

Medical LLM Metacognition Is Multidimensional: A MetaMedQA Reanalysis of Confidence, Missing-Answer Recognition, and Unknown-Answer Detection

Ahmad M. Nazzal
a-nazzal[at]outlook[dot]com

Abstract

Recent work using MetaMedQA argued that large language models (LLMs) lack essential metacognition for reliable medical reasoning. However, metacognition is not a single construct: confidence–correctness discrimination, missing-answer recognition, unknown-answer detection, and abstention behavior may dissociate. Here, we reanalyzed MetaMedQA using a confidence-centered evaluation framework previously developed for a controlled clinical-evidence benchmark. Two GPT-family models, **gpt-4.1-nano** and **gpt-5.5**, were evaluated on 1373 MetaMedQA items using structured outputs containing an answer, numerical confidence, and a more-information-needed judgment. **gpt-4.1-nano** achieved 56.4% accuracy, mean confidence of 79.7%, Brier score of 0.318, expected calibration error of 0.276, and AUROC2 of 0.582. Missing-answer recall was 19.1%, and unknown/unanswerable recall was 25.9%. **gpt-5.5** improved substantially, achieving 84.9% accuracy, mean confidence of 91.2%, Brier score of 0.112, expected calibration error of 0.062, and AUROC2 of 0.819. Missing-answer recall increased to 67.8%, and unknown/unanswerable recall to 56.2%. Nevertheless, incorrect responses from **gpt-5.5** still received high mean confidence. These results suggest that medical-LLM metacognition is better understood as a set of dissociable behavioral capacities rather than as a single absent-or-present property. Stronger models can show improved confidence–correctness discrimination and calibration, while still retaining clinically relevant failures in missing-answer and unknown-answer recognition.

1 Introduction

Large language models (LLMs) are increasingly evaluated for medical reasoning, yet their clinical usefulness depends on more than answer accuracy. A system that is wrong but uncertain may invite review or further information gathering. A system that is wrong while confident may be more dangerous. This distinction has motivated growing interest in whether LLMs can monitor the reliability of their own outputs.

Griot et al. introduced MetaMedQA and concluded that LLMs lack essential metacognition for reliable medical reasoning [2]. Their benchmark emphasized missing-answer recognition, unknown-answer detection, and confidence behavior in medical question answering. This is an important operationalization of metacognition, but it is not the only one. In cognitive science, metacognitive sensitivity is often defined behaviorally as the degree to which confidence tracks correctness [1]. Signal-detection approaches further distinguish first-order task performance from second-order confidence sensitivity and metacognitive efficiency [3].

In prior work, we developed a controlled clinical-evidence benchmark in which diagnostic evidence was parametrically varied to test whether medical-LLM confidence tracked evidence strength, information quality, and correctness [4]. That benchmark suggested that LLM confidence can be

evidence-sensitive, but locally fragile. The present study asks whether the same confidence-centered analysis can clarify the interpretation of MetaMedQA.

The central question is not whether medical LLMs “have” or “lack” metacognition in a broad human sense. Instead, we ask whether different metacognitive-like behaviors dissociate: confidence–correctness discrimination, missing-answer recognition, unknown-answer detection, and the tendency to request more information. We evaluate two GPT-family models, `gpt-4.1-nano` and `gpt-5.5`, on the public MetaMedQA dataset using structured confidence elicitation and AUROC2-based confidence–correctness analysis.

2 Methods

2.1 Dataset

We used the public MetaMedQA dataset released by Griot et al. [2]. MetaMedQA contains medical multiple-choice questions designed to evaluate not only answer selection, but also missing-answer recognition and unknown or unanswerable items. The present analysis included 1373 items. Items were grouped into three operational categories: standard answer-present questions, missing-answer or none-of-the-above questions, and unknown or unanswerable questions.

2.2 Models and prompting

We evaluated `gpt-4.1-nano` and `gpt-5.5` using stateless API calls. Each item was presented as a multiple-choice medical question. The model was instructed to choose the best answer from the listed options, to select the appropriate none-of-the-above or unknown/cannot-answer option when applicable, and to return structured JSON only. The required output fields were: selected answer letter, confidence from 0 to 100, and a Boolean `more_information_needed` judgment. Confidence was defined as the model’s estimated probability that its selected answer was correct based only on the provided question and options.

2.3 Outcome measures

The primary first-order outcome was answer accuracy. For missing-answer items, accuracy corresponds to recall of the none-of-the-above option. For unknown or unanswerable items, accuracy corresponds to correct recognition of the unknown/cannot-answer option. Second-order outcomes included mean confidence, Brier score, expected calibration error (ECE), and AUROC2. AUROC2 treats correctness as the binary target and confidence as the score, summarizing how well confidence discriminates correct from incorrect responses [1]. We also summarized mean confidence separately for correct and incorrect responses and computed `more_information_needed` rates by item type.

2.4 Reproducibility and ethics

No patient data were used in this analysis. All evaluated items came from the publicly available MetaMedQA dataset. Code for the structured prompting and analysis pipeline is available in the project repository. This analysis is intended for model evaluation only and does not provide medical advice.

3 Results

3.1 Overall performance and confidence–correctness discrimination

Both models produced valid structured outputs for all 1373 MetaMedQA items. Overall performance differed substantially between models (Table 1). **gpt-4.1-nano** achieved 56.4% accuracy, with mean confidence of 79.7%, Brier score of 0.318, ECE of 0.276, and AUROC2 of 0.582. Thus, confidence only weakly discriminated correct from incorrect answers.

gpt-5.5 achieved 84.9% accuracy, with mean confidence of 91.2%, Brier score of 0.112, ECE of 0.062, and AUROC2 of 0.819. This indicates substantially stronger confidence–correctness discrimination and better global calibration than **gpt-4.1-nano**. However, incorrect responses from **gpt-5.5** still received high confidence, with mean confidence of 82.6%, compared with 92.7% for correct responses. Therefore, higher AUROC2 should be interpreted as better ranking of correct versus incorrect responses, not as absence of overconfidence on errors (Figure 1).

Table 1: Overall MetaMedQA performance.

Model	Items	Accuracy	Mean conf.	Brier	ECE	AUROC2
gpt-4.1-nano	1373	0.564	0.797	0.318	0.276	0.582
gpt-5.5	1373	0.849	0.912	0.112	0.062	0.819

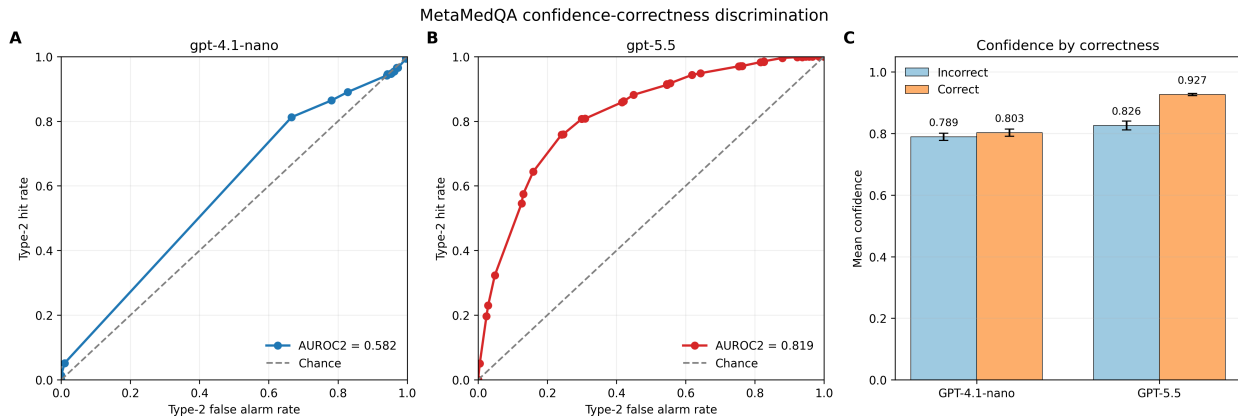


Figure 1: **MetaMedQA confidence–correctness discrimination.** (A) Type-2 ROC curve for **gpt-4.1-nano**. Confidence only weakly discriminated correct from incorrect responses. (B) Type-2 ROC curve for **gpt-5.5**, showing stronger confidence–correctness discrimination. (C) Mean confidence for correct and incorrect responses by model. Although **gpt-5.5** showed higher AUROC2, incorrect responses still received high confidence, indicating residual overconfidence on errors. Error bars indicate 95% confidence intervals.

3.2 Item-type performance

Performance varied strongly by item type (Table 2; Figure 2). On standard answer-present questions, **gpt-4.1-nano** achieved 64.8% accuracy, while **gpt-5.5** achieved 91.0%. Missing-answer recognition was much weaker: **gpt-4.1-nano** identified only 19.1% of missing-answer items correctly, whereas **gpt-5.5** identified 67.8%. Unknown or unanswerable recognition was also limited, with recall of 25.9% for **gpt-4.1-nano** and 56.2% for **gpt-5.5**.

The `more_information_needed` signal also differed by item type. `gpt-4.1-nano` requested more information in 43.8% of unknown/unanswerable cases but only 1.7% of missing-answer cases. `gpt-5.5` requested more information in 56.2% of unknown/unanswerable cases but never did so in missing-answer cases. This suggests that missing-answer recognition and explicit information-sufficiency signaling are behaviorally distinct from unknown-answer detection.

Table 2: MetaMedQA performance by item type.

Model	Item type	n	Accuracy/recall	Mean conf.	More-info rate
<code>gpt-4.1-nano</code>	Answer-present	1096	0.648	0.830	0.013
<code>gpt-4.1-nano</code>	Missing-answer	115	0.191	0.812	0.017
<code>gpt-4.1-nano</code>	Unknown/unanswerable	162	0.259	0.564	0.438
<code>gpt-5.5</code>	Answer-present	1096	0.910	0.917	0.005
<code>gpt-5.5</code>	Missing-answer	115	0.678	0.879	0.000
<code>gpt-5.5</code>	Unknown/unanswerable	162	0.562	0.896	0.562

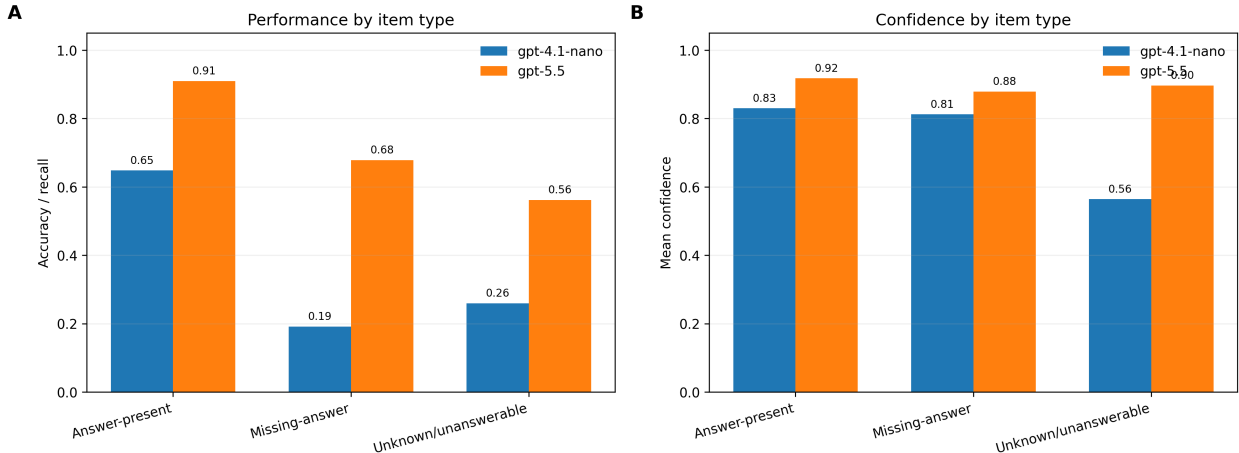


Figure 2: **MetaMedQA performance and confidence by item type.** (A) Accuracy or recall by item type for `gpt-4.1-nano` and `gpt-5.5`. (B) Mean confidence by item type. Although `gpt-5.5` improved substantially across item types, confidence remained high in missing-answer and unknown/unanswerable conditions.

4 Discussion

This MetaMedQA reanalysis supports a multidimensional interpretation of medical-LLM metacognition. `gpt-4.1-nano` showed weak confidence-correctness discrimination and poor missing-answer and unknown-answer recognition. This pattern is broadly consistent with the concern raised by Griot et al. that medical LLMs may fail to recognize gaps in their own answers [2]. However, `gpt-5.5` showed markedly better accuracy, calibration, and AUROC2, suggesting that these capacities can improve substantially with stronger models.

At the same time, the failures did not disappear. Missing-answer and unknown-answer recognition remained incomplete even for `gpt-5.5`. Moreover, incorrect responses still received high confidence. This distinction is important because AUROC2 measures ranking: whether correct responses tend to receive higher confidence than incorrect responses. It does not imply that incorrect responses receive

low confidence in absolute terms. Thus, a model may show improved metacognitive sensitivity while still producing high-confidence errors.

The results help reconcile two views of medical-LLM metacognition. In a controlled evidence-gradient benchmark, confidence can track evidence strength and correctness [4]. In MetaMedQA, however, missing-answer recognition and unknown-answer detection remain difficult. These are not contradictory findings. They suggest that metacognition-like behavior in LLMs is component-specific. Confidence–correctness discrimination, calibration, abstention, and missing-answer recognition should be evaluated separately rather than collapsed into a single claim that models either have or lack metacognition.

This framing has practical implications for medical AI evaluation. A single confidence score or a single benchmark accuracy value is insufficient. Medical LLMs should be tested on whether confidence tracks correctness, whether confidence decreases under uncertainty, whether missing answers are detected, and whether the model can signal that more information is needed. Different models may improve on these dimensions at different rates.

5 Limitations

This study is a reanalysis of a public benchmark using two GPT-family models. It does not exhaustively compare all model families or prompting strategies. MetaMedQA is also a multiple-choice benchmark, and its missing-answer and unknown-answer tasks differ from real clinical uncertainty. The confidence values were explicitly elicited, and such expressed confidence may differ from internal uncertainty proxies or sampling-based measures. Finally, the use of the term metacognition is behavioral and operational: we do not claim that LLMs possess human-like introspection or reflective consciousness.

6 Conclusion

Medical-LLM metacognition should not be treated as a single absent-or-present property. On MetaMedQA, `gpt-5.5` showed substantially stronger confidence–correctness discrimination, calibration, and missing-answer recognition than `gpt-4.1-nano`. Yet even the stronger model retained high-confidence errors and incomplete recognition of missing or unknown answers. These results support a multidimensional evaluation framework in which confidence–correctness discrimination, calibration, missing-answer recognition, and abstention behavior are measured separately.

Data and code availability

The analysis code and related benchmark materials are available at: <https://github.com/anazz-dev/medical-llm-metacognition-benchmark>. MetaMedQA is available from the public Hugging Face repository released by Griot et al. [2].

References

- [1] Stephen M. Fleming and Hakwan C. Lau. How to measure metacognition. *Frontiers in Human Neuroscience*, 8:443, 2014. doi: 10.3389/fnhum.2014.00443.

- [2] M. Griot, C. Hemptinne, J. Vanderdonckt, and D. Yuksel. Large language models lack essential metacognition for reliable medical reasoning. *Nature Communications*, 16:642, 2025. doi: 10.1038/s41467-024-55628-6.
- [3] Brian Maniscalco and Hakwan Lau. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1):422–430, 2012. doi: 10.1016/j.concog.2011.09.021.
- [4] Ahmad M. Nazzal. Large language models show metacognitive sensitivity in medical reasoning: Evidence-sensitive confidence in a controlled clinical benchmark. *Preprint*, 2026. Code and materials available at <https://github.com/anazz-dev/medical-llm-metacognition-benchmark>.