TITLE

Performance of DeepSeek-R1 in Ophthalmology: An Evaluation of Clinical Decision-Making and Cost-Effectiveness

AUTHORS

David Mikhail MD(C) MSc(C)[1], Andrew Farah MDCM(C)[2], Jason Milad BSE(C)[4], Wissam Nassrallah MD PhD[3], Andrew Mihalache MD(C)[1], Daniel Milad MD[3,5,6], Fares Antaki MDCM FRCSC[3,5,6,7,8], Michael Balas MD[9], Marko M. Popovic MD MPH FRCSC[9,10], Rajeev H. Muni MD MSc FRCSC[9,11], Pearse A. Keane MD FRCOphth[12,13], Renaud Duval MD FRCSC[5,6]

AFFILIATIONS

[1]Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada
[2]Faculty of Medicine, McGill University, Montreal, Quebec, Canada
[3]Department of Ophthalmology, Centre Hospitalier de l'Université de Montréal (CHUM), Montreal, Quebec, Canada
[4]Department of Software Engineering, University of Waterloo, Waterloo, Ontario, Canada
[5]Department of Ophthalmology, University of Montreal, Montreal, Quebec, Canada
[6]Centre Universitaire d'Ophtalmologie (CUO), Hôpital Maisonneuve-Rosemont, CIUSSS de l'Est-de-l'Île-de-Montréal, Montreal, Quebec, Canada
[7]The CHUM School of Artificial Intelligence in Healthcare (SAIH), Centre Hospitalier de l'Université de Montréal (CHUM), Montreal, Quebec, Canada
[8]Cole Eye Institute, Cleveland Clinic, Cleveland, OH 44195, USA
[9]Department of Ophthalmology and Vision Sciences, University of Toronto, Toronto, Ontario Canada
[10]Retina Division, Stein and Doheny Eye Institutes, Department of Ophthalmology, University of California, Los Angeles, California, United States of America
[11]Department of Ophthalmology, St. Michael's Hospital/Unity Health Toronto, Toronto, Ontario, Canada
[12]Institute of Ophthalmology, University College London, London, UK
[13]NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust, London, UK

CORRESPONDING AUTHOR

Renaud Duval, MD FRCSC
Department of Ophthalmology, Université de Montréal
2900 Édouard Montpetit Boulevard, Montreal, Quebec, Canada, H3T 1J4
Telephone: (514) 252-3400
Email: renaud.duval@gmail.com

Running Head: Performance of DeepSeek-R1 in Ophthalmology

Figures: 1
Tables: 2
Supplemental Material: None.

Financial Disclosures: None.
Conflicts of Interest: TBD

Keywords: DeepSeek; ChatGPT; artificial intelligence; large language models

Abstract Word Count: 306
Body Word Count: 2401

ABSTRACT

Purpose: To compare the performance and cost-effectiveness of DeepSeek-R1 with OpenAI o1 in diagnosing and managing ophthalmology clinical cases.

Study Design: Cross-sectional evaluation.

Methods: A total of 300 clinical cases spanning 10 different ophthalmology subspecialties were collected from StatPearls. Each case presented a multiple-choice question regarding the diagnosis or management of the case. DeepSeek-R1 was accessed through its public chat-based interface, while OpenAI o1 was queried via an API with a standardized temperature setting of 0.3. Both models were prompted using the Plan-and-Solve+ (PS+) prompt engineering method, instructing them to systematically solve each case. Performance was calculated as the proportion correctly answered multiple choice questions. McNemar's test was employed to compare the two models' performance on paired data. Inter-model agreement for correct diagnoses was evaluated via Cohen's kappa. A token-based cost analysis was performed to estimate the comparative expenditures of running each model at scale, accounting for both input prompts and model-generated output.

Results: DeepSeek-R1 and OpenAI o1 achieved identical overall performance of 82.0% (246/300; 95% CI: 77.3-85.9). Subspeciality-specific analysis revealed numerical variation in performance, though did not reach statistical significance ($p>0.05$). Agreement between the models was moderate overall ($\kappa=0.503$, $p<0.001$), with substantial agreement in Refractive Management/Intervention ($\kappa=0.698$, $p<0.001$) and moderate agreement in Retina/Vitreous ($\kappa=0.561$,$p<0.001$) and Ocular Pathology/Oncology ($\kappa=0.495$, $p<0.01$) cases. Cost analysis indicated an approximately 15-fold reduction in per-query, token-related expenses when using DeepSeek-R1 compared with OpenAI o1 for the same workload.

Conclusions: DeepSeek-R1 demonstrates robust diagnostic reasoning and management decision-making capabilities, performing comparably to OpenAI o1 across a range of ophthalmic subspecialty cases, while also offering a substantial reduction in usage costs. These findings highlight the feasibility of utilizing open-weight, reinforcement learning-augmented LLMs as an accessible, cost-effective alternative to proprietary models. Further investigations should re-evaluate safety guardrails and assess the performance of self-hosted versions of DeepSeek-R1 with domain-specific ophthalmic expertise to optimize clinical utility.

INTRODUCTION

Artificial intelligence (AI) is increasingly being explored in ophthalmology, particularly for its potential to augment diagnostic precision and support clinical decision-making. Large language models (LLMs), a subset of foundation models trained on extensive textual datasets, boast advanced natural language processing (NLP) capabilities and the ability to engage in complex reasoning tasks.[1–3]

In ophthalmology, LLMs have been tested on their ability to answer board-style multiple-choice questions, analyze clinical cases, and interpret ophthalmic images.[4–9] Studies evaluating Generative Pretrained Transformer (GPT) 4 and other advanced models have demonstrated competitive performance with human clinicians in responding to ophthalmic questions, positioning them as valuable tools for clinical reasoning and knowledge retrieval.[10,11] However, while these models excel in structured question-answer formats, studies have highlighted limitations in their ability to manage complex, case-based diagnostic reasoning, particularly in multi-step clinical decision-making and ophthalmic image interpretation.[12,13]

DeepSeek-R1, developed by DeepSeek-AI, has gained global attention for its reasoning-centric design, leveraging reinforcement learning (RL) to improve problem-solving capabilities.[14] Unlike traditional LLMs, which rely heavily on supervised fine-tuning (SFT), DeepSeek-R1 integrates multi-stage training, cold-start data, and self-evolution through RL to refine its logical reasoning and decision-making skills. It was designed with a focus on logical inference, real-time problem-solving, and structured reasoning, making it distinct from generative models that primarily excel in content synthesis. A defining characteristic of DeepSeek-R1 is that it is free to use and reuse, democratizing AI development.[15] DeepSeek-AI reports that it was trained at a fraction of the cost of other leading LLMs, with an estimated $5.6 million compute budget, compared to the $60 million for Meta’s Llama 3.1 405B, $78 million for OpenAI's GPT-4, and $191 million for Google’s Gemini Ultra.[16,17]

Despite its promising capabilities, DeepSeek-R1's performance in ophthalmology has not been evaluated. This study represents the first systematic assessment of DeepSeek-R1’s diagnostic accuracy and decision-making effectiveness in ophthalmology, using clinical cases from StatPearls. By benchmarking DeepSeek-R1 against OpenAI’s state-of-the-art model (o1), we aim to evaluate its performance on diagnosis and management of ocular pathologies, and economic feasibility for potential clinical integration. Given the growing interest in AI-driven diagnostic tools, this study provides critical insights into the viability of open-access, reasoning-based LLMs in ophthalmology and their potential role as an alternative path to proprietary AI models for cost-effective, AI-assisted clinical decision-making.

METHODS

Data Source

This was a cross-sectional study conducted in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) and guidelines.[18] We also conducted this study in accordance with the transparent reporting of a multivariable model for individual prognosis or diagnosis (TRIPOD)-LLM, which is an extension of the TRIPOD + artificial intelligence statement that considers the unique challenges of LLM usage in healthcare.[19] We used 300 cases from an online database of questions from StatPearls.[20] These cases covered a broad range of ophthalmology subspecialities and topics. Each case was classified into one of 13 ophthalmology subspecialties, as defined by the American Academy of Ophthalmology's *Basic and Clinical Science Course*.[21] These included Cataract/Anterior Segment (n=10), Neuro-ophthalmology (n=31), Oculoplastics (n=33), Pediatric Ophthalmology/Strabismus (n=13), Retina and Vitreous (n=69), Cornea/External Disease (n=27), Glaucoma (n=23), Ocular Pathology/Oncology (n=46), Refractive Management/Intervention (n=34), Uveitis (n=14). There were no cases collected on Clinical Optics and Fundamentals.

LLM Access and Parameters

We accessed DeepSeek-R1 through its official chat user interface (UI), which provided direct interaction with the model without requiring the Application Program Interface (API).[22] Unlike API-based implementations, the web platform does not allow for automated batch processing, necessitating manual input for each query. DeepSeek-R1, like other LLMs, uses a "temperature" parameter to control response variability when given identical prompts. The temperature scale ranges from 0 (producing the most deterministic responses) to 1 (yielding highly creative outputs). However, the platform does not provide direct control over this setting on the website platform, meaning responses were generated using the default configuration.

We accessed OpenAI o1, using the API.[23] This enabled us to implement customized automated mass prompting techniques via Google Sheets. Additionally, OpenAI o1's temperature was set to 0.3. While the optimal temperature for our particular application of the model has not been definitively established, our previous research found that a setting of 0.3 resulted in the highest accuracy.[24]

Prompt Engineering

Our previous analysis identified the Zero-Shot Plan-and-Solve + (PS+) prompt as the most effective approach, leading us to select it for this study. Originally developed by Wang et al., the PS+ prompt instructs the model to break down its task into structured steps, executing them sequentially with detailed guidance.[25] Each clinical case was presented with a single prompt, which included the complete case description, a multiple-choice question, and the available answer choices. The questions followed a standardized format, consisting of one correct answer and three distractor options.

Statistical Analysis

Performance was calculated as the proportion of correct responses out of the total number of cases assessed. To compare the performance of DeepSeek-R1 and OpenAI o1, McNemar's test was employed to assess the statistical significance of differences in accuracy between the two models when applied to the same dataset.[26] This test, designed for paired categorical data, evaluates whether one model consistently outperformed the other. To quantify the level of agreement between the models, Cohen's kappa (κ) was calculated, providing insight into the extent to which the two models arrived at similar conclusions beyond what would be expected by chance.[27] Agreement was categorized as the following: 0-0.20 (none to slight agreement), 0.21-0.40 (fair agreement), 0.41-0.60 (moderate agreement), 0.61-0.80 (substantial agreement), and 0.81-1.00 (near perfect agreement). Confidence intervals for accuracy estimates were calculated using the Wilson Score Interval.[28] All statistical tests were two-tailed, with a 5% significance threshold. Statistical analyses were performed with R version 4.4.2 (R Foundation for Statistical Computing).

RESULTS

DeepSeek-R1 achieved an overall accuracy of 82.0% (246/300; 95% CI: 77.3-85.9), which was identical to OpenAI o1's accuracy of 82.0% (246/300; 95% CI: 77.3-85.9) ($p=1.000$) (**Table 1**). Compared to OpenAI o1, DeepSeek-R1 exhibited numerically higher diagnostic accuracy than OpenAI o1 in 4 of the 10 subspecialties analyzed, including Cornea/External Disease (88.9% vs. 85.2%), Glaucoma (95.7% vs. 87.0%), Retina/Vitreous (82.6% vs. 81.2%), and Uveitis (85.7% vs. 71.4%). However, none of these differences reached statistical significance ($p>0.05$). Likewise, OpenAI o1 showed higher numerical accuracy values for Cataract/Anterior Segment (70.0% vs. 60.0%), Neuro-ophthalmology (93.5% vs. 87.1%), Ocular Pathology/Oncology (76.9% vs. 69.2%), Oculoplastics (84.8% vs. 81.8%), and Pediatric Ophthalmology (82.6% vs. 80.4%) (all $p>0.05$). Both models scored the same on Refractive Management/Intervention (73.5%). **Figure 1** compares the performance of both models by subspeciality.

Inter-Model Agreement

Cohen's kappa was calculated to assess agreement between DeepSeek-R1 and OpenAI o1 for correct diagnoses across subspecialties (**Table 2**). Overall agreement across all cases was moderate ($\kappa=0.503$, $p<0.001$). Agreement levels varied by subspecialty, with substantial agreement in Refractive Management/Intervention ($\kappa=0.698$, $p<0.001$) and moderate agreement in Retina/Vitreous ($\kappa=0.561$, $p<0.001$) and Ocular Pathology/Oncology ($\kappa=0.495$, $p<0.01$) cases.

Cost-Effectiveness

Across all 300 cases, there were a total of 386,778 characters used to prompt both models. According to OpenAI's Tokenizer tool, one token is equivalent to approximately 4 characters of text for English text.[29] Thus, the total number of input tokens used to prompt DeepSeek-R1 and

OpenAI o1 was approximately 96,695 tokens. In total, OpenAI o1's responses were 422,046 characters (105,512 tokens), while DeepSeek-R1's responses were 855,961 characters (213,990 tokens) altogether. Under the default pay-per-token rates for OpenAI's "o1-2024-12-17" model ($15.00 USD per 1 million uncached input tokens and $60.00 USD per 1 million uncached output tokens), these 300 prompts would have cost approximately $7.78 USD in total.[30] The cost breaks down to roughly $1.45 for input and $6.33 for output. Alternatively, OpenAI offers two monthly subscription packages for its chat UI, which costs $20/month for up to 50 o1 prompts per week or $200/month for unlimited o1prompts.

By contrast, DeepSeek-R1 is free to use through its public chat UI and mobile app. Thus, we incurred no direct per-query fees during our analysis. DeepSeek also offers an API option for its R1 model, "deepseek-reasoner," which has a token-based billing model—$0.55 USD per 1 million input tokens and $2.19 USD per 1 million output tokens, including both its chain-of-thought (CoT) and final answer.[31] Had the API been used, it would have cost $0.52 USD overall. Thus, comparing the estimated API charges, DeepSeek-R1's cost amounts to 6.71% of OpenAI o1's total, or a 14.91-fold cost reduction per-query expenditures at the prompting volume examined in this study.

DISCUSSION

DeepSeek-R1 demonstrated comparable reasoning capabilities in ophthalmology cases, exactly matching the performance of OpenAI o1 of 82.0% (95% CI: 77.3-85.9). Subspecialty-level analysis revealed similar accuracy levels across each subspeciality, with no significant differences in performance. Despite minor variations, the performance consistency across the subspecialties underscores the capability of both models to handle a broad spectrum of ophthalmic cases. Refractive Management/Intervention, Ocular Pathology/Oncology, and Retina/Vitreous cases revealed moderate to substantial agreement ($p<0.01$). With an overall moderate agreement between the models ($\kappa=0.503$, $p<0.001$), DeepSeek-R1 performed as well as OpenAI o1 while maintaining an approximately 15-fold cost reduction, making it a highly cost-effective alternative. Our results align with those of Mondillo et al., who evaluated DeepSeek-R1 in pediatric decision support and noted that while OpenAI o1 achieved slightly higher accuracy (92.8% vs. 87.0%), DeepSeek-R1 demonstrated superior adaptability and accessibility.[32]

In our study, DeepSeek-R1's comparable performance to OpenAI's highest performing model can be attributed to its innovative training methodology. Initially, DeepSeek-R1 undergoes a "cold start" phase, where it is trained on a diverse set of carefully selected reasoning datasets before RL to establish a foundational understanding.[14] DeepSeek-R1 employs a Mixture of Experts (MoE) strategy, which activates specialized reasoning pathways tailored to specific tasks. This approach improves its handling of domain-specific challenges, enabling it to specialize in complex problem-solving across multiple disciplines.[33] Additionally, DeepSeek-R1 employs rejection sampling and supervised fine-tuning, generating multiple responses to a given problem and selecting only the most accurate and logically coherent outputs for further refinement. This iterative process allows the model to continuously correct mistakes, reinforce high-quality reasoning patterns, and optimize its performance. Prior to its final output, DeepSeek-R1 produces CoT outputs, explicitly discussing its intermediate reasoning steps. Within these responses, the model autonomously revises its reasoning, recognizing errors and correcting them in real-time. This remarkable self-reflection process, referred to as the "aha moment," closely resembles human-like cognitive adjustments, allowing the model to iteratively improve its problem-solving skills.[33,34]

Economically, DeepSeek-R1 presents a compelling advantage. Our analysis revealed that, even when utilizing the official API with token-based billing, DeepSeek-R1's marginal per-query costs are far below those of OpenAI's o1. It is important to note that DeepSeek-R1 is not fully open-sourced. Instead, it is released under an MIT license as an "open-weight" model, meaning users have access to its pre-trained weights and can build upon its architecture, but the underlying training data remains undisclosed.[17] Nonetheless, DeepSeek-R1 remains an attractive option for institutions seeking the flexibility of self-hosting a model to reduce costs. While self-hosting may involve infrastructure expenses, the ability to customize and optimize the deployment can lead to long-term savings, particularly for large-scale clinical operations or resource-limited settings. For example, one study evaluated the replacement of OpenAI's GPT-4 with open-source small language models (SLMs) in a production environment. The research demonstrated that SLMs provided competitive performance while achieving a cost reduction ranging from 5 to 29 times compared to GPT-4.[35]

One of the key advantages of DeepSeek-R1 is its reasoning-centric design, which could make its decision-making process more interpretable and transparent. Traditional LLMs are often "black boxes," where users receive an output without insight into how the model arrived at its conclusion. For clinicians, DeepSeek-R1's can enable verification of whether the model's conclusions align with their own clinical knowledge. Nonetheless, this feature does not eliminate the fundamental issue of AI-generated "hallucinations"—instances where the model produces plausible but incorrect medical information. According to Wang et al., these errors pose a significant risk in patient care, particularly when AI outputs influence diagnoses, treatment recommendations, or research conclusions.[36] Furthermore, the fact that DeepSeek-R1 does not reveal its training corpus means that even when reasoning is made explicit, its responses may still biased by incorrect or misrepresentations of medical knowledge.

DeepSeek-R1's self-reflective capability can even be the means by which it bypasses its own safety constraints, which raises security concerns over the model generating outputs that deviate from established medical guidelines or rationalize incorrect treatments that could endanger patient care. External guardrails such as rule-based reinforcement filters, human-in-the-loop verification, and continuous real-world validation, will be essential for safe deployment.[37] Additionally, retrieval-augmented generation (RAG) techniques and external fact-checking tools could be used to ground the model's responses in up-to-date, peer-reviewed medical literature.[38]

This study has several limitations that warrant discussion. Firstly, our cost analysis was based solely on the final answers generated by the models, excluding CoT outputs. Since CoT reasoning involves additional tokens to articulate intermediate steps, our approach likely underestimated the actual token usage and associated costs. However, this limitation applied uniformly to both DeepSeek-R1 and OpenAI o1, ensuring a fair comparison. Secondly, we did not include the associated images within the cases to test multimodal performance. While both models can handle images, DeepSeek-R1 notes that its image analysis solely extracts text from images. Additionally, we employed the PS+ prompt, which was originally designed to enhance reasoning in models lacking structured problem-solving capabilities, raising the question of whether alternative prompting strategies might yield different results in inherently reasoning-centric models. Furthermore, the publicly available dataset from StatPearls may have been included in the training data of one or both models, leading to potential data leakage. This could artificially inflate model performance by allowing it to recognize or recall previously encountered cases rather than applying genuine clinical reasoning.[39] However, without direct access to the training corpora, we cannot verify whether the identical high performance of both models resulted from prior exposure to the dataset or from the cases being relatively straightforward and well-aligned with the models' capabilities. Thus, future research should aim to evaluate these models using larger, more diverse datasets, including those that incorporate ophthalmic imaging, to better assess their applicability in clinical practice. Emerging multimodal AI models, such as Kimi K1.5, have begun integrating vision-language capabilities, offering new possibilities for ophthalmic imaging analysis.[33,40,41]

CONCLUSION

DeepSeek-R1 demonstrated comparable performance to OpenAI o1 across 300 ophthalmology cases, while also offering a substantial cost advantage. Its enhanced reasoning-centric design appears well suited to a range of clinical scenarios in ophthalmology, supporting its potential as a more accessible AI-driven decision support tool. However, continued research on multimodal prompts and robust external guardrails are necessary to confirm its safety and efficacy in clinical practice.

ACKNOWLEDGEMENTS

None.

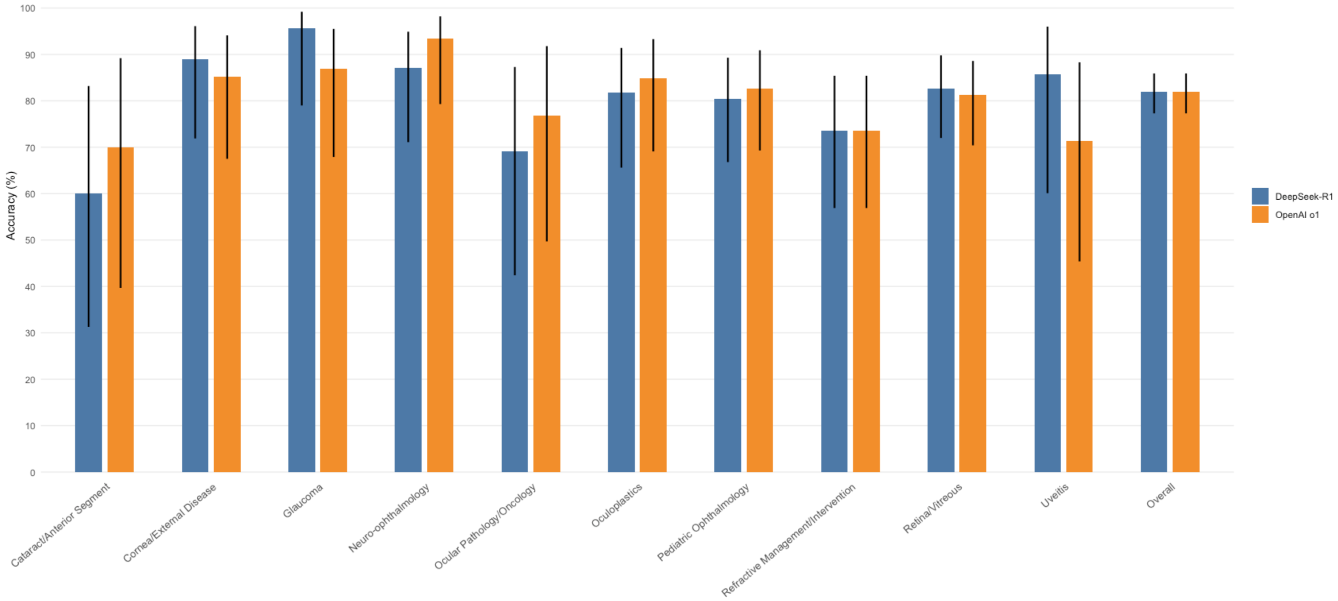

Figure 1. Performance of DeepSeek-R1 and OpenAI by Subspeciality

REFERENCES

1. Chia MA, Antaki F, Zhou Y, Turner AW, Lee AY, Keane PA. Foundation models in ophthalmology. Br J Ophthalmol [Internet]. 2024 Jun 4 [cited 2024 Jun 5]; Available from: https://bjo.bmj.com/content/early/2024/06/04/bjo-2024-325459

2. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, et al. The shaky foundations of large language models and foundation models for electronic health records. Npj Digit Med. 2023 Jul 29;6(1):1–10.

3. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the Opportunities and Risks of Foundation Models [Internet]. arXiv; 2022 [cited 2024 Apr 9]. Available from: http://arxiv.org/abs/2108.07258

4. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. Ophthalmol Sci. 2023 Dec;3(4):100324.

5. Mihalache A, Popovic MM, Muni RH. Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. JAMA Ophthalmol. 2023 Jun 1;141(6):589–97.

6. Milad D, Antaki F, Milad J, Farah A, Khairy T, Mikhail D, et al. Assessing the medical reasoning skills of GPT-4 in complex ophthalmology cases. Br J Ophthalmol. 2024 Oct 1;108(10):1398–405.

7. Mihalache A, Huang RS, Mikhail D, Popovic MM, Shor R, Pereira A, et al. Interpretation of Clinical Retinal Images Using an Artificial Intelligence Chatbot. Ophthalmol Sci. 2024 May 23;100556.

8. Mikhail D, Mihalache A, Huang RS, Khairy T, Popovic MM, Milad D, et al. Performance of ChatGPT in French language analysis of multimodal retinal cases. J Fr Ophtalmol. 2025 Mar 1;48(3):104391.

9. Silhadi M, Nassrallah WB, Mikhail D, Milad D, Harissi-Dagher M. Assessing the performance of Microsoft Copilot, GPT-4 and Google Gemini in ophthalmology. Can J Ophthalmol [Internet]. 2025 Jan 22 [cited 2025 Feb 6];0(0). Available from: https://www.canadianjournalofophthalmology.ca/article/S0008-4182(25)00001-8/fulltext

10. Taloni A, Borselli M, Scarsi V, Rossi C, Coco G, Scorcia V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. Sci Rep. 2023 Oct 29;13(1):18562.

11. Zhang J, Ma Y, Zhang R, Chen Y, Xu M, Rina S, et al. A comparative study of GPT-4o and human ophthalmologists in glaucoma diagnosis. Sci Rep. 2024 Dec 5;14(1):30385.

12. Agnihotri AP, Nagel ID, Artiaga JCM, Guevarra MCB, Sosuan GMN, Kalaw FGP. Large language models in ophthalmology: A review of publications from top ophthalmology journals. Ophthalmol Sci [Internet]. 2024 Dec 17 [cited 2025 Feb 1];0(0). Available from: https://www.ophthalmologyscience.org/article/S2666-9145(24)00217-3/fulltext

13. Antaki F, Chopra R, Keane PA. Vision-Language Models for Feature Detection of Macular Diseases on Optical Coherence Tomography. JAMA Ophthalmol. 2024 Jun 1;142(6):573–6.

14. DeepSeek-AI, Guo D, Yang D, Zhang H, Song J, Zhang R, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning [Internet]. arXiv; 2025 [cited 2025 Feb 4]. Available from: http://arxiv.org/abs/2501.12948

15. deepseek-ai/DeepSeek-R1 · Hugging Face [Internet]. 2025 [cited 2025 Feb 4]. Available from: https://huggingface.co/deepseek-ai/DeepSeek-R1

16. AI Index Report 2024 – Artificial Intelligence Index [Internet]. [cited 2025 Feb 4]. Available from: https://aiindex.stanford.edu/report/

17. Gibney E. China's cheap, open AI model DeepSeek thrills scientists. Nature. 2025 Jan 23;638(8049):13–4.

18. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008 Apr;61(4):344–9.

19. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025 Jan;31(1):60–9.

20. StatPearls [Internet]. [cited 2025 Feb 8]. Available from: https://www.statpearls.com/

21. McCannel CA, Bhatti MT. The Basic and Clinical Science Course of the American Academy of Ophthalmology: The 50th Anniversary of a Unicorn Among Medical Textbooks. JAMA Ophthalmol. 2022 Mar 1;140(3):225–6.

22. DeepSeek [Internet]. [cited 2025 Feb 4]. Available from: https://chat.deepseek.com

23. Introducing OpenAI o1 [Internet]. 2024 [cited 2025 Feb 4]. Available from: https://openai.com/o1/

24. Antaki F, Milad D, Chia MA, Giguère CÉ, Touma S, El-Khoury J, et al. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. Br J Ophthalmol. 2024 Oct 1;108(10):1371–8.

25. Wang L, Xu W, Lan Y, Hu Z, Lan Y, Lee RKW, et al. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models [Internet]. arXiv.org. 2023 [cited 2024 Jun 3]. Available from: https://arxiv.org/abs/2305.04091v3

26. Sundjaja JH, Shrestha R, Krishan K. McNemar And Mann-Whitney U Tests. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 [cited 2025 Feb 4]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK560699/

27. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213–20.

28. Bender R. Calculating confidence intervals for the number needed to treat. Control Clin Trials. 2001 Apr;22(2):102–10.

29. OpenAI Platform [Internet]. [cited 2025 Feb 5]. Available from: https://platform.openai.com

30. Pricing [Internet]. [cited 2025 Feb 6]. Available from: https://openai.com/api/pricing/

31. Models & Pricing | DeepSeek API Docs [Internet]. [cited 2025 Feb 6]. Available from: https://api-docs.deepseek.com/quick_start/pricing

32. Mondillo G, Colosimo S, Perrotta A, Frattolillo V, Masino M. Comparative Evaluation of Advanced AI Reasoning Models in Pediatric Clinical Decision Support: ChatGPT O1 vs. DeepSeek-R1 [Internet]. medRxiv; 2025 [cited 2025 Feb 6]. p. 2025.01.27.25321169. Available from: https://www.medrxiv.org/content/10.1101/2025.01.27.25321169v1

33. Mercer S, Spillard S, Martin DP. Brief analysis of DeepSeek R1 and its implications for Generative AI [Internet]. arXiv; 2025 [cited 2025 Feb 6]. Available from: http://arxiv.org/abs/2502.02523

34. Kounios J, Beeman M. The aha! Moment: The cognitive neuroscience of insight. Curr Dir Psychol Sci. 2009;18(4):210–6.

35. Irugalbandara C, Mahendra A, Daynauth R, Arachchige TK, Dantanarayana J, Flautner K, et al. Scaling Down to Scale Up: A Cost-Benefit Analysis of Replacing OpenAI's LLM with Open Source SLMs in Production [Internet]. arXiv; 2024 [cited 2025 Feb 6]. Available from: http://arxiv.org/abs/2312.14972

36. Safety challenges of AI in medicine in the era of large language models [Internet]. [cited 2025 Feb 6]. Available from: https://arxiv.org/html/2409.18968v2?utm_source=chatgpt.com

37. Pantha N, Ramasubramanian M, Gurung I, Maskey M, Ramachandran R. Challenges in Guardrailing Large Language Models for Science [Internet]. arXiv; 2024 [cited 2025 Feb 6]. Available from: http://arxiv.org/abs/2411.08181

38. Rebedea T, Dinu R, Sreedhar M, Parisien C, Cohen J. NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails [Internet]. arXiv; 2023 [cited 2025 Feb 6]. Available from: http://arxiv.org/abs/2310.10501

39. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023 Aug 4;4(9):100804.

40. Team K, Du A, Gao B, Xing B, Jiang C, Chen C, et al. Kimi k1.5: Scaling Reinforcement Learning with LLMs [Internet]. arXiv; 2025 [cited 2025 Feb 6]. Available from: http://arxiv.org/abs/2501.12599

41. MoonshotAI/Kimi-k1.5 [Internet]. Moonshot AI; 2025 [cited 2025 Feb 6]. Available from: https://github.com/MoonshotAI/Kimi-k1.5

Table 1. Performance of DeepSeek-R1 and OpenAI o1 by Subspeciality

Model	Subspeciality	Accuracy (%)	95% CI (%)	p-value
DeepSeek-R1	Cataract/Anterior Segment	60.0	[31.3, 83.2]	1.000
OpenAI o1		70.0	[39.7, 89.2]	
DeepSeek-R1	Cornea/External Disease	88.9	[71.9, 96.1]	1.000
OpenAI o1		85.2	[67.5, 94.1]	
DeepSeek-R1	Glaucoma	95.7	[79.0, 99.2]	0.625
OpenAI o1		87.0	[67.9, 95.5]	
DeepSeek-R1	Neuro-ophthalmology	87.1	[71.1, 94.9]	0.625
OpenAI o1		93.5	[79.3, 98.2]	
DeepSeek-R1	Ocular Pathology/Oncology	69.2	[42.4, 87.3]	1.000
OpenAI o1		76.9	[49.7, 91.8]	
DeepSeek-R1	Oculoplastics	81.8	[65.6, 91.4]	1.000
OpenAI o1		84.8	[69.1, 93.3]	
DeepSeek-R1	Pediatric Ophthalmology	80.4	[66.8, 89.3]	1.000
OpenAI o1		82.6	[69.3, 90.9]	
DeepSeek-R1	Refractive Management/Intervention	73.5	[56.9, 85.4]	1.000
OpenAI o1		73.5	[56.9, 85.4]	
DeepSeek-R1	Retina/Vitreous	82.6	[72.0, 89.8]	1.000
OpenAI o1		81.2	[70.4, 88.6]	
DeepSeek-R1	Uveitis	85.7	[60.1, 96.0]	0.500
OpenAI o1		71.4	[45.4, 88.3]	
DeepSeek-R1	Overall	82.0	[77.3, 85.9]	1.000
OpenAI o1		82.0	[77.3, 85.9]	

Table 2. Agreement between DeepSeek-R1 and OpenAI o1

Subspecialty	**Cohen's Kappa**	**p-value**
Cataract/Anterior Segment	0.783	0.067
Cornea/External Disease	0.509	0.069
Glaucoma	-0.070	1.000
Neuro-ophthalmology	0.271	0.598
Ocular Pathology/Oncology	**0.495**	**0.004**
Oculoplastics	0.238	0.457
Pediatric Ophthalmology	0.418	0.411
Refractive Management/Intervention	**0.698**	**0.000**
Retina/Vitreous	**0.561**	**0.000**
Uveitis	0.588	0.116
Overall	**0.503**	**0.000**