• Intelligent Ophthalmology •

An evaluation of Gemini, ChatGPT, and DeepSeek large language model responses to ophthalmology case-based questions

Amir–Ali Golrokhian–Sani[1], *Jennifer L. Patnaik*[2,3], *Vishaal Bhambhwani*[4,5]

[1]University of Ottawa, Faculty of Medicine, Ottawa, ON K1H 8M5, Canada
[2]University of Colorado Anschutz Medical Campus, Aurora, CO 80045, United States
[3]Orbis International, New York, NY 10017, United States
[4]Northern Ontario School of Medicine University, Thunder Bay, ON P7B 7A5, Canada
[5]Thunder Bay Regional Health Sciences Centre, Thunder Bay, ON P7B 7A5, Canada
Correspondence to: Vishaal Bhambhwani. 1620 Mountain Road, Thunder Bay, Ontario P7J 1C8, Canada. vishaalb@ymail.com
Received: 2026-03-13 Accepted: 2026-09-18

Abstract

• **AIM:** To evaluate the accuracy and appropriateness of responses of different large language models (LLMs) to ophthalmology case-based questions.

• **METHODS:** Prospective, observational, cross-sectional study conducted in 2025 in Ontario, Canada. Three LLMs (Gemini Advanced 2.5 Pro, ChatGPT 4.5, and DeepSeek DeepThink R1) were asked to respond to all "Clinical Challenge" cases published in *The Journal of the American Medical Association Ophthalmology* from 2021 to 2024 (135 multiple-choice questions). They were evaluated for accuracy and response appropriateness. Chi-square and Kruskal-Wallis tests were used to compare their performances.

• **RESULTS:** The average accuracy across all LLMs was 303/405 (74.8%), with Gemini leading at 106/135 (78.5%), followed by ChatGPT (99/135; 73.3%) and DeepSeek (98/135; 72.6%). The pairwise differences among the LLMs were not statistically significant. The average appropriateness score of all LLMs together was 4.44 out of 6. Gemini, ChatGPT, and DeepSeek had average appropriateness scores of 4.6, 4.2, and 4.5, respectively (P=0.007).

• **CONCLUSION:** These LLMs reveal a moderate-to-high level of accuracy and appropriateness of reasoning when answering complex ophthalmology case-based questions. While ChatGPT and DeepSeek differ by one correct answer, Gemini outperforms them by over 5%. When using LLMs for clinical decision-making, the model's ability to provide quality explanations for the selected course of action is an asset. Although Gemini is the overall leader in appropriateness of reasoning, the fact that DeepSeek outperforms ChatGPT in appropriateness, combined with its open-source, free-to-use, and offline-capable features, makes it an exciting prospect for future use in ophthalmology.

• **KEYWORDS:** artificial intelligence; large language model; ophthalmology; cases; performance
DOI:10.18240/ijo.2026.12.02

Citation: Golrokhian-Sani AA, Patnaik JL, Bhambhwani V. An evaluation of Gemini, ChatGPT, and DeepSeek large language model responses to ophthalmology case-based questions. *Int J Ophthalmol* 2026;19(12):2321-2326

INTRODUCTION

Large language models (LLMs) are artificial intelligence (AI) systems trained on vast quantities of text from articles, books, and websites[1-3]. Developed from advancements in natural language processing, these models use deep learning techniques, especially transformer-based architectures. Their training allows them to learn patterns of word use in a language and predict the following words in a sequence. Ultimately, this enables these LLMs to perform functions in various language-based scenarios. Training these models can be complicated, requiring varying levels of human supervision, resulting in different end products; however, they all share a common focus on language processing.

Since their rise in popularity, LLMs have been explored as valuable tools in ophthalmology[4-12]. These models have continued to improve as they learn, accomplishing various tasks involving complex problem-solving[13]. Two of the most capable models reported consistently in the literature include ChatGPT by OpenAI (OpenAI, Inc., California, USA) and Gemini by Google (Google LLC, California, USA). While both

Table 1 Appropriateness rating scale

Appropriateness	Description
0: Completely inappropriate	The rationale provided is not in line with the discussion The provided response is misleading and/or potentially harmful
1: Mostly inappropriate	The rationale provided is not in line with the discussion The provided response requires many corrections and is not useful
2: Somewhat inappropriate	The rationale provided deviates partially from the discussion While not entirely wrong, the response requires improvement
3: Neither appropriate nor inappropriate	The rationale provided is borderline in similarity to the discussion or lacks sufficient detail to be considered useful
4: Somewhat appropriate	The rationale provided is somewhat the same as the discussion It has moderate deviations in formation. Overall, a limited but not wholly inaccurate response
5: Mostly appropriate	The rationale provided is mostly the same as the discussion It has minor deviations or slight gaps in formation. Overall, a useful response
6: Completely appropriate	The rationale provided is the same as the discussion The provided response is accurate, clear, and very helpful for the specific question

are believed to be transformer-based at their core, ChatGPT focuses more on generating conversational prose, whereas Gemini is highly adaptable to various contexts, making it a capable problem solver[14-17]. These models have been shown to be useful in various areas of ophthalmology, including patient education, diagnosis, triage, and surgical planning[18-20].

In January 2025, the LLM landscape changed with the introduction of a new model from DeepSeek (Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd., Zhejiang Province, China)[21]. The first standout difference between this and the existing models was its open-source nature. This means the model can be customized and even used offline, a major advantage for healthcare applications where privacy is paramount. As a relatively new entrant to the LLM space, its capabilities have not been as thoroughly investigated as those of other models.

Our study compares the responses of the three LLMs, Gemini, ChatGPT, and DeepSeek, to ophthalmology case-based questions and analyzes their decision-making processes. To the best of our knowledge, a comparison of these three models has not been published previously.

MATERIALS AND METHODS

Ethical Approval Formal institutional review board approval was not required, as the study involved no confidential patient information and relied exclusively on publicly available data.

In this prospective, observational, cross-sectional study, the three AI models (Gemini, ChatGPT, and DeepSeek) were tasked with evaluating all "Clinical Challenge" cases published by *The Journal of the American Medical Association* (*JAMA*) *Ophthalmology* from January 1, 2021, to December 31, 2024. This study used Gemini Advanced 2.5 Pro, ChatGPT 4.5, and DeepSeek DeepThink R1.

A total of 135 case-based scenarios were presented to the LLMs from April to July 2025. All text aspects of these cases (*i.e.*, case descriptions, figure captions, and multiple-choice options) were provided to the models. As *JAMA Ophthalmology*'s Clinical Challenges adhered to a standard template, finishing with a "What Would You Do Next" statement before the multiple-choice options, minimal prompt engineering was required to keep the models on track. The prompt selected was "Without searching the web, answer the following case-based question. You must choose only one of the four multiple-choice options." The cases, along with the prompt, were pasted identically in all models. The accounts used for each model were new and had not been used for any other purpose. Each case was pasted into a new discussion when it was asked of the models. The models were not asked any follow-up questions.

The responses of the three LLMs were compared to the correct answers provided by the *JAMA Ophthalmology* Clinical Challenges key. The complete answers of all models to all questions were saved for further analysis of reasoning and possible decision-making processes. A correct answer by the models was defined as explicit mention of the correct multiple-choice option or a clear reference to the correct answer with rephrased wording. Chi-squared tests were used to compare the overall total accuracy (correct answers) across the three models and for pairwise comparisons. The LLM responses were rated on a seven-point Likert scale for "appropriateness" adapted from Balas *et al*[22] (Table 1). Using the discussion section of the *JAMA Ophthalmology* clinical cases as the reference standard for the ideal answer, one of the authors, who was blind to the particular LLM used, rated the appropriateness of the responses from the three models to all questions. The models were then compared using the Kruskal-Wallis test for overall differences and the post hoc Mann-Whitney *U* test for pairwise differences between the models.

All analyses were performed using SPSS 29.0.2.0 and Google Sheets 2025. *P*-values of <0.05 were considered significant.

RESULTS

The prompt structure achieved a 100% success rate in getting all LLMs to respond to every question on the first attempt,

without selecting more than one answer option or choosing an answer not mentioned in the list of options.

The accuracy results (number of correct answers) for Gemini Advanced 2.5 Pro, ChatGPT 4.5, and DeepSeek DeepThink R1 were 106/135 (78.5%), 99/135 (73.3%), and 98/135 (72.6%), respectively. Chi-squared analysis showed that the pairwise differences among the LLMs were not statistically significant, even though Gemini outperformed the other two models by over 5% [ChatGPT versus DeepSeek Pearson Chi-square: 0.891, odds ratio (OR): 1.04, 95% confidence interval (CI): 0.61–1.78; ChatGPT versus Gemini Pearson Chi-square: 0.319, OR: 1.33, 95%CI: 0.76–2.33; DeepSeek versus Gemini Pearson Chi-square: 0.257, OR: 1.38, 95%CI: 0.79–2.41]. The average accuracy across all LLMs was 303/405 (74.8%).

The mean Likert scale ratings on the appropriateness of responses for the LLMs were 4.6 [standard deviation (SD): 1.2], 4.2 (SD: 1.3), 4.5 (SD: 1.2), and for Gemini Advanced 2.5 Pro, ChatGPT 4.5, and DeepSeek DeepThink R1, respectively (median scores of 5, 4, and 5, respectively). These ratings were significantly different across all three LLMs (P=0.007), with Gemini leading and ChatGPT trailing. Mann-Whitney U tests were performed post hoc to identify pairwise differences among the three models, revealing significant differences between the ChatGPT-DeepSeek and ChatGPT-Gemini pairs. ChatGPT was shown to be significantly weaker than the other two models in terms of answer appropriateness [ChatGPT versus DeepSeek test statistic (U): 7792, Z score: -2.13, P=0.04; ChatGPT versus Gemini test statistic (U): 7234, Z score: -3.03, P=0.003]. Additionally, Gemini and DeepSeek were not significantly different from each other [DeepSeek versus Gemini test statistic (U): 8512.5, Z score: -0.97, P=0.35). The overall average appropriateness score of all LLMs together was 4.44 out of 6.

The distribution of appropriateness scores for the models was visualized using a Violin plot and a bar graph for comparison (Figures 1 and 2). Gemini had the most answers rated 5 or 6, and ChatGPT had the most answers rated 1 (Figure 2).

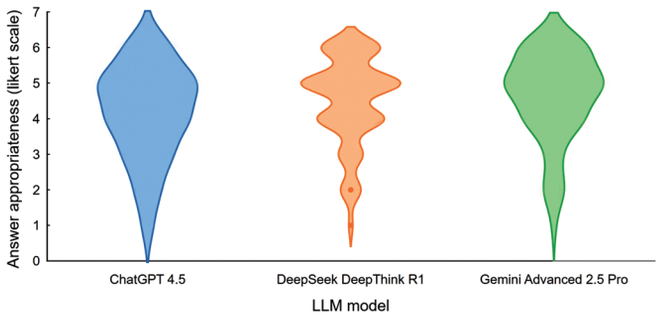


Figure 1 Violin plot of AI appropriateness scores: the AI models are graphed to demonstrate the distribution of their appropriateness scores ChatGPT 4.5 is shown to have a blunted peak at the 4–5 range; DeepSeek DeepThink R1 had multiple peaks, with the largest being at 5, flanked by a peak at both 4 and 6; and Gemini Advanced 2.5 Pro had a moderately pointed peak at the 5 mark. LLM: Large language model; AI: Artificial intelligence.

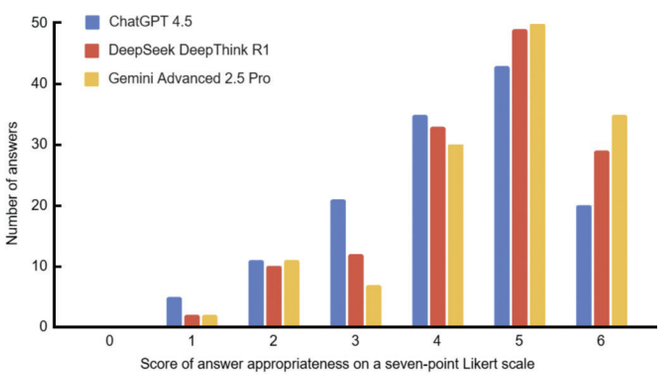


Figure 2 Bar graph of AI appropriateness scores: the AI models are graphed to demonstrate the distribution of their appropriateness scores ChatGPT 4.5 is shown to have the most answers with the scores 1, 3, and 4; DeepSeek DeepThink R1 did not lead at any score, but it was distributed mainly in the 4–6 range; and Gemini Advanced 2.5 Pro had the most answers with the scores 5 and 6. No answers were scored as 0.

DISCUSSION

Our study shows that LLMs, in general, have a moderate to high level of accuracy and appropriateness of reasoning when answering ophthalmology case-based questions, with significant differences in these parameters among the model versions of the LLMs compared.

The Gemini Advanced 2.5 Pro, ChatGPT 4.5, and DeepSeek DeepThink R1 versions of the models were selected for our study because they were the most advanced available in their category at the time of the study. Notably, while the Gemini and ChatGPT models were paid (subscription-based), DeepSeek's DeepThink R1 model was available for free.

The *JAMA Ophthalmology* "Clinical Challenge" cases were selected for multiple reasons. First, being a reputable source, we were confident of the accuracy of the questions and the answers. Second, the questions, answers, and discussions were readily available online. Third, in general, these cases are considered quite challenging by the ophthalmology community, requiring significant ophthalmology knowledge, analysis, and reasoning skills, which we aimed to test in the LLMs. Fourth, as these cases adhered to a standard template, finishing with a "What Would You Do Next" statement before the multiple-choice options, minimal prompt engineering was required to keep the models on track.

The prompt structure presented to the LLMs was chosen for three main reasons. First, it did not restrict the LLMs so heavily as to interfere with their answer presentation. They were not instructed to approach the problem in a step-by-step manner

or even to explain their thinking, leaving them free to answer the questions "naturally". Second, it asked for a single answer to a case, making it clear that these were not real healthcare decisions being made. This was done to prevent situations in which the model might refuse to answer a case because it is not a healthcare professional, as previously observed[23]. Third, it requested that the models not search the web, as Gemini's web search option cannot be disabled. Only text aspects of the cases were presented to the models, as DeepSeek does not accept images, and we did not want the other models to have an unfair advantage.

We found Gemini to provide the greatest number of accurate answers, while ChatGPT and DeepSeek differed only by a single correct answer [Gemini 106/135 (78.5%), ChatGPT 99/135 (73.3%), and DeepSeek 98/135 (72.6%)]. Thus, Gemini outperformed the other two models by over 5%. As ChatGPT and DeepSeek were fairly close in accuracy, this may indicate a similarity in their training materials or architectures, as some reports have alleged[24-25].

Regarding the appropriateness of reasoning, the LLMs had an overall average appropriateness score of 4.44 out of 6. This means that the average rating was between the "somewhat appropriate" and "mostly appropriate" categories. The graph of these appropriateness ratings (Figure 1) further elucidates these findings with a peak at a rating of 5 (*i.e.*, "mostly appropriate"). In our opinion, the performance of these LLMs in this regard was quite impressive, considering that the *JAMA Ophthalmology* clinical cases require significant ophthalmology knowledge, as well as analytical and reasoning skills. Comparing the appropriateness of responses among the three LLMs, Gemini once again led with an average rating 0.12 points higher than DeepSeek, which came in second. A key difference was that ChatGPT heavily trailed the group in appropriateness ratings. The findings of the Kruskal-Wallis and Mann-Whitney *U* tests also showed significant differences in the appropriateness scores of the ChatGPT model compared with the two other models. This differs from the findings on accuracy, where the models were similar.

The difference between the accuracy rating and the appropriateness scores among the models is interesting because the former only requires the model to select the correct letter option, while the latter assesses explanations for the chosen answer. Here, a potential weakness of ChatGPT was identified. The output from ChatGPT fell short when it came to explaining its reasoning correctly and thoroughly. When using AI as a tool in clinical decision-making, its ability to provide quality explanations for the selected course of action would be an asset. Thus, there would be value in further research aimed at analyzing why ChatGPT was unable to perform to the same standards as its contemporary models in this scenario. This was an unusual finding, as ChatGPT has been frequently shown to outperform many other LLMs in the published literature[18,26-30]. One potential explanation is the architecture of the models themselves: while Gemini 2.5 Pro and DeepSeek R1 are reasoning models, meaning they are trained to follow an internal "chain-of-thought" framework that explores multiple options and self-corrects before generating a final output, ChatGPT-4.5 is a standard non-reasoning model, which relies on direct token prediction based on pre-trained statistical associations. This fundamental structural difference likely contributed to its weakness in response appropriateness[31].

Our study has some limitations. First, a single author scored appropriateness. Although the appropriateness score was based on the official solution and discussion of the *JAMA Ophthalmology* Clinical Cases as the reference standard for the ideal answer, and the assessing author was blind to the LLM used, more independent scorers may be desirable. Second, as mentioned earlier, only text aspects of the cases were presented to the models to allow a fair comparison, as DeepSeek does not accept images. Future studies with images would be desirable, once all models accept image input. Third, the study analyses LLM responses to actual complex clinical cases and these actual patients cannot always be put into a single ophthalmology subspecialty categorization. The aim of the study was to examine the role LLMs in ophthalmology as a whole, and not a particular subspecialty of ophthalmology. Future studies with larger sample sizes may be more useful for this purpose.

This paper also demonstrates a strength of DeepSeek over ChatGPT: both models are similar in accuracy but differ in answer appropriateness, with DeepSeek outperforming. While Gemini was the overall leader in both accuracy and appropriateness of reasoning, the fact that DeepSeek performed better than ChatGPT for appropriateness, combined with its open-source nature, free-to-use features, and offline capabilities, makes it an exciting prospect for future use in ophthalmology. At present, however, based on our study results, Gemini Advanced 2.5 Pro would be the LLM of choice for responding to ophthalmology case-based scenarios. It is important to state; however, that while LLMs may serve as a useful tool to supplement medical knowledge, organize thoughts, and explore possible differentials, they should not replace expert clinical judgement or peer consultation.

In conclusion, our study demonstrates the ability of multiple LLMs to answer correctly and reason appropriately when presented with complex ophthalmology cases. The three models were shown to generally behave similarly, with Gemini's Advanced 2.5 Pro leading in terms of accuracy and the appropriateness of its clinical reasoning. DeepSeek was especially interesting as it placed itself close to Gemini

in terms of answer appropriateness. This, combined with its open-source nature, free-to-use feature, and offline capabilities, makes it a deserving candidate for study regarding future use in healthcare settings. Further studies with larger sample sizes and more diverse question sources are desirable to better differentiate the advantages and disadvantages of these different LLMs in the future.

ACKNOWLEDGEMENTS

Thank you to Professor Livio Di Matteo at Lakehead University, Thunder Bay, Ontario, for his pro-bono guidance on statistics.

Authors' Contributions: Golrokhian-Sani AA: Methodology, formal analysis, investigation, data curation, writing–original draft, writing–review & editing, visualization. Patnaik JL: Formal analysis, writing–review & editing. Bhambhwani V: Conceptualization, methodology, writing–original draft, writing–review & editing, supervision.

Data Availability Statement: The full data for this paper is not available as it involves information proprietary to *JAMA Ophthalmology*. Any request for unprocessed numerical data can be submitted to the corresponding author. Information will be stored for 5y.

AI-Generated Content Disclosure: No AI assistance was used in the writing of this manuscript.

Conflicts of Interest: Golrokhian-Sani AA, None; **Patnaik JL**, None; **Bhambhwani V**, Northern Ontario Academic Medicine Association (NOAMA), Orbis Canada, Eyenuk Inc., Icare USA Inc., Bayer, Roche, Biocon Canada, Canadian Ophthalmological Society Foundation.

REFERENCES

1 Thirunavukarasu AJ, Ting DSJ, Elangovan K, *et al*. Large language models in medicine. *Nat Med* 2023;29(8):1930-1940.

2 Khullar D, Wang X, Wang F. Large language models in health care: charting a path toward accurate, explainable, and secure AI. *J Gen Intern Med* 2024;39(7):1239-1241.

3 Denny P, Prather J, Becker BA, *et al*. Computing education in the era of generative AI. *Communications of the ACM* 2024;67(2):56-67.

4 Betzler BK, Chen HC, Cheng CY, *et al*. Large language models and their impact in ophthalmology. *Lancet Digit Health* 2023;5(12):e917-e924.

5 Wong M, Lim ZW, Pushpanathan K, *et al*. Review of emerging trends and projection of future developments in large language models research in ophthalmology. *Br J Ophthalmol* 2024;108(10):1362-1370.

6 Zhang Z, Zhang H, Pan Z, *et al*. Evaluating large language models in ophthalmology: systematic review. *J Med Internet Res* 2025;27:e76947.

7 Artiaga JCM, Guevarra MCB, Sosuan GMN, *et al*. Large language models in ophthalmology: a scoping review on their utility for clinicians, researchers, patients, and educators. *Eye (Lond)* 2025;39(15):2752-2761.

8 Kedia N, Sanjeev S, Ong J, *et al*. ChatGPT and Beyond: an overview of the growing field of large language models and their use in ophthalmology. *Eye* (*Lond*) 2024;38(7):1252-1261.

9 Agnihotri AP, Nagel ID, Artiaga JCM, *et al*. Large language models in ophthalmology: a review of publications from top ophthalmology journals. *Ophthalmol Sci* 2025;5(3):100681.

10 Wei MY, Li YL, Liu SY, *et al*. Evaluating the competence of large language models in ophthalmology clinical practice: a multi-scenario quantitative study. *Front Cell Dev Biol* 2025;13:1704762.

11 Sorin V, Kapelushnik N, Hecht I, *et al*. Integrated visual and text-based analysis of ophthalmology clinical cases using a large language model. *Sci Rep* 2025;15(1):4999.

12 Tailor PD, D'Souza HS, Li H, *et al*. Vision of the future: large language models in ophthalmology. *Curr Opin Ophthalmol* 2024;35(5):391-402.

13 Rao A, Pang M, Kim J, *et al*. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *J Med Internet Res* 2023;25:e48659.

14 Rane N, Choudhary S, Rane J. Gemini versus ChatGPT: applications, performance, architecture, capabilities, and implementation. *J Appl Artif Intell* 2024;5(1):69-93.

15 Alsajri A, Salman HA, Steiti A. Generative models in natural language processing: a comparative study of ChatGPT and Gemini. *Babylon J Artif Intell* 2024;2024:134-145.

16 Rane N, Choudhary S, Rane J. Gemini or ChatGPT Capability, performance, and selection of cutting-edge generative artificial intelligence (AI) in business management. *Stud Econ Bus Relat* 2024;5(1):40-50.

17 Kawinkoonlasate P. A comparative study of Google Gemini and ChatGPT in enhancing English language learning for EFL learners: a case study of the English research writing course. *Pedagogical Res* 2025;10(4):em0251.

18 Sabaner MC, Anguita R, Antaki F, *et al*. Opportunities and challenges of chatbots in ophthalmology: a narrative review. *J Pers Med* 2024;14(12):1165.

19 Huang AS, Hirabayashi K, Barna L, *et al*. Assessment of a large language model's responses to questions and cases about glaucoma and retina management. *JAMA Ophthalmol* 2024;142(4):371-375.

20 Biswas S, Davies LN, Sheppard AL, *et al*. Utility of artificial intelligence-based large language models in ophthalmic care. *Ophthalmic Physiol Opt* 2024;44(3):641-671.

21 Temsah A, Alhasan K, Altamimi I, *et al*. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. *Cureus* 2025;17(2):e79221.

22 Balas M, Janic A, Daigle P, *et al*. Evaluating ChatGPT on orbital and oculofacial disorders: accuracy and readability insights. *Ophthalmic Plast Reconstr Surg* 2024;40(2):217-222.

23 Gill GS, Tsai J, Moxam J, *et al*. Comparison of Gemini advanced and ChatGPT 4.0's performances on the ophthalmology resident ophthalmic knowledge assessment program (OKAP) examination review question banks. *Cureus* 2024;16(9):e69612.

24 Bass D, Ghaffary S. Microsoft Probing If DeepSeek-Linked Group Improperly Obtained OpenAI Data. *Bloomberg*. 2025. https://www.

bloomberg.com/news/articles/2025-01-29/microsoft-probing-if-deepseek-linked-group-improperly-obtained-openai-data

25 Matt Haldane. Is DeepSeek's AI 'distillation' theft OpenAI seeks answers over China's breakthrough. *South China Morning Post*. 2025. https://www.scmp.com/tech/big-tech/article/3296827/deepseeks-ai-distillation-theft-openai-seeks-answers-over-chinas-breakthrough

26 Kasagga A, Sapkota A, Changaramkumarath G, *et al*. Performance of ChatGPT and large language models on medical licensing exams worldwide: a systematic review and network meta-analysis with meta-regression. *Cureus* 2025;17(10):e94300.

27 Wu SA, Koo M, Blum L, *et al*. Benchmarking open-source large language models, GPT-4 and Claude 2 on multiple-choice questions in nephrology. *NEJM AI* 2024;1(2): Aldbp2300092.

28 Amiri I, Jiao C, Zhang AY. Multimodal diagnostic accuracy of GPT-4.o, Claude 3.7, and Gemini 2.5 on real-world retina cases. *J Vitreoretin Dis* 2026:24741264261423759.

29 Sussan TT, Sussan RJ, Atkinson AG, *et al*. A comparative evaluation of GPT-4 turbo and Gemini-pro in medical licensing exams: enhancing artificial intelligence's role in medical education. *Cureus* 2026;18(1):e101101.

30 Abdul Sami M, Abdul Samad M, Parekh K, *et al*. Comparative accuracy of ChatGPT 4.0 and Google Gemini in answering pediatric radiology text-based questions. *Cureus* 2024;16(10):e70897.

31 Rao AS, Esmail KP, Lee RS, *et al*. Large language model performance and clinical reasoning tasks. *JAMA Netw Open* 2026;9(4):e264003.