Epistemic Reliability of Frontier Multimodal LLMs in Subspecialist Ophthalmology: A Factorial Evaluation of Image-Reliant Reasoning

Jaehyeong Kim[2] KIMJAEHYEONG1209@NAVER.COM
Semin Yang[2] SEMINI66@NAVER.COM
Shina Jang[1,3] SAJANG@INHA.AC.KR
Joseph Kim[1,2,*] JKIM0529@NAVER.COM

[1] *Inha University College of Medicine, Incheon, Republic of Korea*
[2] *Department of Ophthalmology, Inha University Hospital, Incheon, Republic of Korea*
[3] *Department of Obstetrics and Gynecology, Inha University Hospital, Incheon, Republic of Korea*
[*] *Corresponding author*

Editors: Under Review for MIDL 2026

Abstract

Existing evaluations of large language models (LLMs) in ophthalmology measure accuracy under idealized, single-condition settings, leaving uncharacterized three epistemic properties—confidence calibration, output self-consistency, and cross-model failure convergence—directly relevant to clinical reliability. We evaluated four frontier multimodal LLMs (GPT-5.4, Claude Opus 4.6, Gemini 2.5 Pro, Gemini 3.1 Pro Preview) on 20 real-world subspecialist-level retinal and uveitic cases using a fully factorial 2×2×3 design (case language × prompt language × input modality) with five repeated runs (4,800 total responses). Overall accuracy was 89.2%–90.3% with no significant between-model differences ($P \geq .109$), substantially outperforming senior residents (50.0%; $P \leq .008$). Diagnostic accuracy was near-ceiling (97.5%–100.0%) while management accuracy was substantially lower (63.9%–72.8%; $P < .001$). Three systematic epistemic patterns emerged invisible to accuracy alone: systematic overconfidence amplified in management tasks (+14.4 to +32.1 pp); a verbosity paradox in which incorrect responses were significantly more elaborate ($P \leq .007$, three of four models); and cross-model convergence on identical errors. These findings reframe LLM evaluation in medical imaging toward epistemic reliability.

Keywords: large language models, ophthalmology, epistemic reliability, confidence calibration

1. Introduction

Despite rapid LLM advances—from GPT-4 achieving 73% on ophthalmic board examinations (Wei et al., 2025) to frontier models exceeding 80% on clinical vignettes—a systematic review of 187 studies found extreme heterogeneity ($I^2 = 94.5\%$) and identified three persistent gaps (Zhang et al., 2025b): **Experimental factors:** Modality, case language, and prompt language have been treated as fixed rather than experimental factors; **Clinical management:** Clinical management reasoning under diagnostic uncertainty is almost unevaluated; **Epistemic reliability:** Epistemic properties of LLM outputs—confidence calibration, self-consistency, and cross-model error convergence—have not been systematically examined, despite their direct relevance to safe clinical use (Hager et al., 2024; Jin

et al., 2024). We address all three gaps with a single fully factorial study, producing the first systematic characterization of *epistemic reliability* across frontier multimodal LLMs in ophthalmology.

2. Methods

Dataset. Twenty cases from the Korean Retina Society and Korean Uveitis Society publications (2020–2025; closed-access), comprising 14 diagnostic and 6 management questions with multimodal imaging (OCT, fundus photography, FA, ICGA).

Models. GPT-5.4, Claude Opus 4.6, Gemini 2.5 Pro, Gemini 3.1 Pro Preview, accessed via official APIs (March 2026).

Factorial design. 2 (case language: Korean/English) × 2 (prompt language: Korean/English) × 3 (modality: text-only [T1] / image-reliant [T2] / combined [T3]) with 5 runs per condition = 4,800 responses. Structured JSON outputs captured answer, reasoning text, and self-reported confidence (0–100%).

Human reference. Three senior ophthalmology residents completed all 20 cases under closed-book conditions using Korean-language T2 inputs.

Statistics. Accuracy: exact McNemar tests; modality: Cochran Q + Holm-corrected post hoc; calibration gap = mean confidence − accuracy (pp); verbosity: Mann-Whitney U; Spearman ρ for consistency–accuracy.

3. Results

Overall accuracy. Response-level accuracy was 90.3% (Claude Opus 4.6), 89.4% (GPT-5.4), 89.3% (Gemini 2.5), 89.2% (Gemini 3.1), with no significant between-model differences (all $P \geq .109$). All four models significantly outperformed the resident majority vote (50.0%; $P \leq .008$).

Diagnostic vs. management gap. Diagnostic accuracy was near-ceiling (97.5%–100.0%) while management accuracy was substantially lower (63.9%–72.8%; all $P < .001$), a gap of 25.1–36.1 pp consistent with difficulty recognizing when prior workup requires escalation (McCoy et al., 2025; Bilalić et al., 2025). T2 was the most challenging modality (84.5%–88.0%) for all models. Prompt language had no significant effect (all $P \geq .608$), supporting Korean–English equivalence. Condition-level heatmaps are shown in Figure 1.

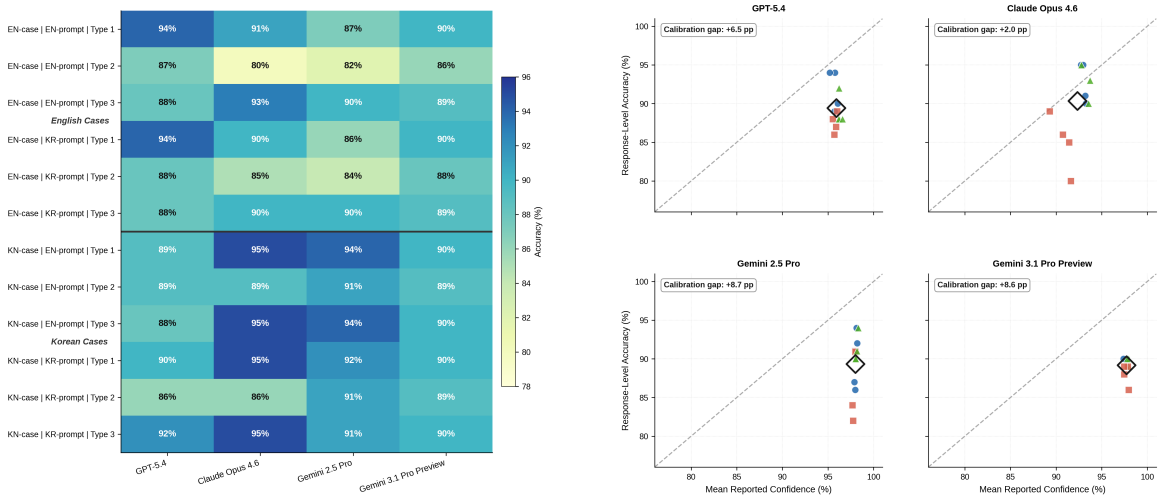
Confidence calibration. All models were systematically overconfident. Claude Opus 4.6 was best calibrated (gap +2.0 pp; correct–incorrect differential 8.9 pp). Both Gemini models expressed ≥90% confidence on 100% of incorrect responses, with differentials of only 1.2–1.4 pp—establishing that confidence cannot serve as an error-detection signal for these architectures (Berner and Graber, 2008). Calibration gaps were further amplified in management tasks (+14.4 to +32.1 pp; Table 1), precisely where accuracy was lowest. Calibration profiles are visualized in Figure 1.

Verbosity paradox. Incorrect responses were significantly longer than correct responses for GPT-5.4 (844 vs. 778 characters, $P = .001$), Claude Opus 4.6 (1,560 vs. 1,446, $P = .007$), and Gemini 2.5 Pro (1,061 vs. 921, $P < .001$). This verbosity paradox (Zhang et al., 2025a)—reasoning elaborateness inversely tracking correctness—implies that the apparent sophistication of LLM outputs cannot serve as a proxy for validity.

Cross-model error convergence. Four questions elicited identical incorrect responses from all four models whenever they erred; one management case produced zero correct responses across all 60 runs at mean confidence 98.9%. This pattern (Peng and Garg, 2025) implies that consensus-based quality filters fail when shared biases drive convergence.

Table 1: Confidence calibration. pp = percentage points.

Model	Overall	Mgmt	Corr–Incorr	5/5
Claude Opus 4.6	+2.0 pp	+14.4 pp	8.9 pp	97.5%
GPT-5.4	+6.5 pp	+22.7 pp	2.5 pp	93.3%
Gemini 2.5 Pro	+8.7 pp	+28.7 pp	1.4 pp	90.4%
Gemini 3.1 Pro Prev.	+8.6 pp	+32.1 pp	1.2 pp	97.1%



(a) Condition-level accuracy heatmap (48 cells). (b) Confidence calibration profiles.

Figure 1: (a) Accuracy across all 48 experimental cells; T2 conditions rank consistently lower. (b) Calibration gaps by model and condition; Gemini models show clinically concerning near-uniform overconfidence.

4. Discussion and Conclusions

Frontier multimodal LLMs demonstrated subspecialist-level accuracy across all modality and language conditions, with Korean–English prompt-language equivalence confirming accessibility for non-Anglophone settings. However, three systematic epistemic patterns—overconfidence amplified in management tasks, verbosity paradox, and cross-model error convergence—are not captured by accuracy benchmarking and must be addressed for safe clinical use. These properties define an epistemic reliability framework that should supplement accuracy in any rigorous evaluation of LLMs for medical imaging.

References

Eta S. Berner and Mark L. Graber. Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, 121(5):S2–S23, 2008. doi: 10.1016/j.amjmed.2008.01.001.

Merim Bilalić, Peter McLeod, and Fernand Gobet. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Scientific Reports*, 15:22940, 2025. doi: 10.1038/s41598-025-XXXXX-X.

Paul Hager, Fabian Jungmann, Kunal Bhagat, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30: 2613–2622, 2024. doi: 10.1038/s41591-024-03097-1.

Qiao Jin, Fangyuan Chen, Yiliang Zhou, et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *npj Digital Medicine*, 7(1):190, 2024. doi: 10.1038/s41746-024-01185-7.

Liam G. McCoy, Felix Bucher, Adam Rodman, et al. Assessment of large language models in clinical reasoning: a novel benchmarking study. *NEJM AI*, 2(10), 2025. doi: 10.1056/AIoa2400570.

Ken Peng and Nikhil Garg. Correlated errors in large language models, 2025.

Jiahui Wei, Xin Wang, Ming Huang, et al. Evaluating the performance of ChatGPT on board-style examination questions in ophthalmology: a meta-analysis. *Journal of Medical Systems*, 49(1):94, 2025. doi: 10.1007/s10916-025-02134-3.

Yue Zhang, Subhabrata S. Das, and Rui Zhang. Demystify verbosity compensation behavior of large language models. In *Proceedings of the UncertaiNLP Workshop*, pages 160–178, 2025a.

Zhiyuan Zhang, Hang Zhang, Zengxin Pan, et al. Evaluating large language models in ophthalmology: systematic review. *Journal of Medical Internet Research*, 27:e76947, 2025b. doi: 10.2196/76947.