Original Paper

Confidence Measurement Metrics in Multimodal Large Language Models for Ultrasound-Based Radiology Cases: Comparative Evaluation Study of Self-Reported, Consistency-Based, and Hybrid Methods

Taewon Han, MD; Jaeseung Shin, MD, PhD; Jeong Hyun Lee, MD; Kyowon Gu, MD

Department of Radiology, Samsung Medical Center, Seoul, Republic of Korea

Corresponding Author:

Jaeseung Shin, MD, PhD
Department of Radiology
Samsung Medical Center
81 Irwon-ro, Irwon-dong - Gangnam-gu
Seoul 06351
Republic of Korea
Phone: 82 10-8714-7650
Email: dr.shinjs@gmail.com

Abstract

Background: Large language models (LLMs) require specialized methodologies to quantify model confidence for safe deployment in health care systems; however, there is a lack of established methods for confidence assessment.

Objective: This study aimed to evaluate confidence metrics for multimodal LLMs interpreting ultrasound-based radiology cases and to compare self-reported, consistency-based, and hybrid methods.

Methods: From a total of 330 quizzes on the Korean Society of Ultrasound in Medicine digital platform, we selected 94 multiple-choice cases. Four multimodal LLMs were evaluated: 3 reasoning models (GPT-5, Claude-4.5-Sonnet, and Gemini-3-Pro) and 1 general model (GPT-4o). Temperature was fixed at 1.0. Multiple confidence metrics were assessed: (1) self-reported metrics generated by LLMs using prompts that elicited direct confidence percentages with answers, including first self-reported confidence and mean self-reported confidence; (2) consistency-based metrics derived from 20 repeated outputs per case, including relative entropy calculated as $1 - H/\log_2 k$ (H=Shannon entropy, k=number of answer choices) and majority-vote percentage; and (3) a Top Weighted Score combining response frequency with self-reported confidence. Receiver operating characteristic analysis for discrimination and Spearman correlation between accuracy and each confidence metric was conducted. Additionally, model calibration was assessed using expected calibration error and Brier score. Processing time and token consumption (input, output, and total) were recorded for each application programming interface call to evaluate resource use across models.

Results: Diagnostic accuracy varied across models, with Gemini-3-Pro achieving the highest accuracy (70/94, 74.47%), surpassing the median human accuracy (59%, IQR 40.3%-75%). Top Weighted Score, a hybrid metric combining response frequency and self-reported confidence, was the only metric achieving statistically significant correlations across all 4 models: Gemini-3-Pro (ϱ=0.52), GPT-5 (ϱ=0.43), Claude-4.5-Sonnet (ϱ=0.30), and GPT-4o (ϱ=0.22). Receiver operating characteristic analysis revealed that Top Weighted Score demonstrated the highest discriminative ability, with area under the curve values of 0.826 (95% CI 0.731-0.920) for Gemini-3-Pro and 0.767 (95% CI 0.668-0.866) for GPT-5. Top Weighted Score was the only metric achieving statistical significance in GPT-4o. Calibration analysis showed that Top Weighted Score achieved the lowest expected calibration error in GPT-5 (0.098) and Claude-4.5-Sonnet (0.192), while Gemini-3-Pro showed comparable calibration between relative entropy (0.119) and Top Weighted Score (0.122). Resource use analysis demonstrated that reasoning models required substantially longer processing times and higher token consumption compared to general models.

Conclusions: In multimodal LLMs applied to ultrasound-based radiology cases, hybrid methods (Top Weighted Score) demonstrated significant associations across all evaluated models and appear to serve as more reliable indicators of diagnostic confidence compared to self-reported or consistency-based metrics alone, although the strength of these associations varied across models, and external validation is warranted before broader clinical application. These findings support integrative

confidence estimation approaches that incorporate response consistency while highlighting the need for resource-efficient sampling strategies to enable practical clinical deployment.

J Med Internet Res 2026;28:e86498; doi: 10.2196/86498

Keywords: artificial intelligence; AI; radiology; medical informatics; diagnostic confidence; large language models; LLMs

Introduction

The integration of large language models (LLMs) into clinical workflows is accelerating from promise to practice [1-3]. These advancements necessitate robust frameworks for evaluating output reliability to ensure patient safety when LLMs are used for medical decision-making [4]. Of particular concern is the calibration of LLMs—the alignment between a model's confidence and its true accuracy—as poorly calibrated LLMs may deliver inaccurate responses with inappropriately high confidence [5], potentially introducing significant risks to patients through downstream diagnostic and treatment errors [6].

Unlike traditional probabilistic classifiers (eg, logistic regression or convolutional neural networks) that expose an explicit class probability for each prediction, LLMs generate text sequentially using probabilities but may present answers confidently despite substantial uncertainty in their underlying probability distributions, resulting in overconfidence issues [4]. This tendency toward overconfidence complicates the safe deployment of LLMs in health care systems and underscores the need for specialized methodologies to quantify model confidence for clinical end users [4].

While existing deep learning models have demonstrated established methods for uncertainty quantification in medical artificial intelligence through various approaches [7-9], architecturally different LLMs still lack standardized methodologies. Several approaches have been proposed to estimate the confidence of LLM outputs [10,11]. In the self-reported method, the model is explicitly prompted to assign a numerical confidence score, typically 0% to 100%, to its own answer [12,13]. Another approach uses sample consistency, leveraging the stochastic behavior of LLMs by running the same prompt multiple times and estimating confidence from the agreement or entropy of the resulting responses [14]. Additional methods include directly using token-level log probabilities to quantify confidence [15]. Despite these various methodologies, a best-practice standard for assessing LLM confidence has not yet been established.

Given the lack of established best practices for confidence assessment, a systematic appraisal of available techniques is essential before these systems can be deployed in clinical practice. Therefore, this study aims to evaluate confidence measurement metrics of multimodal LLMs tasked with ultrasound-based radiological cases assessing whether these approaches can serve as reliable indicators of diagnostic confidence.

Methods

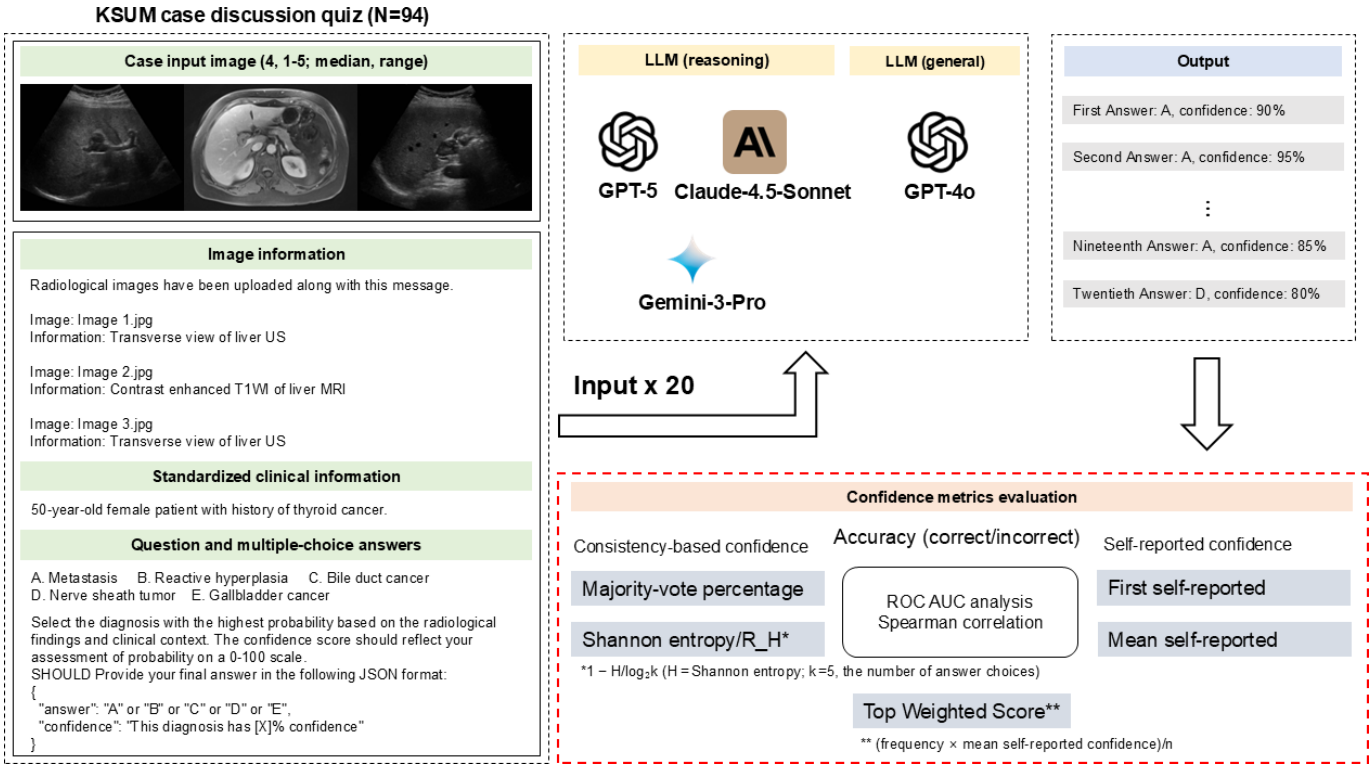
Ethical Considerations

This investigation used publicly available educational datasets, obviating the need for institutional review board approval or informed consent. All quiz materials from the Korean Society of Ultrasound in Medicine (KSUM) website were previously deidentified before public release. No compensation was involved in this study, and no identifiable individuals appear in any images within the manuscript or supplementary materials.

Dataset

A total of 330 case discussion quizzes were extracted by a radiologist (TH, with 4 years of experience) from the KSUM digital platform [16], published between July 28, 2000, and December 25, 2025. The radiologist systematically collected imaging data, question content with corresponding multiple-choice options, and relevant imaging information, including imaging modality and anatomical site. We excluded 236 cases without multiple-choice formats to maintain measurement reliability, resulting in a final dataset of 94 quiz cases (Figure 1). These quiz cases encompassed various imaging modalities, with some cases featuring challenging diagnostic scenarios. To focus on multimodal capabilities, we standardized the brief clinical text to include only patient demographics (age and sex) and chief complaint or previous medical history. The ground truth for our study was established using officially designated answers from the KSUM platform. Human performance benchmarks were derived from response statistics of KSUM platform subscribers, mainly radiologists and radiology trainees with varying degrees of expertise.

Figure 1. Flowchart depicting the evaluation process for consistency metrics across 3 reasoning large language models and 1 general large language model. AUC: area under the curve; KSUM: Korean Society of Ultrasound in Medicine; LLM: large language model; R_H: relative entropy; ROC: receiver operating characteristic.



Multimodal LLMs

We selected four multimodal LLMs, including three reasoning models—(1) GPT-5 (Alias: 2025-08-07; OpenAI) [17]; (2) Claude-4.5-Sonnet (Alias: 2025-09-29; Anthropic) [18]; (3) Gemini-3-Pro (Alias: 2025-11-18; Google) [19]—and one general model—GPT-4o (Alias: 2024-11-20; OpenAI) [20]. The temperature was fixed at 1.0 across all models because the reasoning models do not permit adjustment, and previous literature reports optimal performance at this setting [21]. Additionally, enhanced reasoning capabilities were activated for applicable models: "reasoning effort" and "thinking level" were set to "high" for GPT-5 and Gemini-3-Pro-Preview, and "thinking" mode was enabled for Claude-4.5-Sonnet.

Confidence Measurement Metrics

Radiological images were paired with brief text prompts that described the clinical question and key imaging information. We evaluated self-reported, consistency-based, and hybrid confidence metrics while simultaneously measuring diagnostic accuracy (Figure 1). For each case, each model generated 20 independent outputs, enabling the analysis of consistency-based metrics.

First, a relative entropy–based score (R_H) was calculated as $R\_H=1 - H/\log_2 k$, where $H=-\sum (i=1 \text{ to } k)\ p_i \log_2 p_i$; p_i represents Shannon entropy, p_i represents the relative frequency of option i across repeated model outputs, and k=5 is the number of answer choices (so $\log_2 k \approx 2.322$ bits). R_H ranges from 0 (maximum entropy, complete inconsistency) to 1 (zero entropy, perfect consistency); for example, a response pattern of [A A A A A] yields R_H=1, whereas [A B C D E] yields R_H=0. Intermediate patterns receive scores reflecting their coherence; [A A A B B] would yield a higher R_H value than [A A A B C] because of greater consistency. The raw Shannon entropy (H) was also reported alongside R_H to provide an unnormalized measure.

Second, a majority-vote percentage recorded the proportion of the most frequent response in repeated trials. Both [A A A B B] and [A A A B C], for instance, produce a modal proportion of 60%, illustrating that this metric ignores differences in the dispersion of the remaining responses.

Third, a weighted confidence score was calculated for each answer option as follows: (frequency × mean self-reported confidence)/n, where frequency is the count of that option across repeated trials, mean self-reported confidence is the average confidence rating for that option, and n is the total number of repetitions. The Top Weighted Score was defined as the highest score among all options. For example, if option A appeared 12 times with a mean confidence of 80% and option B appeared 8 times with a mean confidence of 90%, the weighted scores would be (12×80)/20=48 for A and (8×90)/20=36 for B, yielding a Top Weighted Score of 48.

In parallel, each model was prompted to append a numerical self-confidence rating (0%-100%) to each answer. Two self-reported confidence metrics were derived: (1) first self-reported confidence, which used the confidence rating from the first response and (2) mean self-reported confidence, which averaged the confidence ratings across all responses that selected the majority-vote answer. The prompt instructed models to select the diagnosis based on radiological findings

and clinical context before providing the final answer in JSON format, including a confidence score on a 0%-100% scale (eg, "Select the diagnosis with the highest probability… Provide your final answer in the following JSON format: answer: A-E, confidence: 0%-100%"). The exact templates are presented in Table S1 in Multimedia Appendix 1 and Figure 1.

To determine the minimum repetition count required for consistency-based metrics, analyses were additionally conducted with 5, 10, and 15 repeated outputs per case.

For ROC AUC, Spearman correlation, and calibration analyses, each confidence metric was paired with the diagnostic accuracy of its corresponding representative answer. For consistency-based metrics (R_H and majority-vote percentage) and mean self-reported confidence, the representative answer was the majority-voted option across repeated outputs. For the Top Weighted Score, the representative answer was the option receiving the highest weighted score. For the first self-reported confidence, the representative answer was the model's first response. Diagnostic accuracy was assessed by comparing the corresponding representative answer for each metric with the KSUM ground truth.

Resource Use

To evaluate the trade-off between confidence estimation reliability and resource efficiency, we recorded the processing time and token consumption for each model query. Processing time was measured as the duration from application programming interface (API) request submission to response completion. Token usage was recorded as input tokens (text prompt and image data), output tokens (model-generated response), and total tokens consumed per query. We calculated the cumulative processing time and token consumption required to analyze a single quiz case for each repetition count condition (1, 5, 10, 15, and 20).

Statistical Analysis

Accuracy for each model and repetition count (1, 5, 10, 15, and 20) was quantified as the proportion of correct answers. Differences across repetition counts were tested using the Cochran *Q* test [22]. Discriminative ability was evaluated using receiver operating characteristic (ROC) area under the curve (AUC) with 95% CIs estimated by the DeLong method [23], using each confidence metric as a predictor of diagnostic accuracy (correct vs incorrect). Spearman correlation coefficients (ϱ) measured the association between each confidence measurement and diagnostic accuracy (correct vs incorrect). Correlations were interpreted as negligible ($|\varrho|\leq 0.10$), weak ($0.10<|\varrho|\leq 0.39$), moderate ($0.39<|\varrho|\leq 0.69$), strong ($0.69<|\varrho|\leq 0.89$), or very strong ($|\varrho|>0.89$) [24]. Calibration, defined as the alignment between predicted confidence and actual accuracy, was evaluated using expected calibration error (ECE) and Brier scores [25,26], using fixed 10-bin calibration, with 95% CIs estimated via bootstrap resampling (1000 iterations). The Brier score was calculated based on binary diagnostic accuracy (correct vs incorrect) rather than the multiclass probability across all 5 options.

Model response repeatability was evaluated using the Fleiss κ statistic, with results interpreted as follows: >0.8, almost perfect; 0.61 to 0.80, substantial; 0.41 to 0.60, moderate; 0.21 to 0.40, fair; and <0.20, poor [27]. To assess potential training data contamination, majority-vote accuracy at a repetition count of 20 was compared between cases published before February 2025 (n=72) and from February to December 2025 (n=22) using Fisher exact tests. Statistical significance was established at $P<.05$. Statistical analyses were performed using GraphPad Prism (version 10.4.1; GraphPad Software) and Python (version 3.10).

Results

Dataset and Diagnostic Performance

A study of 94 cases was conducted, with a median of 4 (IQR 3-4; range 1-5) input images per case. The distribution of input image modalities comprised radiography (n=8), ultrasonography (n=94), computed tomography (n=18), magnetic resonance imaging (n=26), nuclear medicine imaging (n=6), and other diagnostic techniques, including endoscopic visualization (n=2) and aspiration fluid (n=2). Human accuracy demonstrated substantial variability (median 59%, IQR 40.3%-75%; range 5%-96%; mean 56.8%, SD 22%).

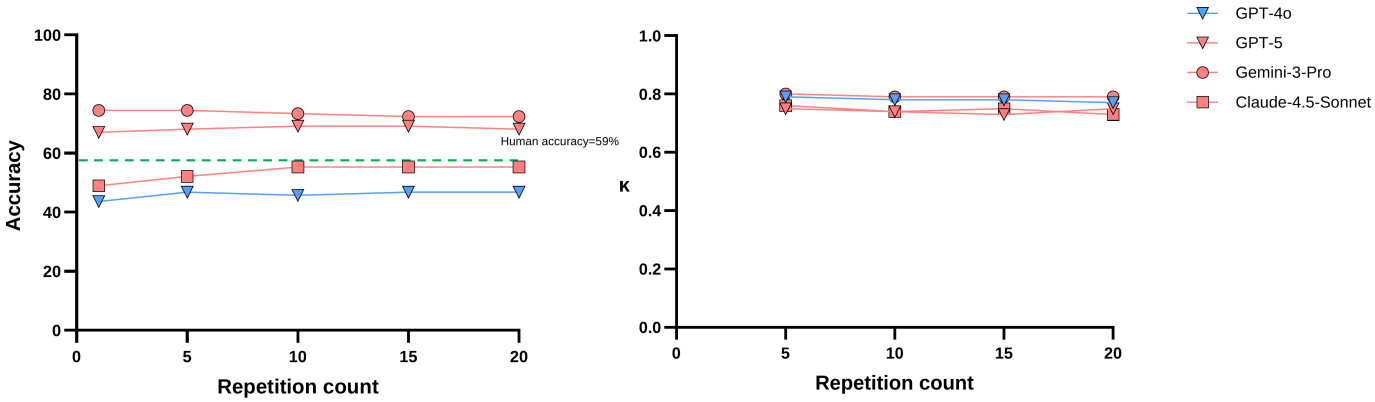
Table 1 and Figure 2 demonstrate the implementation of majority voting across repetition counts. Claude-4.5-Sonnet showed a significant improvement in diagnostic accuracy with majority voting, improving from 48.94% (46/94) to 55.32% (52/94) at 10 repetitions ($P=.01$). In contrast, Gemini-3-Pro, GPT-5, and GPT-4o showed no significant change with majority voting ($P=.67$, $P=.94$, and $P=.08$, respectively), with Gemini-3-Pro showing the highest accuracy ranging from 72.34% (68/94) to 74.47% (70/94).

Table 1. Comparison of first output and majority vote accuracy across multimodal large language models (N=94).[a]

Model	First output, n (%)	Majority vote (5), n (%)	Majority vote (10), n (%)	Majority vote (15), n (%)	Majority vote (20), n (%)	*P* value
Claude-4.5-Sonnet	46 (48.94)	49 (52.13)	52 (55.32)	52 (55.32)	52 (55.32)	.01
Gemini-3-Pro	70 (74.47)	70 (74.47)	69 (73.40)	68 (72.34)	68 (72.34)	.67
GPT-5	63 (67.02)	64 (68)	65 (69.15)	65 (69.15)	64 (68.09)	.94
GPT-4o	41 (43.62)	44 (46.81)	43 (45.74)	44 (46.81)	44 (46.81)	.08

[a]Differences in accuracy were assessed using the Cochran *Q* test. Human accuracy for these cases showed substantial variability (median 59%, IQR 40.3%-75%; mean 56.8%, SD 22%, range 5%-96%).

Figure 2. Accuracy and ϰ values plotted against repetition number.



Fleiss ϰ analysis demonstrated substantial within-model repeatability across all models (Table S2 in Multimedia Appendix 1). Gemini-3-Pro achieved the highest consistency (ϰ=0.79-0.80), followed by GPT-4o (ϰ=0.77-0.79). Claude-4.5-Sonnet and GPT-5 showed comparable repeatability (ϰ=0.73-0.76 and ϰ=0.73-0.75, respectively). All models maintained substantial agreement across all repetition counts, with response consistency remaining stable from 5 to 20 repetitions (Figure 2).

No statistically significant differences in majority-vote accuracy were observed between cases published before and after February 2025 for any of the 4 evaluated models (all *P*>.79; Table S3 in Multimedia Appendix 1).

Confidence Measurement Metrics

The discriminative ability of confidence metrics varied substantially across models (Table 2). Top Weighted Score demonstrated the highest discriminative performance, achieving significant ROC AUC values across all models. Gemini-3-Pro showed the strongest discrimination with Top Weighted Score (ROC AUC=0.826, 95% CI 0.731-0.920, *P*<.001), followed by GPT-5 (ROC AUC=0.767, 95% CI 0.668-0.866, *P*<.001), Claude-4.5-Sonnet (ROC AUC=0.676, 95% CI 0.568-0.785, *P*=.001), and GPT-4o (ROC AUC=0.629, 95% CI 0.509-0.749, *P*=.04). Notably, Top Weighted Score was the only metric achieving statistical significance in GPT-4o.

Table 2. Receiver operating characteristic area under the curve comparing discriminative ability of confidence metrics in multimodal large language models.

Model	Self-reported (first) (ROC[a] AUC[b], 95% CI, *P* value[c])	Self-reported (mean) (ROC AUC, 95% CI, *P* value)	R_H[d] (ROC AUC, 95% CI, *P* value)[e]	Majority-vote percentage (ROC AUC, 95% CI, *P* value)	Top Weighted Score (ROC AUC, 95% CI, *P* value)
Claude-4.5-Sonnet	0.706, 0.602-0.810, <.001	0.636, 0.523-0.748, .02	0.671, 0.565-0.778, .002	0.668, 0.562-0.775, .002	0.676, 0.568-0.785, .001
Gemini-3-Pro	0.532, 0.439-0.625, .50	0.661, 0.546-0.775, .006	0.779, 0.672-0.887, <.001	0.790, 0.682-0.897, <.001	0.826, 0.731-0.920, <.001
GPT-5	0.719, 0.613-0.826, <.001	0.659, 0.547-0.771, .005	0.755, 0.647-0.863, <.001	0.740, 0.631-0.848, <.001	0.767, 0.668-0.866, <.001
GPT-4o	0.597, 0.491-0.703, .07	0.592, 0.476-0.708, .12	0.576, 0.463-0.689, .19	0.577, 0.464-0.689, .18	0.629, 0.509-0.749, .04

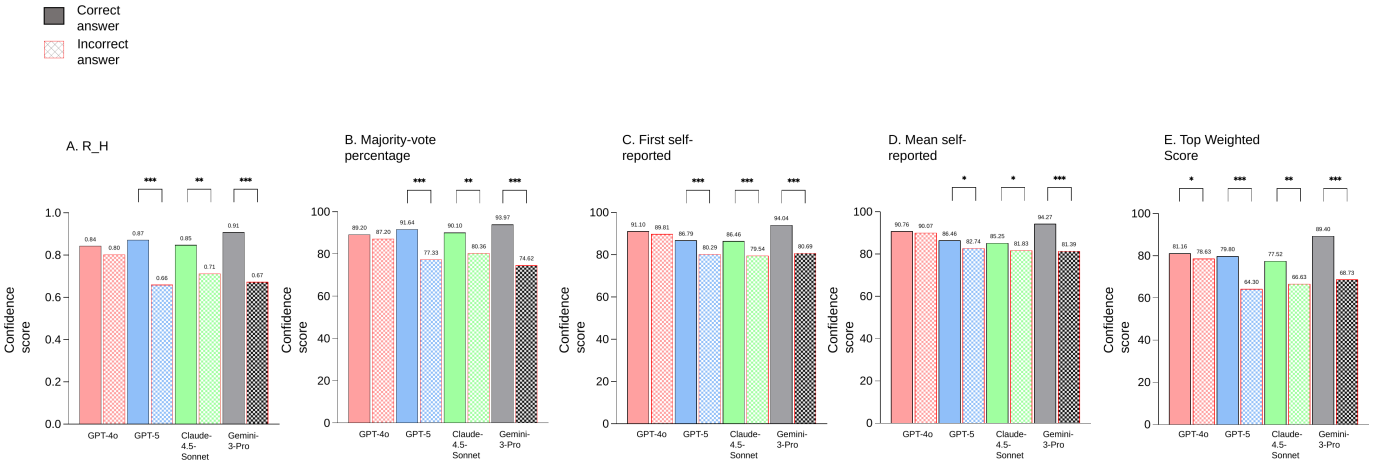
[a]ROC: receiver operating characteristic.
[b]AUC: area under the curve.
[c]Statistically significant results (*P*<.05) are marked with an asterisk.
[d]R_H: relative entropy.
[e]Shannon entropy is not reported separately because it yielded identical AUC values to the relative entropy–based score.

Consistency-based metrics (R_H, majority-vote percentage, and Shannon entropy) showed strong discriminative ability in 3 models. Gemini-3-Pro achieved the highest performance (R_H, ROC AUC=0.779, 95% CI 0.672-0.887, *P*<.001; majority-vote percentage, ROC AUC=0.790, 95% CI 0.682-0.897, *P*<.001; Shannon entropy, ROC AUC=0.779, 95% CI 0.672-0.887, *P*<.001), followed by GPT-5 (R_H, ROC AUC=0.755, 95% CI 0.647-0.863, *P*<.001; majority-vote percentage, ROC AUC=0.740, 95% CI 0.631-0.848, *P*<.001; Shannon entropy, ROC AUC=0.755, 95% CI 0.647-0.863, *P*<.001) and Claude-4.5-Sonnet (R_H, ROC AUC=0.671, 95% CI 0.565-0.778, *P*=.002; majority-vote percentage, ROC AUC=0.668, 95% CI 0.562-0.775, *P*=.002; Shannon entropy, ROC AUC=0.672, 95% CI 0.565-0.778, *P*=.002). However, GPT-4o showed no significant discrimination with consistency-based metrics (R_H, ROC AUC=0.576, *P*=.19; majority-vote percentage, ROC AUC=0.577, *P*=.18; Shannon entropy, ROC AUC=0.576, *P*=.19).

Self-reported confidence demonstrated model-dependent performance. First self-reported confidence achieved significant discrimination in Claude-4.5-Sonnet (ROC AUC=0.706, 95% CI 0.602-0.810, *P*<.001) and GPT-5 (ROC AUC=0.719, 95% CI 0.613-0.826, *P*<.001). However, mean self-reported confidence showed lower discriminative ability in Claude-4.5-Sonnet (ROC AUC=0.636, *P*=.02), GPT-5 (ROC AUC=0.659, *P*=.005), and Gemini-3-Pro (ROC AUC=0.661, *P*=.006).

Figure 3 replicates the ROC findings. Top Weighted Score demonstrated higher confidence scores for correct responses across all 4 models, whereas R_H, majority-vote percentage, first self-reported confidence, and mean self-reported confidence showed this pattern only in GPT-5, Claude-4.5-Sonnet, and Gemini-3-Pro, excluding GPT-4o.

Figure 3. Distribution of confidence scores stratified by diagnostic accuracy (correct vs incorrect) for each multimodal large language model. (A) Relative entropy (R_H), (B) majority-vote percentage, (C) first self-reported confidence, (D) mean self-reported confidence, and (E) Top Weighted Score. Solid bars indicate correct responses; hatched bars indicate incorrect responses. Differences between groups were assessed using the Mann-Whitney *U* test. **P*<.05; ***P*<.01; ****P*<.001.



Spearman correlation analysis revealed similar patterns (Table 3). Notably, Top Weighted Score was the only metric achieving statistical significance across all 4 models, demonstrating moderate correlations in Gemini-3-Pro (ϱ=0.52, 95% CI 0.35-0.65, *P*<.001) and GPT-5 (ϱ=0.43, 95% CI 0.25-0.58, *P*<.001), and weak correlations in Claude-4.5-Sonnet (ϱ=0.30, 95% CI 0.11-0.48, *P*=.003) and GPT-4o (ϱ=0.22, 95% CI 0.02-0.41, *P*=.03). In GPT-4o, Top Weighted Score was the only metric achieving significance. Consistency-based metrics showed moderate correlations in Gemini-3-Pro (R_H, ϱ=0.48; majority-vote percentage, ϱ=0.50; all *P*<.001) and GPT-5 (R_H, ϱ=0.43; majority-vote percentage, ϱ=0.41; all *P*<.001). Self-reported confidence showed weak correlations across all models.

Table 3. Correlation analysis between accuracy and confidence metrics in multimodal large language models at a repetition count of 20.

Model and metrics	ϱ (95% CI)[a]	*P* value[b]
Claude-4.5-Sonnet		
Self-reported (first)	0.36 (0.17 to 0.53)	<.001
Self-reported (mean)	0.23 (0.03 to 0.42)	.02
R_H[c]	0.31 (0.11 to 0.48)	.002
Majority-vote percentage	0.30 (0.11 to 0.48)	.003
Top Weighted Score	0.30 (0.11 to 0.48)	.003
Gemini-3-Pro		
Self-reported (first)	0.06 (–0.14 to 0.26)	.55
Self-reported (mean)	0.25 (0.05 to 0.43)	.014
R_H	0.48 (0.30 to 0.62)	<.001
Majority-vote percentage	0.50 (0.33 to 0.64)	<.001
Top Weighted Score	0.52 (0.35 to 0.65)	<.001
GPT-5		
Self-reported (first)	0.36 (0.17 to 0.52)	<.001
Self-reported (mean)	0.26 (0.06 to 0.44)	.012
R_H	0.43 (0.25 to 0.58)	<.001
Majority-vote percentage	0.41 (0.22 to 0.56)	<.001
Top Weighted Score	0.43 (0.25 to 0.58)	<.001
GPT-4o		
Self-reported (first)	0.18 (–0.02 to 0.37)	.08
Self-reported (mean)	0.16 (–0.05 to 0.35)	.13
R_H	0.14 (–0.07 to 0.33)	.18

Model and metrics	ϱ (95% CI)[a]	*P* value[b]
Majority-vote percentage	0.14 (–0.06 to 0.33)	.18
Top Weighted Score	0.22 (0.02 to 0.41)	.03

[a]Values represent Spearman correlation coefficients (ϱ) with 95% CIs in parentheses.
[b]Significant correlations (*P*<.05) are marked with an asterisk.
[c]R_H: relative entropy.

Calibration

Calibration analysis revealed that the Top Weighted Score demonstrated the best calibration in GPT-5 (ECE=0.098, 95% CI 0.074-0.211; Brier score=0.185, 95% CI 0.140-0.235) and Claude-4.5-Sonnet (ECE=0.192, 95% CI 0.133-0.307; Brier score=0.259, 95% CI 0.203-0.317). In Gemini-3-Pro, R_H and the Top Weighted Score showed comparable calibration (ECE=0.119 vs 0.122; Brier score=0.164 vs 0.163, respectively). Across Claude-4.5-Sonnet, Gemini-3-Pro, and GPT-5, consistency-based metrics, particularly R_H, demonstrated better calibration compared to self-reported metrics. GPT-4o showed the poorest calibration across all metrics (Table 4).

Table 4. Calibration metrics for different confidence measurement methods in multimodal large language models at a repetition count of 20.[a]

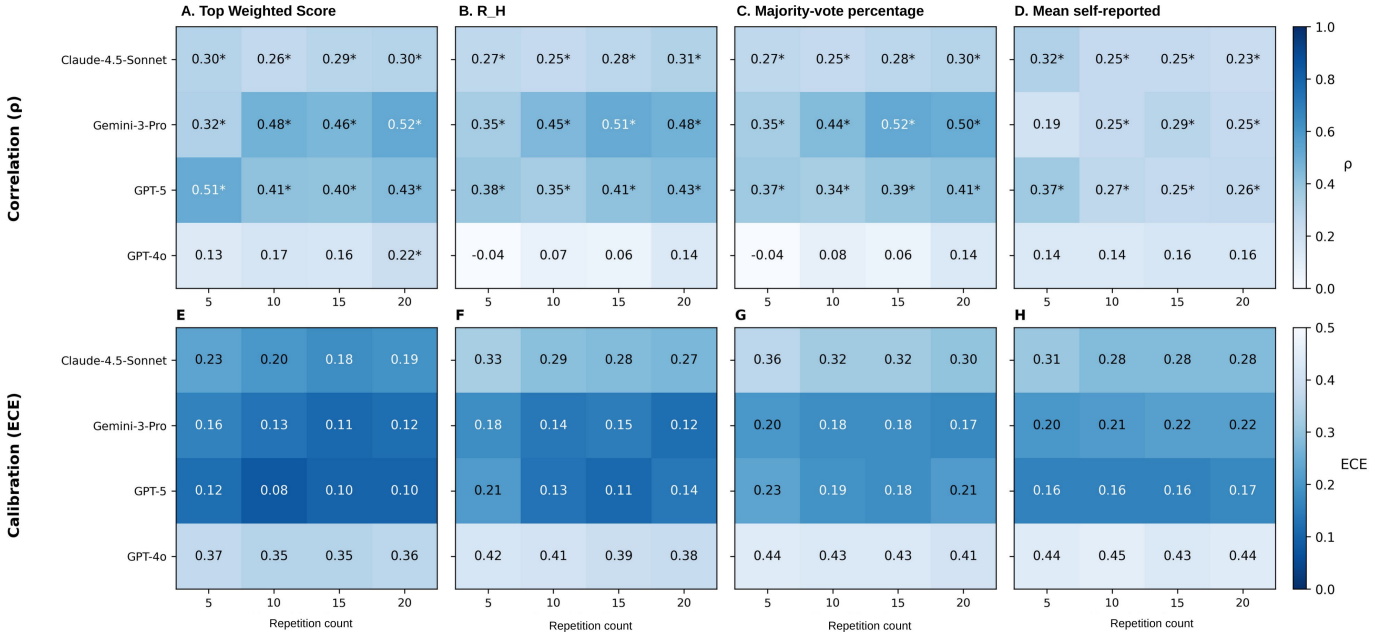
Model and metrics	ECE[b] (95% CI)	Brier score (95% CI)
Claude-4.5-Sonnet		
Self-reported (mean)	0.284 (0.195-0.389)	0.317 (0.253-0.385)
Self-reported (first)	0.340 (0.244-0.439)	0.339 (0.275-0.409)
R_H[c]	0.266 (0.191-0.375)	0.288 (0.216-0.363)
Majority-vote percentage	0.304 (0.226-0.403)	0.321 (0.247-0.399)
Top Weighted Score	0.192 (0.133-0.307)	0.259 (0.203-0.317)
Gemini-3-Pro		
Self-reported (mean)	0.216 (0.128-0.305)	0.243 (0.165-0.322)
Self-reported (first)	0.208 (0.120-0.298)	0.229 (0.152-0.313)
R_H	0.119 (0.078-0.210)	0.164 (0.109-0.226)
Majority-vote percentage	0.168 (0.109-0.264)	0.178 (0.117-0.244)
Top Weighted Score	0.122 (0.071-0.212)	0.163 (0.109-0.220)
GPT-5		
Self-reported (mean)	0.172 (0.097-0.273)	0.235 (0.171-0.303)
Self-reported (first)	0.176 (0.096-0.278)	0.230 (0.164-0.296)
R_H	0.140 (0.099-0.240)	0.191 (0.135-0.252)
Majority-vote percentage	0.206 (0.130-0.299)	0.219 (0.154-0.287)
Top Weighted Score	0.098 (0.074-0.211)	0.185 (0.140-0.235)
GPT-4o		
Self-reported (mean)	0.436 (0.329-0.543)	0.436 (0.349-0.523)
Self-reported (first)	0.468 (0.368-0.570)	0.459 (0.380-0.542)
R_H	0.377 (0.303-0.494)	0.397 (0.313-0.482)
Majority-vote percentage	0.413 (0.332-0.535)	0.435 (0.345-0.525)
Top Weighted Score	0.356 (0.272-0.470)	0.368 (0.295-0.440)

[a]Lower values indicate better calibration. ECE was calculated using 10 bins.
[b]ECE: expected calibration error.
[c]R_H: relative entropy.

Repetition Count Analysis

Correlation analysis stratified by repetition count revealed model-specific patterns in the relationship between consistency-based metrics and diagnostic accuracy (Table S4 in Multimedia Appendix 1). Gemini-3-Pro showed strengthening correlations with increasing repetitions, with majority-vote percentage and R_H improving from weak (ϱ=0.35, *P*<.001) at 5 repetitions to moderate (ϱ=0.51-0.52, *P*<.001) at 15 repetitions, and Top Weighted Score improving from weak (ϱ=0.32, *P*=.002) to moderate (ϱ=0.46, *P*<.001). Mean self-reported confidence in Gemini-3-Pro achieved statistical significance only after 10 repetitions. In contrast, Claude-4.5-Sonnet demonstrated significant weak correlations across all metrics after 5 repetitions. GPT-5 showed improvement in R_H from weak (ϱ=0.38, *P*<.001) at 5 repetitions to moderate (ϱ=0.41, *P*<.001) at 15 repetitions, while Top Weighted Score maintained moderate correlations across all repetition counts (ϱ=0.40-0.51; all *P*<.001). GPT-4o showed nonsignificant correlations across all metrics regardless of repetition count (Figure 4).

Figure 4. Heatmap of Spearman correlation (ϱ) and calibration (ECE) between diagnostic accuracy and confidence metrics across multimodal large language models by repetition count. ECE: expected calibration error; R_H: relative entropy.



Calibration metrics showed similar patterns with increasing repetitions (Table S5 in Multimedia Appendix 1). Top Weighted Score demonstrated decreasing ECE values across repetitions in Gemini-3-Pro (0.158 to 0.112), GPT-5 (0.118 to 0.095), and Claude-4.5-Sonnet (0.230 to 0.184).

Heatmaps display Spearman correlation coefficients (ϱ; top row, panels A-D) and ECE (bottom row, panels E-H) for 4 confidence metrics across repetition counts of 5, 10, 15, and 20. For correlation, darker shading indicates a stronger positive correlation (range 0-1); for ECE, darker shading indicates better calibration with lower error (range 0-0.5). Statistically significant correlations (P<.05) are marked with an asterisk. Shannon entropy exhibited inverse correlations with identical magnitudes to R_H values (data not shown).

Resource Use

Resource consumption varied substantially across models (Table S6 in Multimedia Appendix 1). GPT-4o demonstrated the highest efficiency, requiring a mean processing time of 6.38 (SD 2.26) seconds and a mean of 4353 (SD 1402) total tokens per case at 1 repetition, increasing to 59.43 (SD 12.46) seconds and 43,423 (SD 13,921) tokens at 10 repetitions, and 117.93 (SD 23.44) seconds and 86,877 (SD 27,869) tokens at 20 repetitions. Claude-4.5-Sonnet required a mean processing time of 29.48 (SD 6.37) seconds and a mean of 5551 (SD 1477) total tokens at 1 repetition, increasing to 288.27 (SD 44.76) seconds and 55,207 (SD 14,505) tokens at 10 repetitions, and 750.43 (SD 1694.80) seconds and 110,344 (SD 28,860) tokens at 20 repetitions. Gemini-3-Pro consumed a mean processing time of 54.75 (SD 29.79) seconds and a mean of 8774 (SD 2993) total tokens at 1 repetition, increasing to 543.87 (SD 256.35) seconds and 88,621 (SD 26,743) tokens at 10 repetitions, and 1103.97 (SD 528.83) seconds and 178,114 (SD 54,325) tokens at 20 repetitions. GPT-5 showed the highest resource consumption, requiring a mean processing time of 70.66 (SD 39.39) seconds and a mean of 6348 (SD 2026) total tokens at 1 repetition, increasing to 947.53 (SD 1808.41) seconds and 63,956 (SD 17,991) tokens at 10 repetitions, and 1768.43 (SD 2059.84) seconds and 127,879 (SD 35,210) tokens at 20 repetitions. Notably, sporadic extreme processing times exceeding 500 seconds per individual API call were observed in Claude-4.5-Sonnet (2/1880, 0.11%) calls and GPT-5 (5/1880, 0.27%) calls, whereas no such events occurred in Gemini-3-Pro or GPT-4o.

Discussion

Principal Findings

Our findings indicate that the Top Weighted Score, a composite metric integrating response consistency with self-reported confidence, provided the most consistent assessment of multimodal LLM output reliability for ultrasound-based radiological cases. Notably, it was the only metric to demonstrate statistically significant correlations across all 4 evaluated models, and it exhibited the best calibration in most models. In parallel, consistency-based metrics (R_H, majority-vote percentage, and Shannon entropy) showed strong discriminative performance in Gemini-3-Pro, GPT-5, and Claude-4.5-Sonnet when contrasted with self-reported confidence, underscoring the value of response agreement–derived signals for reliability estimation.

Comparison to Prior Work

These observations align with prior studies suggesting that consistency-based calibration approaches can outperform post hoc verbalized confidence methods for estimating LLM uncertainty and reliability [11,28,29]. In radiology, Huppertz et al [30] similarly reported no significant association between verbalized confidence and diagnostic accuracy, with

accuracy remaining below 50% even at the highest confidence scores. However, in our study, first-response confidence showed relatively high ROC AUC values in GPT-5 and Claude-4.5-Sonnet, whereas mean self-reported confidence demonstrated lower values; notably, Gemini-3-Pro showed the opposite pattern. This inconsistency suggests that the apparent first-response advantage may have been driven by sampling variability. Accordingly, repeated averaging of verbalized confidence scores represents an averaging of scores only partially aligned with actual correctness rather than a more accurate probability estimate [4], and this process may compress case-level variance and reduce discriminative ability. Given the intrinsically stochastic nature of LLM generation, reliance on a single initial output for confidence estimation poses significant risks in clinical settings.

Interestingly, the entropy-based metrics achieved lower ECE and Brier scores than the simple majority-vote percentage metric. This suggests that model response dispersion, rather than only the most frequent response, yields better-calibrated confidence estimates. Given the high κ values observed in our study, the benefit of entropy-based metrics is likely to increase in settings with greater response diversity, such as tasks with larger multiple-choice panels or free-text outputs. However, high interresponse agreement ($\kappa \approx 0.7$-0.8) and fixed sampling parameters (eg, temperature=1) may inflate confidence and reduce discriminative power.

Additionally, although not statistically significant, Gemini-3-Pro, the highest-performing model, showed a decrease in accuracy with majority voting at 15 and 20 repetitions. This may partly reflect the systematic nature of its errors across repetitions. When interresponse agreement is high (κ=0.79-0.80), incorrect responses tend to be consistent across repetitions, which may limit the corrective potential of majority voting and contribute to the marginal decrease in accuracy observed.

Despite broad improvements in capability across successive model releases, recent studies across domains indicate that newer models can retain a systematic tendency toward overconfidence [10,31,32]. Consistent with this literature, we observed that models frequently assigned high confidence to incorrect answers, which poses a direct challenge for deployment in clinical decision support, where confidently presented errors may be disproportionately persuasive and propagate into downstream decision-making [33,34]. Several recent studies have reported that hybrid approaches integrating consistency and verbalized confidence can outperform either method alone in certain models [4,28,35]. In our study, Top Weighted Score operationalizes a related principle by weighting candidate responses using both frequency and self-reported confidence, potentially mitigating 2 complementary limitations: (1) overconfidence inherent to verbalized estimates and (2) reduced sensitivity of pure consistency metrics when interresponse agreement is high. The consistent performance of Top Weighted Score across all tested models supports the utility of such integrative formulations for multimodal radiology tasks.

A major practical limitation of consistency-based estimation is resource intensity, as multiple independent model executions are required to compute agreement and dispersion metrics, increasing API costs and processing time [11]. In our study, processing time varied substantially by model and repetition count, with the burden particularly evident for reasoning models, where extended reasoning traces can materially increase latency and necessitate explicit cost-benefit trade-offs. To mitigate this overhead, adaptive sampling strategies have been proposed that achieve comparable accuracy with significantly reduced computational costs [35,36]. In our repetition-depth analysis, approximately 10 resampling runs per case were sufficient to stabilize discrimination and calibration estimates. Further research is needed to establish minimal sampling schedules for each model, along with inference acceleration strategies leveraging recent advances [37], to balance confidence estimation reliability with resource efficiency. Furthermore, sporadic extreme processing times were observed in Claude-4.5-Sonnet and GPT-5, likely reflecting potential API latency or timeout events. Although this processing time instability affected fewer than 0.3% of total API calls, it disproportionately inflated processing time variability and may represent a real-world barrier to clinical deployment, where predictable response times are essential.

Limitations

This study exhibited several limitations. First, our relatively small sample size (N=94) potentially restricts the generalizability of our findings, and the fixed 10-bin ECE may be unstable in this sample, as confidence score clustering may leave some bins sparsely populated. Second, because model performance was assessed with multiple-choice questions, the evaluation does not fully represent real-world clinical scenarios that typically require free-text responses. Third, our evaluation was limited to the most widely available closed-source models and did not include open-source models that might have high potential applicability in health care. Fourth, as the KSUM educational quizzes are publicly accessible online, there is a possibility that these materials may have been included in the training corpora of the evaluated LLMs. This potential data contamination could inflate model accuracy estimates. However, a temporal holdout analysis revealed no significant accuracy differences between precutoff and postcutoff cases for any model (Table S3 in Multimedia Appendix 1). Furthermore, our previous investigation using the same dataset demonstrated no significant performance differences for GPT-4o between cases published before versus after the model's knowledge cutoff date [38], suggesting that even if contamination occurred, its impact on model performance may be negligible given the vast scale of training parameters. Fifth, a practical limitation of consistency-based estimation is resource intensity, as multiple independent model executions are required, increasing costs and processing time. To address this, we conducted our repetition count analysis, and approximately 10 resampling runs per case appeared to be sufficient to stabilize discrimination and calibration estimates. Sixth, although processing times indicated that

reasoning models engaged in extended computation, the structured JSON output requirement may have partially constrained the models' chain-of-thought process, potentially affecting confidence calibration. Additionally, our prompt was designed such that confidence values were embedded as natural language strings within the JSON output rather than as direct numerical fields, which may introduce parsing instability if models generate slight textual variations. Seventh, our evaluation used general-purpose multimodal LLMs. Although recent studies have evaluated general-purpose multimodal LLMs on radiological cases, these investigations predominantly focus on diagnostic accuracy [3,21,38,39]. The development of radiologic domain-specific multimodal LLMs is in its early stages [40,41], and such models are not yet widely available commercially. Building upon these studies of general-purpose LLMs, our findings may contribute to the future development and evaluation of domain-specific expert models. Finally, the Top Weighted Score was developed post hoc on the study dataset, which introduces a potential risk of overfitting. Therefore, external validation using independent datasets is necessary to confirm the generalizability of this hybrid metric.

Future Directions

Several directions for future research emerge from our findings. First, validation with larger datasets across various radiological conditions would strengthen generalizability and potentially identify additional patterns in model performance and calibration. Second, incorporating free-text tasks would better approximate real-world clinical usage. Third, comparative studies should explore whether open-source models demonstrate similar confidence calibration patterns and assess their applicability in clinical environments. Fourth, further research is needed to establish minimal sampling schedules and develop adaptive sampling strategies to balance confidence estimation reliability with resource efficiency. Additionally, studies using models with accessible internal reasoning processes could enable direct entropy computation from log probabilities, potentially reducing computational overhead while maintaining calibration accuracy. Fifth, comparative studies evaluating human reader confidence alongside LLM confidence metrics during diagnostic tasks could provide valuable insights into the potential for LLMs to augment clinical decision-making. Finally, from a clinical perspective, confidence metrics could augment human decision-making by providing reliability indicators for LLM outputs; however, clinical validation studies are needed to evaluate the practical use of these metrics in real-world diagnostic settings.

Conclusions

In multimodal LLMs applied to ultrasound-based radiology cases, hybrid methods (Top Weighted Score) demonstrated significant associations across all evaluated models and appear to serve as more reliable indicators of diagnostic confidence compared to self-reported or consistency-based metrics alone, although the strength of this association varied across models, and external validation is warranted before broader clinical application.

Acknowledgments

The authors used Claude-4.5-Sonnet (Anthropic) for language editing and grammar checking during manuscript preparation. All content was verified and approved by the authors. The large language models evaluated in this study (GPT-4o, GPT-5, Claude-4.5-Sonnet, and Gemini-3-Pro) are described in the Methods section.

Funding

This work was supported by a grant from the National Research Foundation of Korea (NRF), funded by the Korean government (MSIT) (grant RS-2025-00516874).

Data Availability

The research data supporting the findings of this study are publicly available through the Korean Society of Ultrasound in Medicine (KSUM) [16]. This study used these open-source datasets for analysis. The underlying code for this study, including calibration analysis, is provided in Supplementary Material 1 in Multimedia Appendix 2 and will be made publicly available upon publication.

Authors' Contributions

TH and JS conceived and designed the study. TH, JHL, and KG acquired the data. TH, JHL, KG, and JS analyzed and interpreted the data. TH and JS prepared the first draft of the manuscript. All authors critically revised the manuscript for important intellectual content, approved the final manuscript, and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary table.
[DOCX File (Microsoft Word File), 35 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Calibration new analysis code.

[TXT File (Text, File), 8 KB-Multimedia Appendix 2]

References

1. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA. Sep 5, 2023;330(9):866-869. [doi: 10.1001/jama.2023.14217] [Medline: 37548965]
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. Aug 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
3. Zhou Y, Ong H, Kennedy P, et al. Evaluating GPT-v4 (GPT-4 with vision) on detection of radiologic findings on chest radiographs. Radiology. May 2024;311(2):e233270. [doi: 10.1148/radiol.233270] [Medline: 38713028]
4. Xiong M, Hu Z, Lu X, Li Y, Fu J, He J. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. Presented at: 12th International Conference on Learning Representations, ICLR 2024; May 7-11, 2024; Vienna, Austria. URL: https://researchportal.hkust.edu.hk/en/publications/can-llms-express-their-uncertainty-an-empirical-evaluation-of-con/? [Accessed 2026-05-22]
5. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Presented at: ICML'17: Proceedings of the 34th International Conference on Machine Learning; Aug 6-11, 2017; Sydney, Australia. URL: https://proceedings.mlr.press/v70/guo17a/guo17a.pdf [Accessed 2026-05-22]
6. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). NPJ Digit Med. Jul 8, 2024;7(1):183. [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
7. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. Presented at: ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning; Jun 19-24, 2016; New York, NY, USA. URL: https://proceedings.mlr.press/v48/gal16.pdf [Accessed 2026-05-22]
8. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell. Jan 7, 2019;1(1):20-23. [doi: 10.1038/s42256-018-0004-1]
9. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med. Jan 5, 2021;4(1):4. [doi: 10.1038/s41746-020-00367-3] [Medline: 33402680]
10. Omar M, Agbareia R, Glicksberg BS, Nadkarni GN, Klang E. Benchmarking the confidence of large language models in clinical questions. medRxiv. Preprint posted online on Sep 10, 2024. [doi: 10.1101/2024.08.11.24311810]
11. Savage T, Wang J, Gallo R, et al. Large language model uncertainty measurement and calibration for medical diagnosis and treatment. medRxiv. Preprint posted online on Jun 7, 2024. [doi: 10.1101/2024.06.06.24308399]
12. Tian K, Mitchell E, Zhou A, et al. Just ask for calibration: strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Dec 6-10, 2023; Singapore. [doi: 10.18653/v1/2023.emnlp-main.330]
13. Yang D, Tsai YHH, Yamada M. On verbalized confidence scores for LLMs. arXiv. Preprint posted online on Dec 19, 2024. [doi: 10.48550/arXiv.2412.14737]
14. Raj H, Rosati D, Majumdar S. Measuring reliability of large language models through semantic consistency. arXiv. Preprint posted online on Nov 10, 2022. [doi: 10.48550/arXiv.2211.05853]
15. Fadeeva E, Rubashevskii A, Shelmanov A, et al. Fact-checking the output of large language models via token-level uncertainty quantification. Presented at: Findings of the Association for Computational Linguistics: ACL 2024; Aug 11-16, 2024; Bangkok, Thailand. [doi: 10.18653/v1/2024.findings-acl.558]
16. Education. Korean Society of Ultrasound in Medicine (KSUM). URL: https://www.ksum.or.kr/education/index.php [Accessed 2026-05-23]
17. Singh A, Fry A, Perelman A, Tart A, Ganesh A, El-Kishky A. OpenAI GPT-5 system card. arXiv. Preprint posted online on Dec 19, 2025. [doi: 10.48550/arXiv.2601.03267]
18. Introducing Claude Sonnet 4.5. Anthropic. 2025. URL: https://www.anthropic.com/news/claude-sonnet-4-5 [Accessed 2026-05-22]
19. Model cards. Google DeepMind. 2025. URL: https://deepmind.google/models/model-cards/ [Accessed 2026-05-22]
20. Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A, Clark A. GPT-4o system card. arXiv. Preprint posted online on Oct 25, 2024. [doi: 10.48550/arXiv.2410.21276]
21. 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. Jul 2024;312(1):e240273. [doi: 10.1148/radiol.240273] [Medline: 38980179]
22. Cochran WG. The comparison of percentages in matched samples. Biometrika. Dec 1950;37(3-4):256-266. [doi: 10.1093/biomet/37.3-4.256] [Medline: 14801052]
23. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. Sep 1988;44(3):837-845. [doi: 10.2307/2531595] [Medline: 3203132]

24. Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. Anesth Analg. May 2018;126(5):1763-1768. [doi: 10.1213/ANE.0000000000002864] [Medline: 29481436]
25. Rivera M, Godbout JF, Rabbany R, Pelrine K. Combining confidence elicitation and sample-based methods for uncertainty quantification in misinformation mitigation. Presented at: Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024); Mar 22, 2024; St Julians, Malta. [doi: 10.18653/v1/2024.uncertainlp-1.12]
26. Brier GW. Verification of forecasts expressed in terms of probability. Mon Wea Rev. Jan 1950;78(1):1-3. [doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2]
27. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. Mar 1977;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
28. de Oliveira R, Garber M, Gwinnutt JM, et al. A study of calibration as a measurement of trustworthiness of large language models in biomedical natural language processing. JAMIA Open. 2025;8(4):ooaf058. [doi: 10.1093/jamiaopen/ooaf058] [Medline: 40655536]
29. Lyu Q, Shridhar K, Malaviya C, et al. Calibrating large language models with sample consistency. Presented at: Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25); Feb 25 to Mar 4, 2025; Philadelphia, Pennsylvania, USA. [doi: 10.1609/aaai.v39i18.34120]
30. Huppertz MS, Siepmann R, Topp D, et al. Revolution or risk? Assessing the potential and challenges of GPT-4V in radiologic image interpretation. Eur Radiol. Mar 2025;35(3):1111-1121. [doi: 10.1007/s00330-024-11115-6] [Medline: 39422726]
31. Naderi N, Safavi-Naini SAA, Savage T, Atf Z, Lewis P, Nadkarni G. Self-reported confidence of large language models in gastroenterology: analysis of commercial, open-source, and quantized models. arXiv. Preprint posted online on May 24, 2025. [doi: 10.48550/arXiv.2503.18562]
32. Vashurin R, Fadeeva E, Vazhentsev A, et al. Benchmarking uncertainty quantification methods for large language models with LM-Polygraph. Trans Assoc Comput Linguist. Mar 19, 2025;13:220-248. [doi: 10.1162/tacl_a_00737]
33. Omar M, Nassar S, Hijazi K, Glicksberg BS, Nadkarni GN, Klang E. Generating credible referenced medical research: a comparative study of OpenAI's GPT-4 and Google's Gemini. Comput Biol Med. Feb 2025;185:109545. [doi: 10.1016/j.compbiomed.2024.109545] [Medline: 39667055]
34. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care. Mar 21, 2023;27(1):120. [doi: 10.1186/s13054-023-04393-x] [Medline: 36945051]
35. Taubenfeld A, Sheffer T, Ofek E, et al. Confidence improves self-consistency in LLMs. Presented at: Findings of the Association for Computational Linguistics; Jul 27 to Aug 1, 2025; Vienna, Austria. [doi: 10.18653/v1/2025.findings-acl.1030]
36. Aggarwal P, Madaan A, Yang Y. Let's sample step by step: adaptive-consistency for efficient reasoning and coding with LLMs. Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Dec 6-10, 2023; Singapore. [doi: 10.18653/v1/2023.emnlp-main.761]
37. He C, Huang Y, Mu P, Miao Z, Xue J, Ma L. WaferLLM: large language model inference at Wafer scale. Presented at: 19th USENIX Symposium on Operating Systems Design and Implementation (OSDI '25); Jul 7-9, 2025; Boston, Massachusetts, USA. URL: https://www.usenix.org/system/files/osdi25-he.pdf [Accessed 2026-05-22]
38. 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. May 2025;44(3):220-231. [doi: 10.14366/usg.25012] [Medline: 40235070]
39. 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. Jan 2025;314(1):e240689. [doi: 10.1148/radiol.240689] [Medline: 39835982]
40. Hong EK, Ham J, Roh B, et al. Diagnostic accuracy and clinical value of a domain-specific multimodal generative AI model for chest radiograph report generation. Radiology. Mar 2025;314(3):e241476. [doi: 10.1148/radiol.241476] [Medline: 40131111]
41. Hong EK, Roh B, Park B, et al. Value of using a generative AI model in chest radiography reporting: a reader study. Radiology. Mar 2025;314(3):e241646. [doi: 10.1148/radiol.241646] [Medline: 40067108]

Abbreviations

API: application programming interface
AUC: area under the curve
ECE: expected calibration error
KSUM: Korean Society of Ultrasound in Medicine
LLM: large language model
R_H: relative entropy

ROC: receiver operating characteristic

Edited by Andrew Coristine; peer-reviewed by Avijit Mitra, Danqing Hu, Elliot L Epstein, Merlijn Sevenster; submitted 25.Oct.2025; final revised version received 15.May.2026; accepted 18.May.2026; published 02.Jun.2026

Please cite as:
Han T, Shin J, Lee JH, Gu K
Confidence Measurement Metrics in Multimodal Large Language Models for Ultrasound-Based Radiology Cases: Comparative Evaluation Study of Self-Reported, Consistency-Based, and Hybrid Methods
J Med Internet Res 2026;28:e86498
URL: https://www.jmir.org/2026/1/e86498
doi: 10.2196/86498