Beyond Single Diagnostic Accuracy: Three-Level Hierarchical Evaluation of Accuracy and Reliability in Multimodal LLM for Ultrasound Interpretation

Taewon Han, Jaeseung Shin✉

Department of Radiology and Center for Imaging Sciences, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
dr.shinjs@gmail.com

Abstract. Multimodal large language models (LLMs) are increasingly explored for radiology applications, yet standard benchmarks typically report single diagnostic accuracy without assessing whether models produce consistent reasoning across repeated queries. We propose a Three-Level Hierarchical Evaluation Framework that sequentially assesses (1) anatomical location, (2) imaging feature description, and (3) diagnosis, measuring both correctness and cross-repetition consistency at each level. A hierarchical gating mechanism advances cases only when both criteria are satisfied, quantifying the fraction of cases achieving "reliably correct understanding." Three state-of-the-art multimodal LLMs—Claude Opus 4.6, Gemini 3 Pro, and Kimi K2.5—were evaluated on 121 ultrasound cases under image-only, text-only, and image + text conditions, each repeated three times (N = 3,267 total responses). While ungated diagnostic accuracy reached up to 43.5%, the full hierarchical gate pass rate never exceeded 8.3%. Within-condition similarity analysis revealed that image-only inputs yielded the lowest response reproducibility, whereas text-only inputs exhibited the highest stability, suggesting current vision encoders struggle with the inherent noise and visual complexity of ultrasound images, potentially making image inputs a primary source of response instability. These findings demonstrate that conventional metrics substantially overestimate the clinical reliability of multimodal LLMs, underscoring that hierarchical frameworks jointly assessing accuracy and reproducible reasoning must become standard in medical image interpretation benchmarking.

Keywords: Multimodal Large language models · Radiology · Ultrasound · Hierarchical Evaluation · Consistency · Reliability.

1 Introduction

Multimodal large language models (LLMs) have demonstrated remarkable capabilities across diverse domains, and their application to medical image interpretation—particularly in radiology—has attracted substantial research interest [1]. Early studies assessed general-purpose models on radiology board-style questions, typically reporting top-1 or top-3 accuracy on multiple-choice or free text formats [2-4]. More recent

studies have examined multimodal models capable of directly processing radiological images alongside clinical text, demonstrating promising diagnostic accuracy on radiological question-answering benchmarks, fueling speculation that these models may eventually augment or even substitute human expertise in specific diagnostic tasks [5-8]. However, particularly in highly complex and operator-dependent modalities like ultrasound, a critical limitation of current evaluation is their near-exclusive reliance on single-instance correctness [7]. This superficial metric creates a profound illusion of competence, masking potential clinical hazards [9,10].

This conventional evaluation strategy overlooks a fundamental requirement for clinical deployment: reproducibility [9-12]. In radiology, a diagnostic tool that produces the correct answer on one occasion but yields a discordant—potentially erroneous—response when queried again under identical conditions cannot be considered reliable, regardless of its aggregate accuracy. The inherent stochastic variance of LLMs introduces a level of response variability that standard top-K accuracy metrics fail to capture [12-14]. Recent work has highlighted the necessity of repeatability assessments for LLM-based clinical tools, demonstrating that run-to-run variability remains high even when performance appears acceptable [9-15].

Moreover, existing benchmarks typically evaluate only diagnostic accuracy, neglecting the soundness of the underlying reasoning process [16]. A model might arrive at the correct final diagnosis despite incorrect anatomical localization or inappropriate feature characterization [16-19]. Such "right for the wrong reasons" outputs represent flawed reasoning pathways that are unacceptable in real clinical practice. To ensure genuine medical image understanding, it is imperative to move beyond flat evaluations and adopt a hierarchical evaluation that rigorously verifies the correctness of intermediate clinical steps—from anatomy to morphology—prior to evaluating the final diagnostic conclusion [16-22].

To address these limitations, we propose a Three-Level Hierarchical Evaluation Framework that jointly assesses accuracy and consistency at three sequential levels: (Level1, L1) anatomical localization, (Level2, L2) imaging feature description, and (Level3, L3) final diagnosis. A hierarchical gating mechanism requires that models satisfy both correctness and cross-repetition consistency criteria at each level before advancing to the next, thereby distinguishing models that achieve reliably correct understanding from those that merely produce correct responses.

Our contributions: (1) we introduce a hierarchical evaluation benchmark with gating that quantifies the reliability gap between single diagnostic accuracy and reproducible correct output, establishing a minimum standard for evaluating multimodal LLMs in ultrasound medical image interpretation; (2) we provide factorial design evidence (image-only vs. text-only vs. image + text) demonstrating that image inputs can serve as a source of response instability, reducing reproducibility relative to text-based conditions in multimodal LLMs; and (3) we provide empirical evidence that conventional single-instance accuracy metrics overestimate the clinical reliability of current models, underscoring the need for reliability-aware evaluation standards in medical artificial intelligence.

2 Methods

2.1 Study Design and Dataset

This study employed a retrospective experimental design to evaluate whether multimodal LLMs achieve genuine understanding of ultrasound medical imaging. To assess response stability, each model was queried three independent times per case under each experimental condition.

All ultrasound cases publicly available on the Korean Society of Ultrasound in Medicine digital platform (https://www.ultrasound.or.kr/) were initially considered, yielding an initial pool of 329 cases collected between July 28, 2000 and January 25, 2026. After excluding 16 cases with poor image quality, 1 video-based case, 10 non-diagnostic cases, and 181 cases involving non-ultrasound imaging modalities, 121 cases were included for analysis. These cases span diverse anatomical regions and representative pathologies in clinical ultrasound practice. Each case consists of a representative ultrasound image, accompanying short clinical information, and an expert-verified ground-truth answer. No patient-identifiable information was accessed or utilized at any stage of this study; all data were obtained in their publicly available, de-identified form.

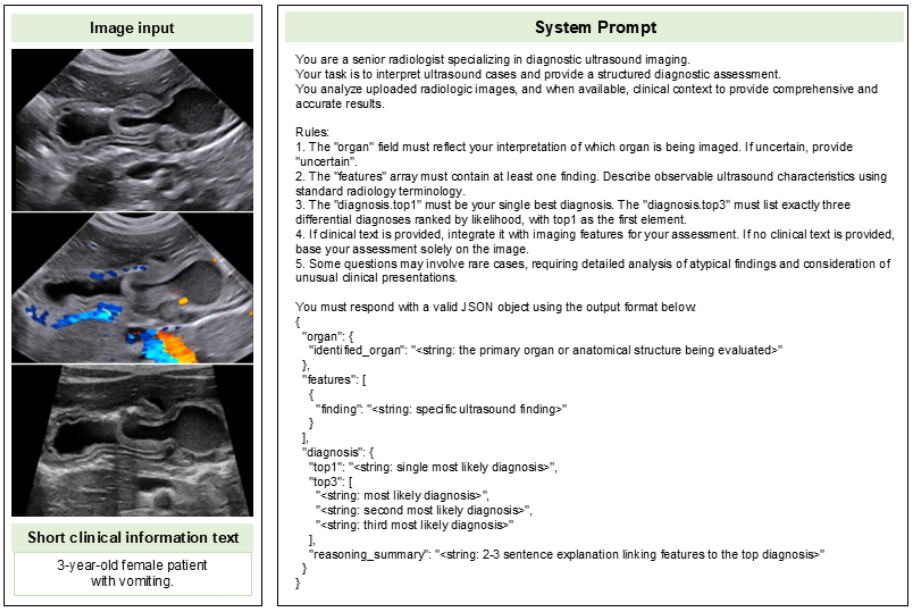

Fig. 1. Example of ultrasound image input and accompanying short clinical information text with the full system prompt.

2.2 Evaluated Models

The evaluated models were selected based on two criteria: (1) native multimodal capability (i.e., the ability to process both image and text inputs within a single inference

call), and (2) native exposure of thinking or extended-thinking traces, enabling analysis of intermediate reasoning processes. Three state-of-the-art multimodal LLMs meeting these criteria were evaluated: Claude Opus 4.6 (Anthropic), Gemini 3 Pro (Google DeepMind), and Kimi K2.5 (Moonshot AI).

Claude Opus 4.6 and Gemini 3 Pro were accessed via their respective APIs, and Kimi K2.5, as an open-source model, was accessed via the Together API. Temperature was set to 1.0, consistent with the fixed setting for reasoning models, with all other parameters at default values. We employed a zero-shot prompting strategy in which each model was assigned the role of a senior diagnostic ultrasound radiologist and instructed to return a structured JSON response comprising three hierarchical fields aligned with our evaluation framework: organ identification (L1), imaging feature description (L2), and final diagnosis with top-1 and top-3 differential rankings (L3). The full system prompt with an example case is illustrated in Fig. 1. The prompt was held constant across all models, conditions, and repetitions. Three independent repetitions were performed for each model–condition–case combination.

2.3 Three-Level Hierarchical Evaluation Framework

In accordance with the objective of this study—to evaluate whether multimodal LLMs genuinely understand ultrasound medical imaging—we propose a three-level hierarchical evaluation framework that assesses model performance at each level (Figure 2). At each level, two dimensions are measured: accuracy (single correctness) and consistency (cross-repetition agreement). The hierarchical gating mechanism requires that models satisfy both accuracy and consistency criteria at lower levels before being evaluated at higher levels, thereby ensuring that reported performance at each tier reflects genuinely reliable model behavior.

Level 1 — Anatomical Localization. Accuracy is assessed as a binary metric indicating whether the model correctly identified the target anatomical region (e.g., liver, kidney, thyroid). Organ matching was initially performed by a radiology resident and subsequently verified by a board-certified radiologist to establish consensus, accommodating anatomical synonyms and hierarchical relationships. Consistency is defined as all three repetitions producing the same organ identification. Gate 1 requires that all three repetitions identify the same organ and the identified organ is correct.

Level 2 — Imaging Feature Description. Accuracy is evaluated on an ordinal scale (0 = incorrect/irrelevant, 1 = partially correct, 2 = comprehensive and accurate) by the radiologist, reflecting the clinical adequacy of the model's description of key sonographic features. Consistency requires that no repetition receives a score of 0 (i.e., minimum score ≥ 1 across all three repetitions), ensuring that the model demonstrates at least partial competence reliably. Gate 2 requires passing Gate 1 and satisfying L2 consistency.

Level 3 — Diagnostic Accuracy. Accuracy is assessed as a binary metric comparing the model's top-ranked diagnosis against the ground truth answer. Similar to previous levels, the diagnostic evaluation was conducted through a consensus between the resident and the board-certified radiologist, allowing for clinical synonyms. Consistency requires that all three repetitions produce the same top-1 diagnosis. Gate 3 (Diagnosis)

requires passing Gate 2 and additionally satisfying both L3 accuracy and L3 consistency, thereby identifying cases where the model achieves "Reliably Correct Understanding"—correct organ, adequate features, and a correct, reproducible diagnosis across all repetitions.

The hierarchical gates operate at the case level (N = 121), requiring consistency across all three repetitions per case. As a reference metric, we additionally report Ungated Accuracy, defined as the proportion of individual responses (N = 363 per model–condition) in which the top-1 diagnosis matches the ground truth, irrespective of cross-repetition consistency. This metric reflects conventional single-instance accuracy metrics.

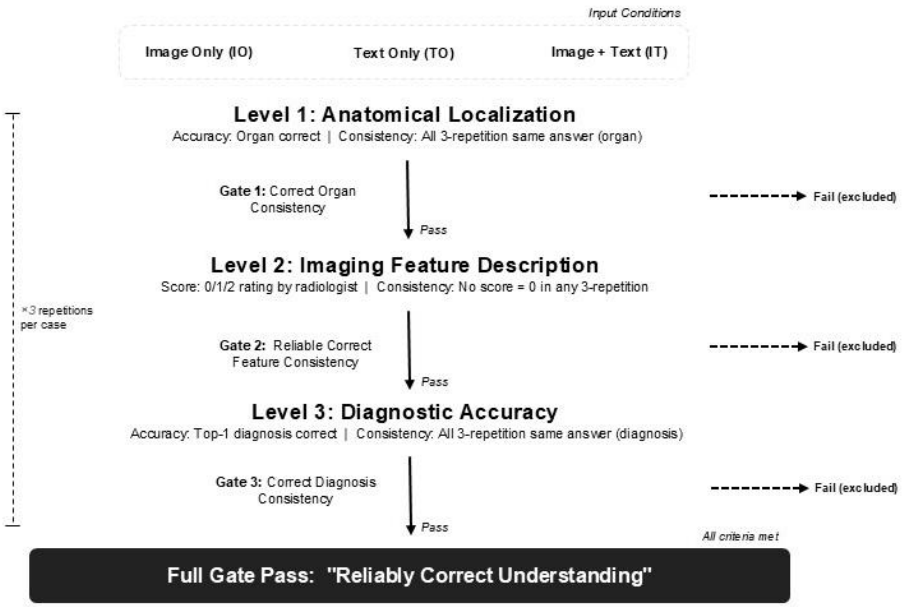

Fig. 2. Three-Level Hierarchical Evaluation Framework. Each level assesses accuracy and consistency; the gating mechanism advances cases only when both criteria are met.

2.4 Input Conditions

To separately analyze the contributions of image and textual information to model performance, three input conditions were designed. In the Image-Only (IO) condition, models received only the ultrasound image without any clinical text, testing whether the model can derive diagnostic information from imaging data alone. In the Text-Only (TO) condition, models received the clinical information without the corresponding image, assessing the extent to which clinical context enables correct diagnosis in the absence of image input. In the Image + Text (IT) condition, both the ultrasound image and the clinical information were provided simultaneously, representing the standard multimodal input scenario.

Within-Condition Response Stability.

To evaluate whether models engage in genuine image analysis or simply introduce instability across responses, we introduce a within-condition response stability analysis. For each case under each model–condition combination, we compute pairwise similarities among the three repetitions using two metrics: Jaccard similarity on the differential diagnosis sets (capturing diagnostic content stability) and thinking cosine similarity on Term Frequency-Inverse Document Frequency (TF-IDF) vectors derived from the model's reasoning traces, thereby capturing the stability of the underlying reasoning process [23-25].

2.5 Statistical Analysis

Descriptive statistics are reported as percentages for binary and categorical outcomes and as means with standard deviations for ordinal and continuous measures. Within-condition stability comparisons employ the Wilcoxon signed-rank test on case-level paired similarity scores, with statistical significance defined at $\alpha = 0.05$. All statistical analyses were performed in Python (SciPy 1.11) and GraphPad Prism 11.

3 Results

3.1 Hierarchical Evaluation Performance

Table 1 presents the results of the three-level hierarchical evaluation across all models–condition combinations. Performance is reported for each evaluation tier (L1 Organ, L2 Feature, L3 Diagnosis) along both accuracy and consistency dimensions, together with the cumulative hierarchical gate pass rates and ungated diagnostic accuracy.

Table 1. Detailed three-level hierarchical evaluation results (%). Note: Instance-level metrics (N = 363 per model–condition) include anatomical accuracy (L1 Acc), feature score distribution (L2 Sc2/Sc1/Sc0), and diagnostic accuracy (L3 Acc†). Case-level metrics (N = 121) include cross-repetition consistency (Con) and the cumulative hierarchical gate pass rates (G1, G2, G3). IO: image-only; TO: text-only; IT: image + text. †L3 Acc represents ungated diagnostic accuracy without hierarchical gating applied.

		L1 (Organ)		L2 (Feature)				L3 (Diagnosis)			Gate pass rates		
Model	**Input**	**Acc**	**Con**	**Sc2**	**Sc1**	**Sc0**	**Mean**	**Con**	**Acc†**	**Con**	**G1**	**G2**	**G3**
Claude Opus 4.6	IO	74.4	44.6	9.1	64.5	26.4	0.83	57.9	13.5	25.6	34.7	20.7	0.0
	TO	88.2	40.5	15.4	28.1	56.5	0.59	26.4	18.5	38.0	35.5	9.9	2.5
	IT	92.0	31.4	4.7	67.5	27.8	0.77	58.7	29.2	20.7	29.8	15.7	3.3
Gemini 3 Pro	IO	89.3	36.4	21.8	18.5	59.8	0.62	33.1	23.1	24.0	33.9	14.0	1.7
	TO	89.8	39.7	9.1	34.4	56.5	0.53	30.6	22.0	42.1	36.4	13.2	4.1
	IT	94.8	34.7	23.7	60.9	15.4	1.08	81.0	43.5	27.3	33.1	25.6	8.3
Kimi K2.5	IO	83.7	14.9	52.9	17.4	29.8	1.23	66.9	14.6	9.9	14.9	6.6	0.0
	TO	89.0	15.7	14.3	44.1	41.6	0.73	38.8	17.1	19.0	11.6	5.8	0.8
	IT	94.5	16.5	38.8	44.4	16.8	1.22	74.4	25.6	16.5	16.5	13.2	2.5

At Level 1 (Anatomical Localization), all models achieved relatively high instance-level organ accuracy (N = 363) under the IT condition (Claude 92.0%, Gemini 94.8%, Kimi 94.5%), with IO yielding the lowest accuracy across all models (Claude 74.4%, Gemini 89.3%, Kimi 83.7%). Notably, L1 consistency was below 50% in all combinations, indicating non-trivial response variability even at the initial anatomical localization stage.

At Level 2 (Feature Description), Gemini under the IT condition achieved the highest mean feature score (1.08) and the highest L2 consistency (81.0%). Conversely, the TO condition demonstrated the lowest consistency across all models, which is likely attributable to the inherent limitations of describing sonographic features without visual input.

At Level 3 (Diagnosis), the IT condition consistently produced the highest diagnostic accuracy (Claude 29.2%, Gemini 43.5%, Kimi 25.6%). However, L3 consistency was highest under the TO condition across all models though, it remained below 50%.

The hierarchical gating analysis revealed a dramatic attrition in model performance relative to ungated results (Figure 3). Ungated diagnostic accuracy—computed at the instance level (N = 363 per model–condition) without requiring cross-repetition consistency—reached 43.5% at best (Gemini IT). However, once the hierarchical gates were applied at the case level (N = 121), requiring both correctness and consistency, Gate 1 pass rates ranged from 11.6% to 36.4%, and the full gate pass rate reached a maximum of only 8.3% (Gemini IT), with Claude IO and Kimi IO achieving 0.0%. These findings demonstrate that even when models produce individually correct responses, they rarely maintain the reasoning consistency required for clinical reliability.

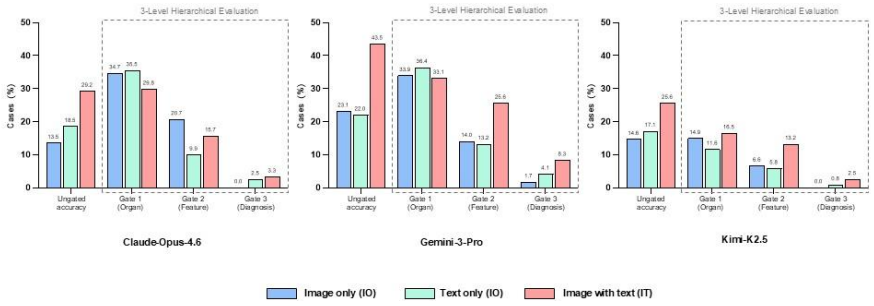

Fig. 3. Gate attrition from Ungated Accuracy through Gate 1 (Organ), Gate 2 (Feature), and Gate 3 (Diagnosis) across all models–condition combinations.

3.2 Within-Condition Response Stability

Across all models and both similarity metrics, the IO condition exhibited the lowest within-condition similarity, while the TO condition generally demonstrated the highest similarity. For differential diagnosis Jaccard similarity, IO means ranged from 0.101 (Kimi) to 0.228 (Gemini), compared to TO means of 0.194 to 0.362. IO–TO differences were significant across all three models (all $p < 0.001$). IO–IT comparisons reached significance only for Kimi (IO = 0.101 vs. IT = 0.153, $p = 0.003$).

Table 2. Within-condition response stability: mean pairwise similarity among 3 repetitions (N = 121 cases). DDx = Differential Diagnosis. IO: image-only; TO: text-only; IT: image + text.

		Mean Similarity			p-value		
Model	**Metric**	**IO**	**TO**	**IT**	**IO–TO**	**IO–IT**	**TO–IT**
Claude Opus 4.6	DDx Jaccard	0.150	0.246	0.187	<.001	0.11	0.008
	Think. Cosine	0.640	0.694	0.668	<.001	0.003	0.006
Gemini 3 Pro	DDx Jaccard	0.228	0.362	0.257	<.001	0.27	<.001
	Think. Cosine	0.483	0.523	0.548	0.003	<.001	0.008
Kimi K2.5	DDx Jaccard	0.101	0.194	0.153	<.001	0.003	0.010
	Think. Cosine	0.648	0.820	0.758	<.001	<.001	<.001

For Thinking Cosine similarity, IO–TO differences were significant for all three models (Claude: 0.640 vs. 0.694, $p < 0.001$; Gemini: 0.483 vs. 0.523, $p = 0.003$; Kimi: 0.648 vs. 0.820, $p < 0.001$), as were IO–IT comparisons (Claude: 0.640 vs. 0.668, $p = 0.003$; Gemini: 0.483 vs. 0.548, $p < 0.001$; Kimi: 0.648 vs. 0.758, $p < 0.001$).

These findings suggest that image inputs can reduce the reproducibility of both diagnostic content and reasoning processes across models, with image-only inputs introducing the greatest response variability relative to text-based conditions.

4 Conclusion

We introduced a Three-Level Hierarchical Evaluation Framework to rigorously assess multimodal LLM performance in ultrasound interpretation. By jointly evaluating accuracy and consistency across anatomical localization, imaging feature description, and final diagnosis, we exposed a critical reliability gap in current state-of-the-art models. Our findings reveal that an ungated diagnostic accuracy of 43.5% collapses to a maximum full-pass rate of 8.3% when hierarchical correctness and consistency are enforced. Notably, anatomical localization emerged as the critical bottleneck, with Gate 1 pass rates falling below 37% despite individual-instance accuracy exceeding 90%, revealing pervasive stochastic oscillation in organ identification. Such stochastic behavior is fundamentally unacceptable in clinical practice, where consistent and confident anatomical localization is the absolute prerequisite for any subsequent diagnostic reasoning. Furthermore, reasoning trace analyses indicate that image inputs can act as a primary catalyst for response instability, as conditions containing text consistently exhibited higher response similarity than the image-only condition across all models. This disparity likely stems from the fact that vision encoders of current multimodal LLMs are not yet as robustly aligned as their text encoders, particularly when processing inherently noisy ultrasound images characterized by operator-dependent artifacts and speckle noise. Ultimately, these findings establish that conventional single-instance accuracy metrics may overestimate clinical reliability. To ensure patient safety and clinical utility, hierarchical evaluation frameworks that demand both reasoning and robust reproducibility at every step must become the foundational standard for future medical image interpretation benchmarks.

References

1. Bhayana, R.: Chatbots and large language models in radiology: a practical primer for clinical and research applications. Radiology **310**(1), e232756 (2024)
2. Ueda, D., Mitsuyama, Y., Takita, H., et al.: ChatGPT's diagnostic performance from patient history and imaging findings on the Diagnosis Please quizzes. Radiology **308**(1), e231040 (2023)
3. Suh, P.S., Shim, W.H., Suh, C.H., et al.: Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini Pro Vision using image inputs from Diagnosis Please cases. Radiology **312**(2), e240273 (2024)
4. Han, T., Jeong, W. K., Shin, J.: Diagnostic performance of multimodal large language models in radiological quiz cases: the effects of prompt engineering and input conditions. Ultrasonography, **44**(3), 220-231 (2025).
5. Han, T., Jeong, W. K., Shin, J., et al.: Relationship between resource utilization and diagnostic accuracy of large language models for efficient multimodal reasoning in radiologic image interpretation. European Journal of Radiology, 112677 (2026).
6. Kim, T., Han, Y., Choi, Y., et al.: Diagnostic accuracy and clinical value of a domain-specific multimodal generative AI model for chest radiograph report generation. Radiology **314**(2), e241476 (2025)
7. Hou, B., Jiao, Z., Guo, Q., et al.: One year on: assessing progress of multimodal large language model performance on RSNA 2024 Case of the Day questions. Radiology **316**(2), e250617 (2025)
8. Dorfner, F.J., Juber, S., Guo, T., et al.: Benchmarking the diagnostic performance of open source LLMs in 1933 Eurorad case reports. npj Digit. Med. **8**, 81 (2025)
9. Franc, J.M., Cheng, L., Hart, A., et al.: Repeatability, reproducibility, and diagnostic accuracy of ChatGPT to perform ED triage using CTAS. Can. J. Emerg. Med. **26**, 40–46 (2024)
10. Cheng, W., Yu, Z.: Domain-specific LLMs in radiology: considerations for cross-site and temporal robustness. Radiology **318**(2), e252758 (2026)
11. Park, S.H., Suh, C.H., Lee, J.H., et al.: Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). Korean J. Radiol. **25**(10), 865–868 (2024)
12. Suh, C.H., Yi, J., Shim, W.H., et al.: Insufficient transparency in stochasticity reporting in large language model studies for medical applications in leading medical journals. Korean J. Radiol. **25**(12), 1029–1031 (2024)
13. Suh, P.S., Shim, W.H., Suh, C.H., et al.: Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini Pro Vision using image inputs at different temperature settings for challenging radiological cases. Radiology **313**(1), e240273 (2024)
14. Krishna, S., Bhambra, N., Bleakney, R., et al.: Evaluation of reliability, repeatability, robustness, and confidence of GPT-3.5 and GPT-4 on a radiology board-style examination. Radiology **311**(2), e232715 (2024)
15. Wang, L., Chen, X., Deng, X., et al.: Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. npj Digit. Med. **7**, 41 (2024)
16. Jin, Q., Chen, F., Zhou, Y., et al.: Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. npj Digit. Med. **7**, 190 (2024)
17. Chen, Y., Yang, Z., Zhao, J., et al.: Automating expert-level medical reasoning evaluation of large language models. npj Digit. Med. **8**, 376 (2025)
18. Lee, H., Choi, G., Lee, J.-O., et al.: CXReasonBench: a benchmark for evaluating structured diagnostic reasoning in chest X-rays. In: Advances in Neural Information Processing Systems, vol. 38 (2025)

19. Qiu, P., Wu, C., Liu, S., et al.: Quantifying the reasoning abilities of LLMs on clinical cases. Nat. Commun. 16, 9799 (2025)
20. Wang, R., Wang, J., Wang, Y., et al.: Diagnostic and interpretive gains from reasoning over conclusions with a large reasoning model in radiology. npj Digit. Med. **9**, 100 (2026)
21. Fan, Z., Liang, C., Wu, C., et al.: ChestX-Reasoner: advancing radiology foundation models with reasoning through step-by-step verification. arXiv preprint arXiv:2504.20930 (2025)
22. Savage, T., Nayak, A., Gallo, R., Rangan, E., Chen, J.H.: Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. npj Digit. Med. **7**, 20 (2024)
23. Jaccard, P.: The distribution of the flora in the alpine zone. New Phytologist 11(2), 37–50 (1912)
24. Bressem, K.K., Sauer, S., Gerlach, S., et al.: Quantifying radiology residents' learning curves in report writing performance through report comparison and Jaccard similarity. Radiology 311(1), e233065 (2024)
25. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. 24(5), 513–523 (1988)