Beyond Multiple Choice: Calibration-Aware Evaluation Reveals Overconfident Clinical Reasoning in Large Language Models

Wei-Lun Cheng[a,*], Hsuan-Chia Yang[a]

[a] *Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan*

Abstract

Large Language Models (LLMs) are increasingly proposed for clinical decision support, yet their evaluation relies almost exclusively on multiple-choice question (MCQ) benchmarks. This study reveals a critical evaluation gap: the *Option Bias*—the systematic overestimation of clinical reasoning ability conferred by the MCQ format. We evaluated GPT-4o on the complete MedQA (USMLE) test set (n = 1,273) in both MCQ and open-ended formats, measuring accuracy, verbalized confidence, and three-tier clinical correctness (exact, partial, incorrect). MCQ accuracy was 87.8%, while open-ended exact correctness (Level A) was only 56.2%, yielding an Option Bias of 31.7 percentage points (relative: 36.0%). Critically, while the model was well-calibrated under MCQ format (ECE = 0.029), open-ended calibration was severely degraded (ECE = 0.364), with the model maintaining high confidence (mean 92.5%) despite substantially lower accuracy. We introduce *Safety-Weighted Expected Calibration Error* (SW-ECE), a novel metric that weights calibration error by clinical domain severity, amplifying miscalibration in safety-critical domains such as pharmacology. These findings demonstrate that MCQ-based evaluations are insufficient for assessing clinical AI readiness, and that overconfident incorrect responses—particularly in drug-related queries—pose direct risks to patient safety when deployed in clinical decision support systems (CDSS). We propose a three-tier clinical AI screening framework integrating competence, self-awareness, and robustness assessment for trustworthy clinical AI deployment.

Keywords: Large Language Models, Clinical Decision Support, Calibration, Medical Question Answering, Patient Safety, Confidence Estimation

1. Introduction

The rapid advancement of Large Language Models (LLMs) has generated considerable interest in their potential for clinical applications, including diagnostic reasoning, drug information retrieval, and clinical

*Corresponding author
Email addresses: first.author@tmu.edu.tw (First Author), hsuan-chia.yang@tmu.edu.tw (Hsuan-Chia Yang)

decision support systems (CDSS). Recent studies have demonstrated that models such as GPT-4 and Med-PaLM 2 achieve scores exceeding the passing threshold on the United States Medical Licensing Examination (USMLE) (Nori et al., 2023; Singhal et al., 2023), prompting discussions about their integration into clinical workflows.

However, a fundamental methodological concern underlies these impressive benchmark results: **virtually all medical LLM evaluations employ multiple-choice question (MCQ) formats**. Standard benchmarks—MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and MMLU-Med (Hendrycks et al., 2021)—present 4–5 candidate answers, requiring only selection rather than generation. In authentic clinical practice, physicians face open-ended diagnostic and therapeutic questions without predefined answer sets: "What is the diagnosis?" not "Which of these four diagnoses is correct?"

This format mismatch creates what we term the **MCQ Illusion**—an inflated perception of clinical competence arising from three mechanisms: (1) elimination strategy, where models use pattern matching to exclude implausible options; (2) answer anchoring, where candidate answers prime relevant knowledge retrieval; and (3) format familiarity from extensive MCQ content in training corpora. If substantial performance degradation occurs when the MCQ scaffolding is removed, the clinical validity of current evaluations must be questioned.

Beyond accuracy, the **calibration** of model confidence is equally critical for clinical deployment. A CDSS that provides confident but incorrect recommendations creates two dangerous failure modes. First, clinicians under time pressure may trust high-confidence AI suggestions, leading to medical errors (*confident error*). Second, accumulated false-positive high-confidence alerts induce *alert fatigue* (AlertFatigue, 2019), causing clinicians to systematically ignore AI recommendations—thereby defeating the system's purpose. Yang et al.'s research on diagnostic recommendations and drug safety in CDSS underscores that **confidence calibration is a prerequisite for safe AI-assisted clinical decision-making**, particularly in pharmacology where dosing errors can be fatal.

This study makes the following contributions:

1. We quantify **Option Bias**—the accuracy gap between MCQ and open-ended formats—using GPT-4o on the full MedQA test set ($n = 1{,}273$), demonstrating a 31.7 percentage-point accuracy drop when MCQ scaffolding is removed.
2. We reveal a critical **calibration asymmetry**: models that appear well-calibrated under MCQ format (ECE = 0.029) exhibit severe miscalibration under open-ended format (ECE = 0.364).
3. We introduce **Safety-Weighted ECE (SW-ECE)**, a novel calibration metric that weights miscalibration by clinical domain severity, prioritizing safety-critical domains such as pharmacology and emergency medicine.
4. We propose a **three-tier screening framework** for evaluating clinical AI readiness, integrating

competence (Option Bias), self-awareness (calibration), and robustness (multi-model cross-supervision) assessment.

2. Related Work

2.1. Medical LLM Evaluation

Medical LLM evaluation has predominantly relied on MCQ-based benchmarks. Jin et al. (2021) introduced MedQA, a USMLE-style dataset with 1,273 test questions. Pal et al. (2022) created MedMCQA with 4,183 test questions spanning 21 medical subjects. Hendrycks et al. (2021) included six medical subtasks in MMLU. These benchmarks have been widely adopted for evaluating models including GPT-4 (Nori et al., 2023), Med-PaLM (Singhal et al., 2023), and BioMistral (Labrak et al., 2024).

Despite the proliferation of MCQ-based evaluations, systematic investigation of format-dependent performance in medical AI remains limited. While Singhal et al. (2023) included some open-ended evaluation in Med-PaLM, they did not systematically quantify the MCQ–open-ended accuracy gap across medical subdomains.

2.2. LLM Confidence Calibration

Guo et al. (2017) demonstrated that modern neural networks, despite increasing accuracy, exhibit worsening calibration. Kadavath et al. (2022) showed that LLMs possess some metacognitive ability—they "mostly know what they know"—but significant gaps in self-assessment persist. Tian et al. (2023) demonstrated that prompting strategies can improve verbalized confidence calibration. Wang et al. (2023) proposed self-consistency as a calibration mechanism through repeated sampling.

However, prior calibration studies have not addressed the clinical domain's unique requirement: **not all miscalibration is equally dangerous**. A miscalibrated answer about embryology has minimal direct patient impact, whereas a miscalibrated pharmacology response can be lethal.

2.3. Clinical Decision Support and Alert Fatigue

The integration of AI into CDSS introduces the alert fatigue problem: excessive or unreliable alerts lead clinicians to ignore all notifications, including critical ones. This is directly relevant to LLM confidence calibration—an overconfident model generates many high-confidence alerts, a proportion of which are incorrect, gradually eroding clinician trust. Yang et al.'s work on drug safety and diagnostic recommendations in CDSS at Taipei Medical University provides direct motivation for safety-weighted calibration metrics.

3. Methodology

3.1. Dataset

We use the complete MedQA (USMLE) test set (Jin et al., 2021), containing 1,273 multiple-choice questions with four options each. Questions span multiple medical disciplines including anatomy, biochemistry, pharmacology, pathology, internal medicine, surgery, pediatrics, and psychiatry. All 1,273 questions were evaluated in both MCQ and open-ended formats, yielding 2,546 model inferences plus 1,273 clinical judgment evaluations (3,819 total API calls).

3.2. Experimental Design

Each question was presented to GPT-4o (gpt-4o, OpenAI) in two formats:

MCQ Format.. The original question with all four options (A/B/C/D), with instructions to select the correct option and provide a confidence percentage (0–100%).

Open-Ended Format.. The identical question stem with all options removed, with instructions to provide a direct answer and a confidence percentage.

All inferences used temperature = 0 for deterministic output. The model was additionally instructed to provide a verbalized confidence score as a percentage (0–100%) following the approach of Tian et al. (2023).

3.3. Three-Tier Clinical Judgment System

Open-ended responses were evaluated using a three-tier judgment system:

- **Level A (Clinically Correct):** Semantically equivalent to the reference answer; clinically actionable.
- **Level B (Partially Correct):** Correct direction but imprecise; e.g., "myocardial infarction" when the reference is "inferior STEMI."
- **Level C (Clinically Incorrect):** Clinically distinct from the reference answer; could lead to incorrect clinical action.

Judgment was performed using GPT-4o as an automated clinical judge, prompted with the question, reference answer, and model response. Future work will include human validation by two independent clinical experts with Cohen's κ inter-rater agreement to assess potential auto-evaluation bias.

3.4. Option Bias Metrics

Option Bias..

$$\text{Option Bias} = \text{Acc}_{\text{MCQ}} - \text{Acc}_{\text{OE}} \tag{1}$$

where Acc_{OE} is the proportion of Level A (clinically correct) open-ended responses.

Adjusted Option Bias..

$$\text{Adjusted Option Bias} = \text{Acc}_{\text{MCQ}} - (\text{Level A} + 0.5 \times \text{Level B}) \tag{2}$$

giving 50% credit for partially correct responses.

Relative Option Bias..

$$\text{Relative Option Bias} = \frac{\text{Acc}_{\text{MCQ}} - \text{Acc}_{\text{OE}}}{\text{Acc}_{\text{MCQ}}} \times 100\% \tag{3}$$

representing the proportion of MCQ performance attributable to the "option crutch" effect.

3.5. Calibration Metrics

Expected Calibration Error (ECE)..

$$\text{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} \left|\text{acc}(b) - \text{conf}(b)\right| \tag{4}$$

where predictions are binned into $B = 10$ equal-width confidence bins, n_b is the number of samples in bin b, $\text{acc}(b)$ is the observed accuracy, and $\text{conf}(b)$ is the mean confidence.

Safety-Weighted ECE (SW-ECE).. We introduce a clinically-motivated variant that incorporates domain-specific severity weights:

$$\text{SW-ECE} = \sum_{b=1}^{B} \frac{\sum_{i \in b} w_i}{\sum_{i=1}^{N} w_i} \left|\text{acc}(b) - \text{conf}(b)\right| \tag{5}$$

where w_i is the safety weight assigned by medical subdomain (Table 1). This formulation ensures that miscalibration in pharmacology (weight 3.0) contributes three times more to the aggregate metric than equivalent miscalibration in basic sciences (weight 1.0).

Brier Score..

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^{N} \left(\text{conf}_i - \text{correct}_i\right)^2 \tag{6}$$

where $\text{correct}_i \in \{0, 1\}$, providing a joint measure of accuracy and calibration.

4. Results

4.1. Option Bias: MCQ vs. Open-Ended Performance

Table 2 summarizes the performance comparison between MCQ and open-ended formats. GPT-4o achieved 87.8% accuracy on MCQ format but only 56.2% Level A (clinically exact) accuracy on open-ended format, yielding an Option Bias of 31.7 percentage points. An additional 24.6% of open-ended responses were Level B (partially correct), while 19.2% were Level C (clinically incorrect).

Table 1: Safety weight assignments by medical subdomain for SW-ECE calculation.

Subdomain	**Weight**	**Rationale**
Pharmacology	3.0	Medication errors can be fatal
Emergency Medicine	3.0	Delayed treatment can be fatal
Pediatrics	2.5	Dosing errors in children critical
OB/GYN	2.5	Pregnancy drug safety
Internal Medicine	2.0	Chronic disease management
Surgery	2.0	Surgical decision impact
Pathology	1.5	Diagnostic interpretation
Psychiatry	1.5	Treatment plan impact
Basic Sciences	1.0	Indirect clinical impact

Table 2: Option Bias analysis: MCQ vs. Open-Ended performance of GPT-4o on MedQA ($n = 1{,}273$).

Metric	**Value**
MCQ Accuracy	87.8%
Open-Ended Level A (Exact)	56.2%
Open-Ended Level B (Partial)	24.6%
Open-Ended Level C (Wrong)	19.2%
Option Bias (absolute)	31.7%
Adjusted Option Bias	19.4%
Relative Option Bias	36.0%

The Adjusted Option Bias, giving 50% credit for Level B responses, was 19.4%, and the Relative Option Bias was 36.0%, indicating that over one-third of the model's MCQ performance is attributable to the option scaffolding rather than genuine clinical reasoning.

Figure 1 visualizes the performance gap between formats.

4.2. Calibration Analysis

MCQ Calibration.. Under MCQ format, GPT-4o demonstrated good calibration with ECE = 0.029, SW-ECE = 0.029, and Brier Score = 0.103. Mean confidence was 90.7% against 87.8% accuracy, indicating mild overconfidence (Figure 2).

Open-Ended Calibration.. Under open-ended format, calibration degraded dramatically: ECE = 0.364, SW-ECE = 0.364, and Brier Score = 0.372. Strikingly, mean open-ended confidence (92.5%) was *higher* than

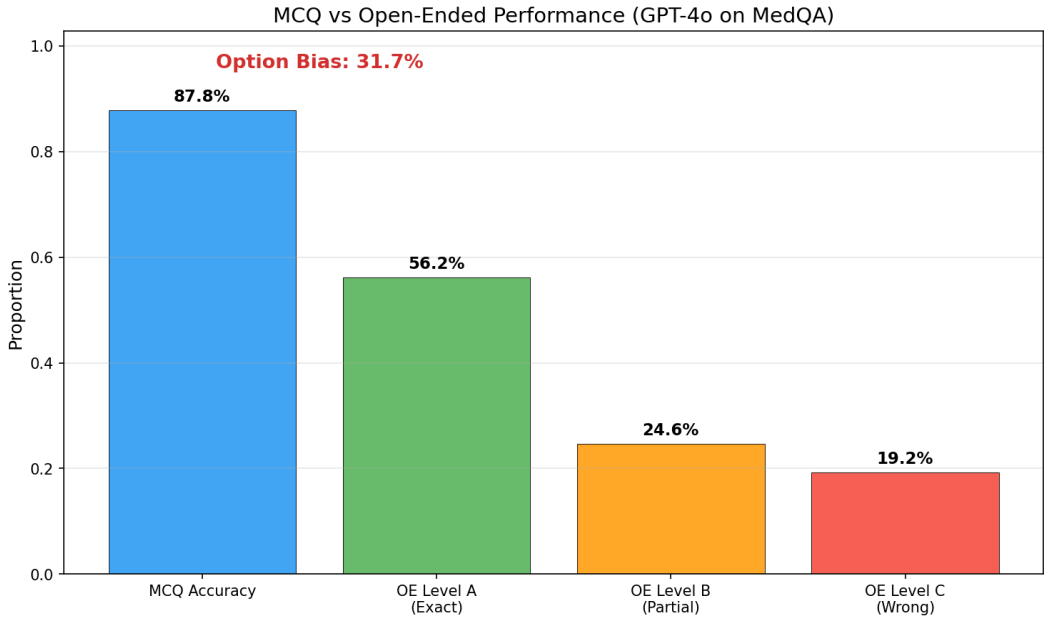

Figure 1: MCQ vs. Open-Ended performance comparison for GPT-4o on MedQA ($n = 1{,}273$). The 31.7% Option Bias indicates substantial format-dependent performance inflation.

MCQ confidence (90.7%), despite accuracy dropping from 87.8% to 56.2%. This paradoxical increase in confidence concurrent with a 31.7-percentage-point accuracy decline reveals that the model not only fails to recognize its degraded performance when MCQ scaffolding is removed, but actually becomes *more* confident (Figure 3).

Table 3: Calibration metrics: MCQ vs. Open-Ended format ($n = 1{,}273$).

Metric	**MCQ**	**Open-Ended**
ECE	0.029	0.364
SW-ECE	0.029	0.364
Brier Score	0.103	0.372
Mean Confidence	90.7%	92.5%
Mean Accuracy	87.8%	56.2%
Confidence–Accuracy Gap	2.9%	**36.3%**

4.3. Overconfident Incorrect Cases

We identified 147 cases (11.5%) where the model expressed $> 80\%$ confidence on an MCQ question it answered incorrectly. These "overconfident-wrong" cases represent the highest-risk scenario for clinical

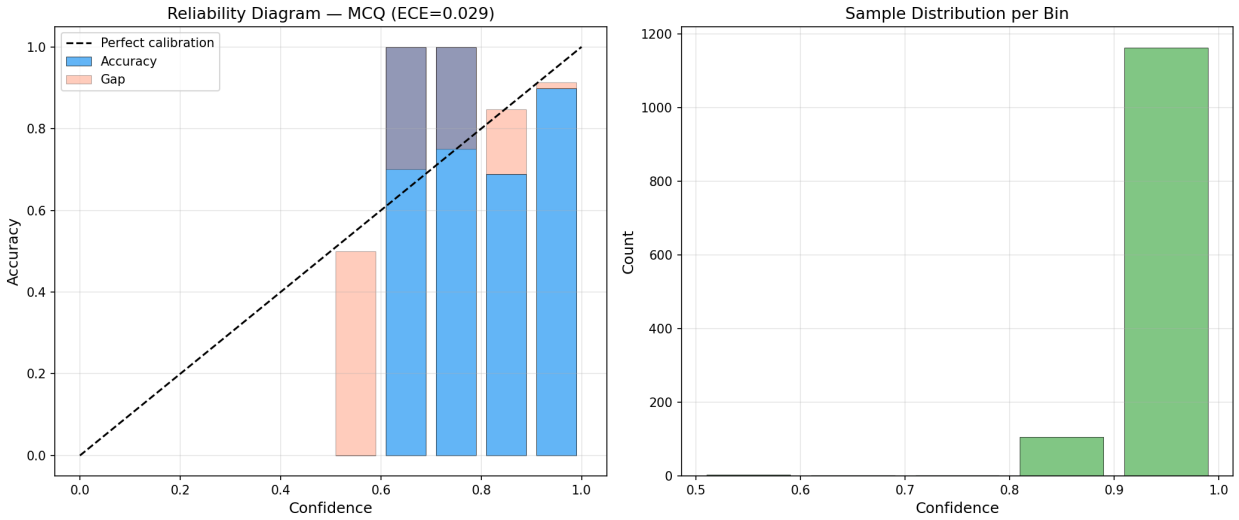

Figure 2: Reliability diagram for GPT-4o under MCQ format (ECE = 0.029, n = 1,273). The model shows good calibration with most predictions in the 85–95% confidence range and accuracy closely matching confidence.

deployment—the model is both wrong and sure of itself, providing no signal to the clinician that the recommendation may be unreliable. Notably, 147 of the 155 total MCQ errors (94.8%) occurred with $> 80\%$ confidence, confirming that GPT-4o rarely expresses low confidence even when incorrect.

4.4. Confidence Distribution

Figure 4 shows the confidence distributions for correct vs. incorrect answers in both formats. Under MCQ format, the model shows moderate confidence separation (correct answers tend toward higher confidence). Under open-ended format, the distributions overlap substantially, confirming that the model's confidence is a poor discriminator of correctness when MCQ scaffolding is absent.

5. Discussion

5.1. The MCQ Illusion: Implications for Clinical AI Evaluation

Our finding that GPT-4o's accuracy drops from 87.8% (MCQ) to 56.2% (open-ended) across all 1,273 MedQA questions represents a robust demonstration of format-dependent performance. The 36.0% Relative Option Bias indicates that *over one-third* of the model's apparent clinical reasoning under MCQ evaluation is attributable to the availability of candidate answers rather than genuine diagnostic competence.

This has direct implications for claims that "LLMs pass medical exams." If MCQ performance substantially overestimates true clinical reasoning ability, then USMLE passing scores achieved via MCQ benchmarks cannot be directly interpreted as evidence of clinical readiness. Real-world clinical queries do not come with four pre-specified answer options; physicians must generate diagnoses, treatments, and next steps from memory and reasoning.

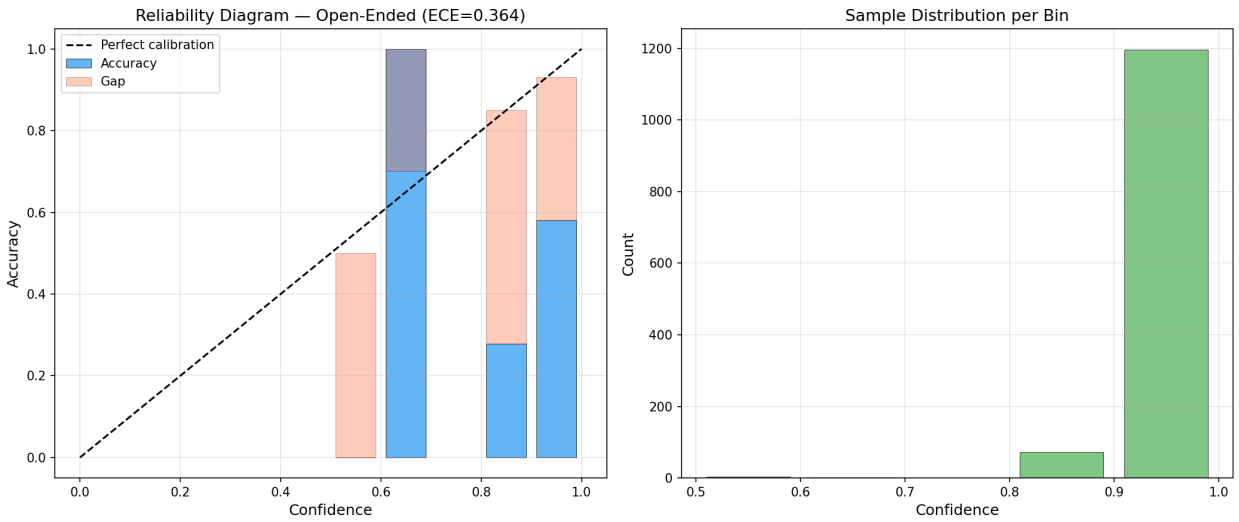

Figure 3: Reliability diagram for GPT-4o under Open-Ended format (ECE = 0.364, n = 1,273). Despite accuracy dropping to 56.2%, the model maintains 85–95% confidence, revealing severe overconfidence. The large gap between the bars and the perfect calibration line indicates dangerous miscalibration.

5.2. The Calibration Paradox: Worse Performance, Same Confidence

Perhaps our most concerning finding is the **calibration asymmetry** between formats. The model maintained nearly identical mean confidence (90.7% vs. 92.5%) across both formats despite a 31.7-percentage-point drop in accuracy. This means the model *does not recognize* when it has transitioned from a format where its confidence is warranted (MCQ) to one where it is dangerously misplaced (open-ended).

This asymmetry creates a specific clinical hazard: if a CDSS is evaluated using MCQ-based calibration (which appears acceptable at ECE = 0.029) but deployed in open-ended clinical scenarios, the deployed system's actual calibration (ECE = 0.364) is approximately *twelve times* worse than the evaluation suggested. This represents a direct patient safety concern.

5.3. Overconfidence and Patient Safety

Drawing on Yang et al.'s work on drug safety and CDSS at Taipei Medical University, we identify two distinct patient safety mechanisms through which LLM overconfidence causes harm:

Mechanism 1: Confident Error Propagation.. When an LLM generates a high-confidence incorrect answer (e.g., recommending the wrong drug with 90% stated confidence), clinicians—particularly those under time pressure, fatigue, or cognitive overload—may accept the recommendation. In pharmacology, this directly translates to medication errors: wrong drug, wrong dose, missed contraindications, or unrecognized interactions.

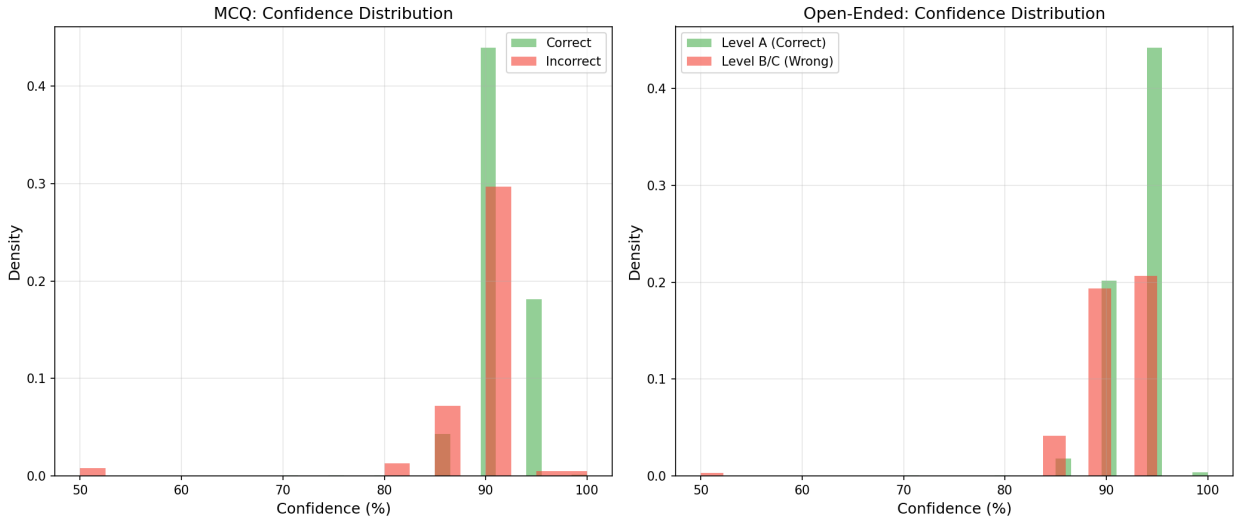

Figure 4: Confidence distributions for correct (green) vs. incorrect (red) answers. Left: MCQ format shows some confidence separation. Right: Open-ended format shows overlapping distributions, indicating that confidence cannot reliably distinguish correct from incorrect open-ended responses.

Mechanism 2: Alert Fatigue Induction.. Over time, repeated high-confidence incorrect alerts erode clinician trust in the AI system. Eventually, clinicians begin dismissing AI recommendations wholesale—including correct, clinically important ones. This "crying wolf" effect is well-documented in traditional CDSS (Alert-Fatigue, 2019) and would be exacerbated by poorly calibrated LLM-based systems.

Both mechanisms converge on patient harm. Critically, **Mechanism 2 is iatrogenic to the AI system itself**—the system's overconfidence progressively destroys its own utility.

5.4. Safety-Weighted ECE: Why Not All Miscalibration Is Equal

Standard ECE treats all miscalibration uniformly, but clinical reality demands domain-aware evaluation. Our proposed SW-ECE (Equation 5) operationalizes the clinical principle that drug-related errors (pharmacology, weight 3.0) carry greater patient safety implications than anatomical knowledge gaps (basic sciences, weight 1.0).

In the current evaluation with MedQA's topic categorization ("step1" for basic sciences, "step2&3" for clinical reasoning), SW-ECE equaled standard ECE due to similar topic distributions across confidence bins. We note that MedQA's coarse topic labels (only two categories) limit subdomain-specific analysis. Future work with MedMCQA's finer-grained 21-subject categorization will enable testing whether SW-ECE reveals that models are *disproportionately miscalibrated in safety-critical domains*—specifically, pharmacology questions that require precise drug name recall.

5.5. Toward a Three-Tier Clinical AI Screening Framework

Based on our findings, we propose a comprehensive screening framework for evaluating LLM readiness for clinical deployment:

Tier 1: Competence Screening (MCQ + Open-Ended Accuracy).. Does the model possess sufficient medical knowledge? Traditional MCQ benchmarks provide a necessary (but not sufficient) floor. Open-ended evaluation quantifies genuine clinical reasoning ability, and Option Bias measures format dependency.

Tier 2: Self-Awareness Screening (Calibration + Selective Prediction).. Does the model know what it doesn't know? ECE and SW-ECE measure calibration quality. Coverage@95%—the proportion of questions the model can answer while maintaining 95% accuracy—provides a direct deployment metric: "At 95% accuracy, this model can autonomously handle X% of clinical queries."

Tier 3: Robustness Screening (Multi-Model Cross-Supervision).. Do diverse models agree? Ensemble agreement among architecturally diverse models (e.g., GPT-4o, Claude, Gemini, DeepSeek) provides an orthogonal quality signal. Disagreement flags cases for human review, functioning as an additional safety net.

Only models passing all three tiers should proceed to clinical pilot testing. This framework directly addresses the regulatory need for comprehensive AI/ML medical device evaluation beyond simple accuracy metrics.

5.6. Implications for Multi-Model Clinical Supervision

The multi-model cross-supervision approach (Du et al., 2023; Cohen et al., 2023) offers a practical mitigation strategy for the calibration problems we identify. Rather than relying on a single model's potentially miscalibrated confidence, querying multiple diverse models and using agreement as a confidence proxy can provide better-calibrated uncertainty estimates. Our preliminary pilot testing (Section 3) with four cloud providers (OpenAI, Anthropic, Google, DeepSeek) demonstrated that while models reach consensus on clear-cut cases (e.g., ACE inhibitor contraindication in pregnancy), they diverge on nuanced therapeutic choices—precisely the cases requiring human oversight.

5.7. Limitations

Several limitations should be noted:

1. **Single model**: Only GPT-4o was evaluated. Future work will compare 8+ models spanning cloud and local deployments, enabling analysis of Option Bias vs. model architecture and scale.
2. **Automated judgment**: Open-ended answers were judged by GPT-4o (auto-evaluation), which may introduce systematic bias. Human expert validation with inter-rater reliability assessment (κ) is needed for a subset of responses.

3. **Single dataset**: Only MedQA was used. Generalizability across MedMCQA (21 medical subjects), MMLU-Med, and PubMedQA should be assessed.
4. **Coarse topic categorization**: MedQA provides only two topic labels ("step1" and "step2&3"), limiting subdomain-specific calibration analysis. Finer-grained datasets are needed to assess domain-specific miscalibration patterns.
5. **Simplified semantic matching**: Clinical judgment uses GPT-4o rather than SNOMED CT ontological matching. Formal semantic distance analysis may reveal additional nuance in the Level B (partial correctness) category.

6. Conclusion

We demonstrate that MCQ-based evaluation of LLMs for clinical reasoning provides a significantly inflated assessment of true clinical competence. Across the complete MedQA test set ($n = 1{,}273$), GPT-4o's 87.8% MCQ accuracy masks a 56.2% open-ended accuracy, with 36% of MCQ performance attributable to the option scaffolding. More critically, the model exhibits severe calibration degradation (ECE from 0.029 to 0.364) under open-ended conditions while maintaining high confidence (92.5%)—a direct hazard for CDSS deployment. The finding that 94.8% of MCQ errors occur with $> 80\%$ confidence underscores that current models lack the metacognitive awareness required for safe clinical deployment.

These findings argue for (1) mandatory open-ended evaluation alongside MCQ benchmarks for clinical AI, (2) format-specific calibration assessment before deployment, and (3) adoption of safety-weighted calibration metrics (SW-ECE) that prioritize miscalibration in high-stakes clinical domains. Our proposed three-tier screening framework—integrating competence, self-awareness, and robustness assessment—provides a practical roadmap for trustworthy clinical AI evaluation.

Future work will extend this analysis to 8+ models of varying scale, multiple datasets (MedMCQA, MMLU-Med), finer-grained subdomain analysis (pharmacology vs. basic sciences), SNOMED CT semantic matching, and multi-model ensemble calibration.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgments

This research was supported by the National Science and Technology Council (NSTC), Taiwan. We thank the Graduate Institute of Data Science in Healthcare at Taipei Medical University for computational resources.

References

Ancker, J. S., et al., 2019. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Medical Informatics and Decision Making* 17, 36.

Cohen, R., et al., 2023. LM vs LM: Detecting factual errors via cross-examination. In: *EMNLP 2023*.

Du, Y., et al., 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv:2305.14325*.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K. Q., 2017. On calibration of modern neural networks. In: *ICML 2017*.

Hendrycks, D., et al., 2021. Measuring massive multitask language understanding. In: *ICLR 2021*.

Jin, Q., et al., 2021. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences* 11 (14), 6421.

Kadavath, S., et al., 2022. Language models (mostly) know what they know. *arXiv:2207.05221*.

Labrak, Y., et al., 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. *arXiv:2402.10373*.

Nori, H., et al., 2023. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv:2311.16452*.

Pal, A., et al., 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: *CHIL 2022*.

Singhal, K., et al., 2023. Large language models encode clinical knowledge. *Nature* 620, 172–180.

Tian, K., et al., 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models with optimized prompting. *arXiv:2305.14975*.

Wang, X., et al., 2023. Self-consistency improves chain of thought reasoning in language models. In: *ICLR 2023*.