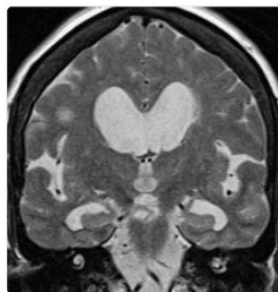
A descriptive analysis of diagnostic accuracy across the three independent ChatGPT-4o accounts revealed different patterns based on the type and combination of input information (Table 1). The lowest mean accuracies were observed when the model was presented with diagnostic options alone (20.67%) and with “image only” inputs (19.90%). In contrast, performance significantly improved with the inclusion of multiple input types. The “clinical information + options” configuration achieved a mean accuracy of 75.45%, whereas the “image + clinical information + options” combination yielded the highest overall performance at 80.88%. Accuracy scores were relatively consistent across accounts, indicating stable model behavior when sufficient contextual information was available. These results highlight the importance of multimodal input for maximizing diagnostic accuracy and reliability in LLMs within radiology.

To determine the distribution characteristics of the diagnostic accuracy data, the Shapiro–Wilk test was performed on case-level

ChatGPT 4o ▾

You will be presented with a radiologic image. Based solely on this image, determine the most likely diagnosis. Do not assume any clinical information or consider answer choices. Justify your decision based on visible imaging findings only.

Sure, please upload the radiologic image you'd like me to evaluate.



The coronal T2-weighted MRI image reveals **marked dilation of the lateral ventricles**, particularly involving the frontal horns, with **periventricular hyperintensities** suggestive of transependymal CSF flow. There is **no evidence of mass effect or obstructive lesion** at the foramen of Monro, third ventricle, or aqueduct visible on this slice.

Most Likely Diagnosis:

Normal Pressure Hydrocephalus (NPH)

Justification Based on Imaging:

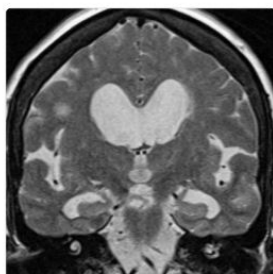
- **Ventriculomegaly** out of proportion to cortical atrophy (disproportionately enlarged ventricles).
- **Rounded configuration of the frontal horns** of the lateral ventricles.
- **Periventricular T2 hyperintensities** indicating interstitial edema due to CSF transudation.
- Absence of cortical sulcal effacement or mass effect, ruling out acute obstructive hydrocephalus or mass lesion.

This imaging pattern is classic for NPH, though definitive diagnosis requires clinical correlation and response to CSF diversion.

Figure 2. ChatGPT-4o query for the “image only” condition. No clinical information or answer options are provided. MRI, magnetic resonance imaging; CSF, cerebrospinal fluid.

You will be provided with a radiologic image and associated clinical information. Based on both inputs, determine the most likely diagnosis. Do not use any external answer choices. Explain your reasoning by integrating imaging and clinical features.

Of course—please upload the radiologic image and provide the associated clinical information so I can begin the analysis.



A 70-year-old woman presented to the neurology clinic with a 2-year history of gait disturbance, cognitive impairment, and urinary incontinence. What is the diagnosis?

The clinical triad of **gait disturbance**, **cognitive impairment**, and **urinary incontinence** in an elderly patient strongly suggests **normal pressure hydrocephalus (NPH)**.

Radiologic Findings:

The MRI (T2-weighted coronal image) demonstrates:

- **Ventriculomegaly:** Markedly enlarged lateral ventricles, especially the frontal horns, without corresponding cortical atrophy.
- **Disproportionately enlarged subarachnoid space hydrocephalus (DESH)** pattern:
 - Enlarged Sylvian fissures.
 - Narrowed high-convexity sulci, indicating compression.
- No mass lesion or obvious obstruction at the level of the aqueduct or foramen of Monro.

Figure 3. ChatGPT-4o query for radiological image + clinical information condition. No options are given. ChatGPT-4o, Chat Generative Pre-trained Transformer; MRI, magnetic resonance imaging.

mean accuracy scores for each of the seven input combinations. The results revealed that none of the distributions conformed to normality (all $P < 0.001$). This finding supports the use of non-parametric statistical methods, such as Fleiss' kappa and the Wilcoxon signed-rank test, for further analyses.

To assess interobserver consistency across different ChatGPT-4o accounts, Fleiss' kappa was calculated for each input combination (Table 2). The interobserver agreement varied considerably, with the highest consistency observed in the "image + clinical + options" condition ($\kappa = 0.733$), indicating substantial agreement. Conversely, the "options only" condition resulted in the lowest agreement ($\kappa = 0.023$), suggesting high variability in model outputs when contextual

information was not provided. Intermediate levels of agreement were seen in the "image + clinical" ($\kappa = 0.606$) and "image + options" ($\kappa = 0.652$) groups. These findings further emphasize the stabilizing effect of rich contextual input on the model's behavior.

Post-hoc comparisons using Wilcoxon signed-rank tests revealed statistically significant differences in diagnostic accuracy between most input combinations ($P < 0.001$) (Figure 6). Both "image only" and "options only" configurations performed significantly worse than any multimodal combination. No statistically significant difference was observed between "image + clinical" and "image + options" ($P = 0.051$) or between "clinical + options" and "image + clinical + options" ($P = 0.058$), indicating diminishing marginal

returns from image inclusion when comprehensive clinical context and differential options are already available. Together, these results emphasize the dominant contribution of clinical information and the synergistic value of multimodal input in enhancing diagnostic accuracy and reproducibility of LLM-driven image interpretation.

Although these results reflect intramodel variability, human-level diagnostic performance was further assessed for context. A benchmarking analysis was conducted involving nine radiology residents across the same 129-case dataset, grouped by input condition (Figure 7). Given the exploratory nature and limited size of the human comparator group, statistical inference was intentionally avoided to prevent overinterpretation. In the "image + options" scenario, the mean accuracy of residents (63.04%) was marginally higher than that of ChatGPT-4o (61.24%), suggesting that human readers retain a slight advantage in visual interpretation. However, when residents had access to ChatGPT-4o's image-based outputs, their accuracy improved markedly to 74.16%, indicating a clear benefit from AI-assisted interpretation. In the "clinical information + options" condition, ChatGPT-4o substantially outperformed the residents (75.45% vs. 42.89%). When residents were allowed to use ChatGPT-4o before answering, their performance rose significantly (74.93%), almost matching that of the model. In the full-input configuration (image + clinical + options), ChatGPT-4o again outperformed the resident group (80.88% vs. 71.57%), and the combination of both yielded the highest accuracy overall (83.47%). In all input combinations, ChatGPT-4o access provided significant benefit to residents, especially in structured clinical decision-making scenarios.

Discussion

The findings of this study provide several key insights into the diagnostic capabilities and limitations of LLMs, particularly ChatGPT-4o, within the context of radiological image interpretation. First, diagnostic performance was found to be remarkably limited when the model was presented with radiological images or diagnostic options in isolation. The "image only" and "options only" combinations resulted in the lowest mean accuracies, at 19.90% and 20.67%, respectively. In contrast, the inclusion of clinical information led to a marked improvement in performance, especially when paired with diagnostic options. This pattern may reflect the inherent design of instruction-tuned LLMs, which perform

You will receive clinical information and a list of possible diagnoses. Select the most likely diagnosis based only on the clinical details and the listed options. Do not incorporate any imaging data in your reasoning.

Understood. Please provide the clinical information and the list of diagnostic options.

A 70-year-old woman presented to the neurology clinic with a 2-year history of gait disturbance, cognitive impairment, and urinary incontinence. What is the diagnosis?

- a. Normal-pressure hydrocephalus
- b. Alzheimer's disease
- c. Parkinson's disease
- d. Benign intracranial hypertension
- e. Multiple system atrophy

The most likely diagnosis is:

a. Normal-pressure hydrocephalus (NPH)

Reasoning:

This patient presents with the classic triad of **gait disturbance, cognitive impairment, and urinary incontinence**, which is strongly suggestive of **normal-pressure hydrocephalus**. This triad is often summarized as:

- **"Wet"** – urinary incontinence
- **"Wobbly"** – gait disturbance (often magnetic gait)
- **"Wacky"** – cognitive decline

While other conditions like Alzheimer's disease (cognitive dominant), Parkinson's disease (bradykinesia, tremor), and multiple system atrophy (parkinsonism with autonomic features) may share overlapping features, none present as prominently with this triad.

Benign intracranial hypertension (idiopathic intracranial hypertension) is not typical in this demographic and usually presents with headache and visual symptoms rather than gait and cognitive issues.

demonstrated poor reproducibility across different accounts. Interestingly, although the addition of radiological images to the "clinical information + options" input did not improve accuracy, it did enhance interobserver agreement. This suggests that image input, although not increasing correctness, may contribute to more consistent model behavior across different interactions, thereby stabilizing diagnostic output when the clinical context is already rich.

As part of this study, we also conducted a supplementary analysis involving radiology residents to contextualize the performance of ChatGPT-4o. Doing so allowed for indirect insights into question difficulty, the impact of image and clinical information on decision-making, and the presence of guiding clinical hints embedded in the cases. Residents demonstrated relatively stronger performance in image-provided scenarios, reflecting their visual training and possible reliance on pattern recognition developed through clinical exposure; this contrasted with ChatGPT-4o, which performed particularly well on text-based tasks when clinical information was provided, likely due to its broad access to medical knowledge and its advanced capacity for textual reasoning. When residents were allowed to consult ChatGPT during interpretation, their accuracy improved further, especially when all inputs—images, clinical context, and diagnostic options—were available. Interestingly, in the "image + options" configuration, residents assisted by ChatGPT were able to outperform the model itself, whereas in other combinations, they barely matched its performance. This observation suggests that, at their current stage, LLMs may enhance human diagnostic accuracy in visually anchored tasks by offering auxiliary reasoning pathways or differential diagnosis support, rather than serving as standalone diagnosticians.

Although the presence of implicit diagnostic cues in clinical descriptions (commonly referred to as the "hint effect") may have contributed to the model's performance, this alone is unlikely to account for the substantial gap observed between the model and human readers, as radiology residents would also notice these cues. The markedly higher accuracy of ChatGPT-4o compared with residents in the "clinical information + options" scenario (75.45% vs. 42.79%) suggests that the model may be leveraging subtle linguistic patterns that elude human interpretation, as noted by earlier studies.¹⁹

Figure 4. ChatGPT-4o query for clinical information + diagnostic options condition. No image is given. ChatGPT-4o, Chat Generative Pre-trained Transformer.

best when decision space is constrained and explicitly framed. This relationship between input structuring and model performance is further illustrated when examining how different combinations interact. Although adding clinical information to the "image + options" condition significantly improved accuracy, the reverse scenario—adding images to the "clinical information + options" setting—did not yield a statistically significant gain. This asymmetry may indicate that ChatGPT-4o derives proportionally more diagnostic value from structured textual data than from single-slice imaging inputs, consistent with its language-dominant architecture. Furthermore, although a ceiling effect may have limited further gains in the "clinical

information + options" setting, the findings suggest that ChatGPT-4o's ability to interpret radiologic images independently remains limited. Accuracy was lowest with "image only" input, and adding images to text-based prompts did not yield significant improvements. These results indicate that the model relies more on structured text than on visual reasoning, which may be restricted by current architectural constraints.

Interobserver agreement analyses demonstrated that the highest Fleiss' kappa scores were observed in multimodal combinations, particularly when clinical information and diagnostic options were included. In contrast, scenarios lacking a clinical context

Consistent with prior research, our study found that the combination of clinical information and diagnostic options yielded relatively high diagnostic accuracy among

the tested input configurations, whereas the addition of radiological images to this setting did not lead to a significant improvement. This outcome mirrors the findings of Li et al.², who reported enhanced diagnostic performance when clinical context was added to imaging input. Similarly, Güneş et al.²⁰ and Horiuchi et al.²¹ observed substantially higher accuracy in text-based tasks than in image-based ones, highlighting the model's reliance on structured language. Elek et al.²² also demonstrated that although ChatGPT could recognize anatomical structures on computed tomography and MRI scans, its diagnostic accuracy remained low in the absence of guiding clinical information. These findings suggest that when structured textual context is available, visual input contributes only marginally to the model's decision-making process. Collectively, this evidence supports the presence of a text-biased processing pattern in current LLMs and underscores their suitability for structured clinical reasoning tasks, particularly those involving multiple-choice formats. Moreover, other studies have reported high accuracy rates in text-based radiology tasks that incorporate imaging findings as written input.^{21,23,24} Nevertheless, this strength does not generalize to all text-based formats. For instance, Sonoda et al.²⁵ compared ChatGPT-4o, Claude 3 Opus, and Gemini 1.5 Pro, and found that ChatGPT-4o performed poorly compared with Claude 3 Opus, even when provided with richly structured radiology findings. Similarly, studies by Fervers et al.²⁶ and Perchik et al.²⁷ have shown low model performance in narrative or open-ended scenarios, indicating that LLM effectiveness in radiology is strongly influenced by task design, question format, and the structure of the input prompt.

While the addition of radiological images did not significantly enhance diagnostic accuracy, our findings revealed a notable improvement in interobserver agreement with the inclusion of visual input. This suggests that image data, although insufficient

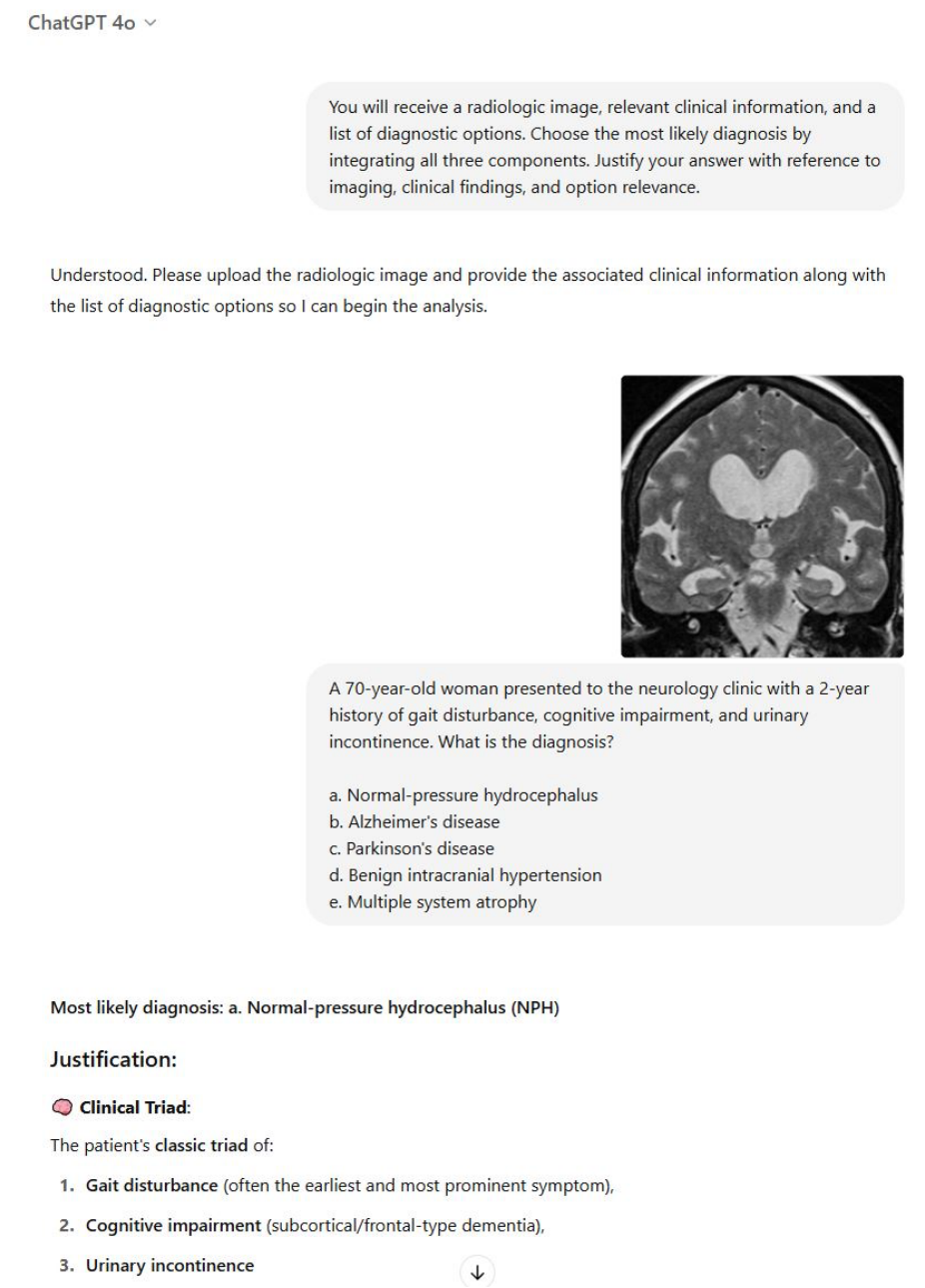


Figure 5. ChatGPT-4o query with radiological images, clinical information, and diagnostic options provided. ChatGPT-4o, Chat Generative Pre-trained Transformer.

Table 1. Accuracy rates of distinct accounts for each of the different input combinations				
Input combination	Account ¹	Account ²	Account ³	Average
Radiological image	19.38%	17.05%	23.26%	19.90%
Clinical information	41.86%	42.64%	49.61%	44.70%
Diagnostic options	10.85%	31.01%	20.16%	20.67%
Radiological image + clinical information	53.49%	50.39%	54.26%	52.71%
Radiological image + diagnostic options	59.69%	62.02%	62.02%	61.24%
Clinical information + diagnostic options	72.87%	76.74%	76.74%	75.45%
Radiological image + clinical information + diagnostic options	80.62%	79.07%	82.95%	80.88%

to improve accuracy, may help stabilize the model's internal decision processes across repeated runs. In our study, Fleiss' kappa scores were consistently higher in multimodal settings that combined images with clinical information and diagnostic options, indicating greater consistency across different

model instances. Similarly, Schramm et al.¹¹ reported that ChatGPT-4V produced more consistent responses in complex neuroimaging tasks when prompted with images supported by textual annotations. In contrast, Krishna et al.³ found that repeated prompts with identical text-only input often yielded

inconsistent outputs; however, they lacked investigations into image-included inputs. To our knowledge, this study is among the first to examine intramodel consistency systematically in a radiologic diagnostic setting.

Several studies have utilized open-access radiology datasets to evaluate LLM performance,^{4,28,29} raising the question of whether high accuracy in certain settings reflects genuine reasoning or simple memorization. In our study, the consistently low diagnostic accuracy in the "image only" condition—despite the use of publicly available cases—combined with the marked improvement following the addition of clinical context, strongly argues against memorization. These findings support the view that ChatGPT-4o relies on contextual understanding rather than the retrieval of previously encountered content.

Our comparative analysis involving radiology residents provided additional insight into the complementary strengths of human expertise and LLM-based diagnostic reasoning. Human readers outperformed the model in "image only" tasks, likely due to their training in visual pattern recognition—an area where current LLMs still fall short. Horiuchi et al.²¹ similarly found that ChatGPT did not reach the performance level of either residents or board-certified radiologists in challenging neuroradiology cases, consistent with our findings. Notably, our study also showed that ChatGPT-4o, when used as a supportive tool, improved the diagnostic accuracy of residents. In contrast, Mukherjee et al.⁴ reported no significant benefit when ChatGPT-4V was used to support human readers. This discrepancy may be explained by methodological differences; their study employed an earlier, less capable version of the model and included a smaller number of diagnostic cases, which may have limited statistical power to detect performance improvements.

Our study did not formally stratify question difficulty; however, the heterogeneous nature of our dataset inherently encompassed a wide range of case complexity. We did not label individual items as "easy" or "difficult"; however, the diagnostic accuracy achieved by radiology residents offers an indirect measure of relative difficulty. Recent studies have emphasized that LLMs, including ChatGPT-4 variants, tend to perform substantially worse on complex or ambiguous cases. For example, Horiuchi et al.²¹ reported significantly reduced LLM performance in challenging diagnostic tasks, highlighting

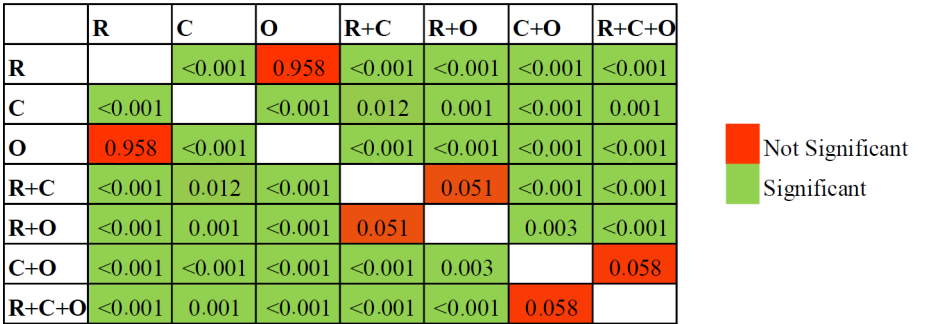


Figure 6. Wilcoxon signed rank test results for post-hoc pairwise comparisons. R, radiological image; C, clinical information; O, diagnostic options. Green: statistically significant ($P < 0.05$); Red: not significant.

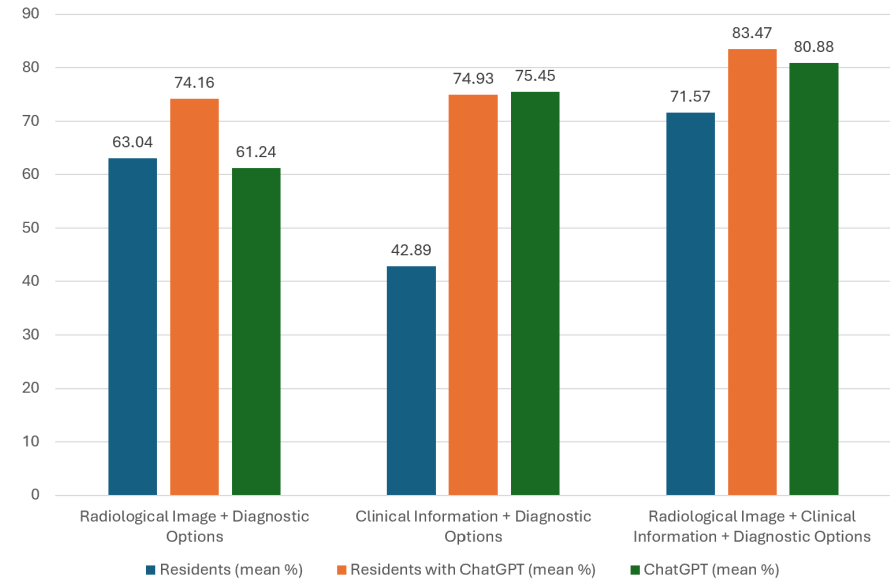


Figure 7. Mean diagnostic accuracy (%) of radiology residents, ChatGPT-4o, and residents with access to ChatGPT-4o responses ("wChatGPT") across three input configurations: "radiological image + diagnostic options," "clinical information + diagnostic options," and "radiological image + clinical information + diagnostic options." ChatGPT-4o outperformed residents in all settings except the "image only" evaluation, where residents had a slight advantage. In all conditions, residents showed a marked improvement when aided by ChatGPT-4o responses. ChatGPT-4o, Chat Generative Pre-trained Transformer.

Table 2. Interobserver agreement rates between accounts	
Input combination	Fleiss' kappa
Radiological image	0.497
Clinical information	0.613
Diagnostic options	0.023
Radiological image + clinical information	0.606
Radiological image + diagnostic options	0.652
Clinical information + diagnostic options	0.540
Radiological image + clinical information + diagnostic options	0.733

Notes: The Fleiss' kappa table shows interobserver variability for different input combinations. Substantial agreement is seen in the "radiological images + clinical information + options" combination.

persistent limitations under cognitive strain. Similarly, Bhayana et al.¹ identified question complexity as a key determinant of model accuracy. Some authors have further suggested that model performance can be optimized by adjusting temperature settings and prompt design strategies in accordance with question difficulty.³⁰ Collectively, these findings underscore the need for future benchmarking efforts to incorporate standardized complexity metrics to better evaluate LLM capabilities across varying levels of clinical challenge.

In comparison with prior literature, our study makes a unique contribution by systematically evaluating ChatGPT-4o's diagnostic accuracy across seven controlled input combinations and three independent repetitions. Unlike previous studies that focused on narrow clinical scenarios or language-specific settings, our methodology deliberately isolated each input type and assessed reproducibility through repeated testing.^{4,25} This design enabled a comprehensive evaluation of both diagnostic performance and output consistency, alongside a direct human comparison. Additionally, we included an input configuration containing only diagnostic options, without clinical context or image input. In contrast to our main input types, this allowed us to observe frequency bias and random selection tendencies in the model's decision-making indirectly—an aspect that, to our knowledge, has not been systematically explored in prior LLM-radiology studies. Moreover, none of the prior studies has made a post-hoc analysis regarding the inclusion of visual data input to already-provided clinical information and answer options. As a subtle but important difference, our study found that the addition of image data to clinical information and diagnostic options input did not improve accuracy; this contrasts with some prior assumptions suggesting that multimodal inputs would enhance performance, and, thus, offers a more nuanced understanding of ChatGPT's current limitations.

Overall, these findings highlight that although LLMs currently show strong potential in text-based reasoning tasks, their ability to interpret radiologic images independently remains limited. Although prior studies have reported mixed results regarding their role in radiologic decision-making, particularly in multimodal settings, it is evident that these models are still in the early stages of adapting to the complexities of visual diagnostic reasoning. As such, they may be more appropriately positioned as assistive tools, rather than autonomous diagnostic agents.

Several studies have suggested that LLMs can be beneficial when incorporated into structured decision-support systems, where clinical context is clearly defined and human oversight is preserved.^{2,14,28} However, their diagnostic precision declines markedly in the absence of textual input, reflecting a strong reliance on language-based information. This dependency raises important concerns about the safety, reproducibility, and clinical applicability of these models in radiological practice.

This study has several notable strengths. First, it employed a comprehensive and controlled experimental design that systematically evaluates ChatGPT-4o's diagnostic performance across seven different input combinations. This allowed us to assess not only diagnostic accuracy but also how the model responds to varying levels of contextual information. Second, by using three independent accounts for each scenario, we were able to measure interobserver variability, a rarely addressed aspect in previous LLM-focused radiology research. Third, the use of peer-reviewed multiple-choice diagnostic radiology questions enhances the clinical validity of our findings. Notably, to our knowledge, this is the first study to assess a prompt configuration that includes only diagnostic options, without accompanying clinical or imaging data, in the context of LLM performance in radiology. Finally, the study was designed and reported in accordance with the Minimum Information about Clinical Artificial Intelligence Modeling-LLM checklist,¹⁶ ensuring methodological rigor, transparency, and alignment with emerging standards in LLM evaluation.

Nevertheless, this study has several limitations. The radiological images provided to the model were static and isolated from the dynamic, interactive environment that is typical of clinical practice. Furthermore, even though the prompts were designed to isolate the effect of each input type, the interpretability of ChatGPT-4o's internal reasoning remains unclear. Finally, the forced-choice format used in the diagnostic tasks may not reflect the complexity of real-life radiological interpretation and reporting, which often involves open-ended reasoning and the integration of longitudinal patient data. Although case difficulty was not formally stratified to avoid subjective grading and maintain study focus, the comparison with radiology residents may serve as an indirect benchmark for overall difficulty, as previously discussed. In addition, although our dataset included a broad spectrum of

imaging modalities and anatomical regions, the case distribution was not balanced across these categories, potentially introducing subtle biases in modality or subspecialty representation. Furthermore, this study focused exclusively on ChatGPT-4o—currently among the most advanced and accessible multimodal LLMs—to ensure consistency in performance benchmarking under controlled conditions. Future studies should explore whether similar patterns hold across different models to assess the generalizability of our findings.

As LLMs continue to evolve, their integration into radiology is expected to become increasingly specialized and clinically aligned. A central priority for future development is improving the visual processing capabilities of these models through domain-specific multimodal training. Some studies underscore the potential of task-specified pre-trained models that are fine-tuned on radiology-specific image-text pairs, rather than relying solely on general-purpose architectures, such as ChatGPT.³¹ Another key area for advancement is the role of prompt engineering and fine-tuning strategies. Although our study used zero-shot prompting without any domain-specific customization, growing evidence suggests that carefully designed prompts and iterative fine-tuning can greatly influence diagnostic performance and consistency.^{9,12,32,33} As such, future research should systematically investigate how these factors can be optimized to balance model flexibility with clinical reliability, especially in high-stakes decision-making.

Beyond diagnostic reasoning, LLMs may also contribute to broader aspects of radiology practice, including structured reporting, generating patient-friendly summaries, prioritizing studies based on urgency, and integrating clinical context into triage or decision-support systems. Hybrid approaches that combine the interpretive strengths of LLMs with image-based AI tools or human oversight may provide a safer and more effective pathway toward real-world deployment. Finally, the use of LLMs in medical education deserves further exploration, particularly in simulation-based diagnostics and real-time AI-assisted feedback for trainees.

In conclusion, this study demonstrates that ChatGPT-4o's diagnostic accuracy in radiology is highly dependent on the presence of contextual information, particularly clinical data and structured diagnostic options. Its limited performance in "image only" scenarios and the lack of added value when ra-

diological images are paired with rich textual input emphasize the model's current limitations in autonomous radiological image interpretation. Even though ChatGPT-4o shows promise as a supportive tool in structured clinical environments—especially for triage or decision support—it cannot yet replace human expertise in image-driven diagnostic tasks. Future advancements must address the need for integrated, multimodal training and transparent reasoning mechanisms before these models can be safely adopted for real-world clinical radiology applications.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary Materials 1 and 2:
<https://d2v96xfpocvxx.cloudfront.net/bda9171a-fae8-4995-8276-2138323f1e16/content-images/dfefbd7e-f7a2-43da-927b-e889250128eb.pdf>

References

- Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology*. 2023;307(5):230582. [Crossref]
- 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(1):232411. [Crossref]
- Krishna S, Bhambra N, Bleakney R, Bhayana R. Evaluation of reliability, repeatability, robustness, and confidence of GPT-3.5 and GPT-4 on a radiology board-style examination. *Radiology*. 2024;311(2):232715. [Crossref]
- Mukherjee P, Hou B, Suri A, et al. Evaluation of GPT large language model performance on RSNA 2023 case of the day questions. *Radiology*. 2024;313(1):240609. [Crossref]
- Hayden N, Gilbert S, Poisson LM, Griffith B, Klocho C. Performance of GPT-4 with vision on text- and image-based ACR diagnostic radiology in-training examination questions. *Radiology*. 2024;312(3):240153. [Crossref]
- Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology*. 2023;307(4):230424. [Crossref]
- Nguyen D, Rao A, Mazumder A, Succi MD. Exploring the accuracy of embedded ChatGPT-4 and ChatGPT-4o in generating BI-RADS scores: a pilot study in radiologic clinical support. *Clin Imaging*. 2025;117:110335. [Crossref]
- Doddi S, Hibshman T, Salichs O, et al. Assessing appropriate responses to ACR urologic imaging scenarios using ChatGPT and Bard. *Curr Probl Diagn Radiol*. 2024;53(2):226-229. [Crossref]
- 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):240273. [Crossref]
- Brin D, Sorin V, Barash Y, et al. Assessing GPT-4 multimodal performance in radiological image analysis. *Eur Radiol*. 2024;35(4):1959-1965. [Crossref]
- Schramm S, Preis S, Metz MC, et al. Impact of multimodal prompt elements on diagnostic performance of GPT-4V in challenging brain MRI cases. *Radiology*. 2025;314(1):240689. [Crossref]
- Kim TT, Makutonin M, Sirous R, Javan R. Optimizing large language models in radiology and mitigating pitfalls: prompt engineering and fine-tuning. *RadioGraphics*. 2025;45(4):240073. [Crossref]
- Kao JP, Kao HT. Large language models in radiology: a technical and clinical perspective. *Eur J Radiol Artif Intell*. 2025;2:100021. [Crossref]
- Kuzan BN, Meşe İ, Yaşar S, Kuzan TY. A retrospective evaluation of the potential of ChatGPT in the accurate diagnosis of acute stroke. *Diagn Interv Radiol*. 2025;31(3):187-195. [Crossref]
- Kahalian S, Rajabzadeh M, Öçbe M, Medisoglu MS. ChatGPT-4.0 in oral and maxillofacial radiology: prediction of anatomical and pathological conditions from radiographic images. *Folia Med (Plovdiv)*. 2024;66(6):863-868. [Crossref]
- 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(10):865-868. [Crossref]
- The New England Journal of Medicine. June 26, 2025 [Internet]. Boston (MA): Massachusetts Medical Society; 2025 [cited 2025 May 10]. [Crossref]
- Home | RSNA Case Collection. Accessed May 10, 2025. [Crossref]
- Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology*. 2024;310(1):232756. [Crossref]
- Güneş YC, Cesur T, Çamur E, Günbey Karabekmez L. Evaluating text and visual diagnostic capabilities of large language models on questions related to the Breast Imaging Reporting and Data System Atlas 5th edition. *Diagn Interv Radiol*. 2025;31(2):111-129. [Crossref]
- Horiuchi D, Tatekawa H, Oura T, 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(4):779-787. [Crossref]
- Elek A, Ekizaloğlu DD, Güler E. Evaluating Microsoft Bing with ChatGPT-4 for the assessment of abdominal computed tomography and magnetic resonance images. *Diagn Interv Radiol*. 2025;31(3):196-205. [Crossref]
- Mitsuyama Y, Tatekawa H, Takita H, et al. Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors. *Eur Radiol*. 2024;35(4):1938-1947. [Crossref]
- Kim SH, Schramm S, Adams LC, et al. Benchmarking the diagnostic performance of open source LLMs in 1933 Eurorad case reports. *NPJ Digit Med*. 2025;8(1):97. [Crossref]
- Sonoda Y, Kurokawa R, Nakamura Y, et al. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" cases. *Jpn J Radiol*. 2024;42(11):1231-1235. [Crossref]
- Fervers P, Hahnfeldt R, Kottlors J, et al. ChatGPT yields low accuracy in determining LI-RADS scores based on free-text and structured radiology reports in German language. *Front Radiol*. 2024;4:1390774. [Crossref]
- Perchik J, Godwin R, West J, Summerlin D, Zahid M, Smith A. Does ChatGPT pass the LIRADS test? Comparing quality of AI generated impressions to human reports. *Jpn J Gastroenterol*. 2023;10(5):1-5. [Crossref]
- Ueda D, Mitsuyama Y, Takita H, et al. Diagnostic Performance of ChatGPT from patient history and imaging findings on the diagnosis please quizzes. *Radiology*. 2023;308(1):231040. [Crossref]
- Suthar PP, Kounsai A, Chhetri L, Saini D, Dua SG. Artificial Intelligence (AI) in Radiology: a deep dive into ChatGPT 4.0's Accuracy with the American Journal of Neuroradiology's (AJNR) "case of the month." *Cureus*. 2023;15(8):43958. [Crossref]
- Milad D, Antaki F, Milad J, et al. Assessing the medical reasoning skills of GPT-4 in complex ophthalmology cases. *Br J Ophthalmol*. 2024;108(10):1398-1405. [Crossref]
- Lee RW, Lee KH, Yun JS, Kim MS, Choi HS. Comparative analysis of M4CXR, an LLM-based chest X-ray report generation model, and ChatGPT in radiological interpretation. *J Clin Med*. 2024;13(23):7057. [Crossref]
- 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):220-231. [Crossref]
- Ray PP, Majumder P. ChatGPT in radiology: a deeper look into its limitations and potential pathways for improvement. *Can Assoc Radiol J*. 2024;75(1):202. [Crossref]