AMERICAN ACADEMY OF OPHTHALMOLOGY®

Can OpenAI's New o1 Model Outperform Its Predecessors in Common Eye Care Queries?

Krithi Pushpanathan, MSc,[1,2,] Minjie Zou, MMed,[1,2,*] Sahana Srinivasan, BEng,[1,2] Wendy Meihua Wong, MMed,[1,2,4] Erlangga Ariadarma Mangunkusumo, MD,[1,2,4] George Naveen Thomas, MMed,[1,2,4] Yien Lai, MMed,[1,2,4] Chen-Hsin Sun, MD,[1,2,4] Janice Sing Harn Lam, MMed,[1,2,4] Marcus Chun Jin Tan, MMed,[1,2,4] Hazel Anne Hui'En Lin, MMed,[1,2,4] Weizhi Ma, PhD,[5] Victor Teck Chang Koh, MMed,[1,2,4] David Ziyou Chen, MMed,[1,2,4,*] Yih-Chung Tham, PhD[1,2,3,6,*]*

Objective: The newly launched OpenAI o1 is said to offer improved reasoning, potentially providing higher quality responses to eye care queries. However, its performance remains unassessed. We evaluated the performance of o1, ChatGPT-4o, and ChatGPT-4 in addressing ophthalmic-related queries, focusing on correctness, completeness, and readability.

Design: Cross-sectional study.

Subjects: Sixteen queries, previously identified as suboptimally responded to by ChatGPT-4 from prior studies, were used, covering 3 subtopics: myopia (6 questions), ocular symptoms (4 questions), and retinal conditions (6 questions).

Methods: For each subtopic, 3 attending-level ophthalmologists, masked to the model sources, evaluated the responses based on correctness, completeness, and readability (on a 5-point scale for each metric).

Main Outcome Measures: Mean summed scores of each model for correctness, completeness, and readability, rated on a 5-point scale (maximum score: 15).

Results: O1 scored highest in correctness (12.6) and readability (14.2), outperforming ChatGPT-4, which scored 10.3 ($P = 0.010$) and 12.4 ($P < 0.001$), respectively. No significant difference was found between o1 and ChatGPT-4o. When stratified by subtopics, o1 consistently demonstrated superior correctness and readability. In completeness, ChatGPT-4o achieved the highest score of 12.4, followed by o1 (10.8), though the difference was not statistically significant. o1 showed notable limitations in completeness for ocular symptom queries, scoring 5.5 out of 15.

Conclusions: While o1 is marketed as offering improved reasoning capabilities, its performance in addressing eye care queries does not significantly differ from its predecessor, ChatGPT-4o. Nevertheless, it surpasses ChatGPT-4, particularly in correctness and readability.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2025;5:100745 © 2025 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

Large language models (LLMs) are a crucial aspect of natural language processing, a branch of artificial intelligence focused on language prediction.[1] Through extensive training on vast amounts of text, LLMs can perform various tasks, including question-answering and summarization.[1] Their capacity to engage in human-like interactions positions LLMs as transformative tools in numerous sectors, including health care.[2–4]

Large language models have the potential to revolutionize medical fields by improving how patients access and understand medical information. By delivering immediate, tailored responses, these models facilitate patient understanding of complex medical concepts, treatment options, and decision-making processes.[3,5] This bridges existing gaps in health care literacy while enhancing the overall patient experience.[6] Among these models, OpenAI's ChatGPT has gained prominence for its ability to make medical information more accessible,[7,8] including ophthalmology-related information.[9–11]

The latest iteration of ChatGPT, OpenAI o1 (known as o1), features advanced reasoning capabilities that facilitate nuanced responses to complex queries,[12] representing a potential advancement over its predecessors. To that end, building on the previous improvements seen in ChatGPT-4o,[13] o1 has the potential to more effectively address common eye care queries. However, this aspect has not been evaluated.

The aim of this study was to evaluate the comparative performance of o1 to its predecessors, ChatGPT-4 and ChatGPT-4o, in the quality of responses to ophthalmic inquiries across 3 evaluation dimensions. Findings from this

https://doi.org/10.1016/j.xops.2025.100745
ISSN 2666-9145/25

study may provide insights into the effectiveness of the newly launched o1 in addressing common eye care queries.

Methods

Study Design and Setting

The study was conducted between September 16 and 26, 2024, at the Ophthalmology Department of the National University Hospital, Singapore.

Sixteen commonly inquired patient queries related to eye care were selected from prior studies (conducted in 2023, covering myopia-, ocular symptom-, and retina-related domains) where ChatGPT-4 previously demonstrated suboptimal ratings (poor or borderline) for accuracy in its responses.[14–16] As depicted in Figure 1, we evaluated 3 iterations of ChatGPT—GPT-4 (Legacy Model), GPT-4o, and o1 (referring to the OpenAI o1-preview model)—to compare their performance in responding to eye care queries. Each question was posed to the models as an isolated prompt in individual chats. Memory and model improvement functions were disabled to ensure the models did not retain information from previous interactions or adapt based on earlier responses. In this study, we used the web interface of ChatGPT, where the temperature setting was typically set by OpenAI as 0.70 to 1.00. We did not make further changes to the temperature throughout the course of our study.

A standardized prompt strategy was employed for consistency. It included the instruction, "You may browse the web if needed," appended to the end of each question. Since the queries in this study did not inherently require real-time updates, initiating web searches was explicitly prompted to encourage the models to leverage external sources for more relevant responses.[17] To prevent grader bias, we omitted references, citations, and model-specific information (e.g., training cutoff dates) from the responses. Outputs were anonymized and randomized before evaluation.

Nine attending-level ophthalmologists participated in the evaluation, comprising 3 pediatric ophthalmologists (YL, CHS, JSHL) who evaluated responses on myopia-related questions, 3 general ophthalmologists (MCJT, HAHL, DZC) who assessed answers concerning ocular symptoms, and 3 retinal specialists (WMW, EAM, GNT) who graded responses related to retinal conditions. Each ophthalmologist independently rated the responses for correctness, completeness, and readability on a 5-point scale (1, very poor; 2, poor; 3, borderline; 4, good; 5, very good), as detailed in Table S1 (available at www.ophthalmologyscience.org). They were also encouraged to provide additional comments on the responses if deemed necessary to provide rationale for high or low scores. All evaluations were completed within a 1-week period to ensure consistency in assessment conditions. Although the graders were aware that 3 models were being assessed, they were masked to the specific types of LLMs used.

Statistical Analysis

Statistical analyses were performed using R (version 4.3.1, R Foundation for Statistical Computing). Descriptive statistics summarized the mean, standard deviation, and range of summed scores for each model across the metrics of correctness, completeness, and readability. Differences in mean summed scores across the 3 models were assessed using the Kruskal-Wallis rank sum test as the data did not satisfy parametric assumptions. After this, post hoc pairwise comparisons were conducted using the Dunn test with Bonferroni correction to control for type 1 error in multiple comparisons. Statistical significance was set at $P < 0.05$.

Results

Table 2 presents the performance for each LLM across the metrics of correctness, completeness, and readability. It details the mean scores, standard deviation, and range of scores aggregated across the graders and averaged over the 16 evaluated questions.

Overall, across all evaluated questions, we observed that o1 excelled in generating accurate responses, achieving a mean summed correctness score of 12.6 ± 1.5 out of 15. Pairwise Dunn post hoc test revealed that o1 significantly outperformed ChatGPT-4 (mean score $= 10.3 \pm 2.4$, $P = 0.010$), but there was no significant difference with ChatGPT-4o (11.6 ± 2.0, $P = 0.461$).

On the other hand, in completeness, ChatGPT-4o led with a mean score of 12.4 ± 1.9, significantly outperforming ChatGPT-4 (10.5 ± 2.0, $P = 0.021$). Although ChatGPT-4o achieved a higher mean score than o1 (10.8 ± 4.1), the difference was not statistically significant ($P = 0.854$). For readability, o1 achieved the highest mean score of 14.2 ± 1.2, significantly better than ChatGPT-4 (12.4 ± 1.0, $P < 0.001$) but not significantly different when compared with ChatGPT-4o (13.2 ± 1.1, $P = 0.080$).

Figure 2 illustrates the mean summed scores for each LLM across the subtopics. In addition, the raw individual scores for each question are further detailed in Table 3. Across the subtopics of myopia-, ocular symptom-, and retina-related queries, we consistently observed that o1 scored higher in correctness and readability. In completeness, on the contrary, ChatGPT-4o outperformed o1 for myopia- and ocular symptom—related queries. Notably, o1's completeness score for ocular symptoms was particularly low (mean of 5.5 out of 15, across the 4 ocular symptom-related questions). However, in the retinal subgroup, o1 (mean score = 13.2) achieved a higher completeness score than ChatGPT-4o (11.8), a finding that diverges from the overall trend.

Discussion

This study examined whether advancements in o1's model architecture and its general reasoning capabilities also led to improvements in responses for common eye care inquiries compared with its predecessors, ChatGPT-4 and ChatGPT-4o. As one of the first studies to evaluate the newly released o1 model in this context, our findings provide early insights into its performance for eye care queries. Our results indicate that o1 excelled in generating accurate and highly readable responses, particularly surpassing ChatGPT-4. However, o1 generally performed similarly to ChatGPT-4o across all metrics.

The superior performance of o1 in correctness and readability could be attributed to its enhanced reasoning capabilities.[12] This was particularly evident in the question regarding vision therapy for myopia prevention, where o1 scored highest in this question across all metrics (Table S4, available at www.ophthalmologyscience.org). As shown in the example, its reasoning process highlighted the importance of "current, accurate, and

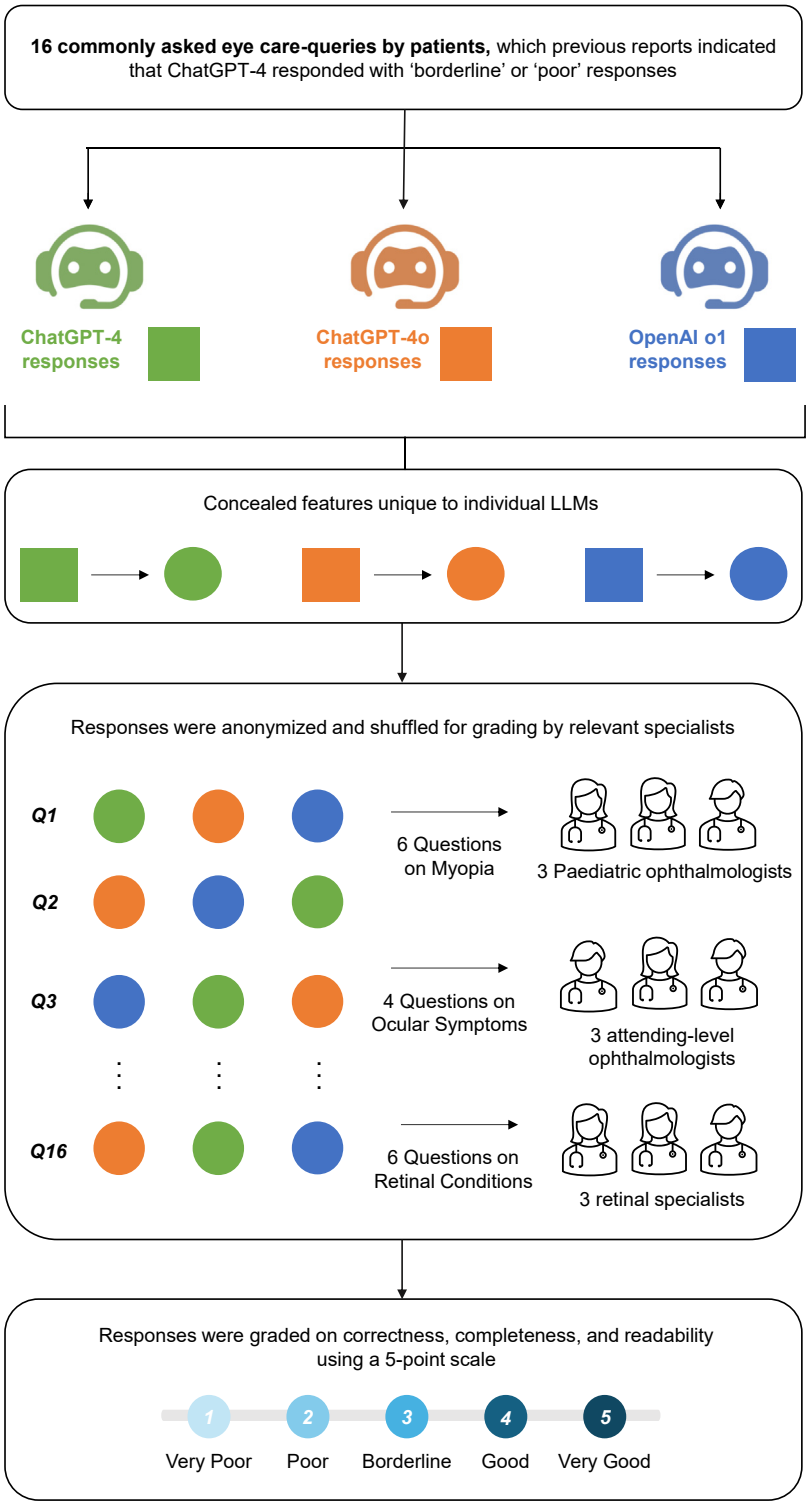

Figure 1. Study design flowchart. LLMs = large language models.

source-cited data" and acknowledged the "limited evidence for eye exercises." Consequently, this yielded a more structured response that differentiated between evidence-based recommendations and those lacking empirical support. Moreover, o1's response seemed to provide a more organized flow of presentation with improved readability,

Table 2. Comparative Performances of ChatGPT Models across Correctness, Completeness, and Readability

Grading Metric	Model		
	ChatGPT-4	*ChatGPT-4o*	*OpenAI o1*
Correctness			
Mean ± SD	10.3 ± 2.4	11.6 ± 2.0	12.6 ± 1.5
Range	6–14	6–14	9–15
Completeness			
Mean ± SD	10.5 ± 2.0	12.4 ± 1.9	10.8 ± 4.1
Range	6–13	8–14	3–14
Readability			
Mean ± SD	12.4 ± 1.0	13.2 ± 1.1	14.2 ± 1.2
Range	11–14	11–15	11–15

SD = standard deviation.

contrasting with the continuous list format used by the other models.

However, o1 still exhibited room for improvement in correctness. For instance, it received its lowest accuracy score for this myopia-related question: "My child has not developed myopia; should he/she start using atropine?" In this case, the model incorrectly stated, "Current evidence does not support the routine use of atropine in children without myopia" (Table S5, available at www.ophthalmologyscience.org). However, a randomized controlled trial published in 2023 demonstrated that nightly use of 0.05% atropine eye drops among premyopes aged 4 to 9 could reduce the incidence of myopia and the rate of myopic shifts over 2 years.[18] Although the model's knowledge cutoff is later than the finding's publication and it has web browsing capability, misinformation persists. This highlights the need for better precision in sourcing and verifying up-to-date clinical evidence.

In terms of completeness, o1 demonstrated suboptimal performance in the ocular symptoms subtopic, where it notably scored low. This shortfall may be attributed to the phrasing of the queries (i.e., the prompts), which were more generic and used broad descriptive terminology such as "sudden blurring of vision" (Table S6, available at www.ophthalmologyscience.org). Consequently, the response generated by o1 failed to provide further elaboration, unlike in responses of other subtopics. In 3 of 4 instances, the model provided equally generic responses like "Sorry you're experiencing this … consult a health care professional." Although these responses were not incorrect, they lacked the clinical specificity to offer meaningful guidance to the patient, and thus were deemed incomplete by the evaluators and received lower scores.

Furthermore, we also conducted an additional qualitative review of questions where ChatGPT-4 previously generated "good-rated" responses in our past 3 studies.[14–16] In this additional review, 1 question was randomly selected from each subtopic, and o1-generated responses were compared with the corresponding GPT-4-generated responses from the prior studies. Two evaluators (K.P. and M.Z.) qualitatively assessed the responses (detailed in Tables S7–S9, available at www.ophthalmologyscience.org). Overall, responses from both models demonstrated similar accuracy and readability. However, o1 exhibited greater depth in myopia- and retinal disease–related queries, providing more specific information on research findings for myopia and offering deeper elaborations on retinal disease severity.

The strengths of our study include masking and randomization to minimize grading bias, thereby enhancing the reliability of our findings. Second, this study built upon insights from our previous research,[14–16] which identified 16 ophthalmic questions where ChatGPT-4's performance was suboptimal. These insights provided a strong foundation for the overall study design, allowing for targeted evaluation of improvements in the newer models (o1 and ChatGPT-4o). Third, we adopted a comprehensive grading approach across 3 dimensions: correctness, completeness, and readability.

Nevertheless, a limitation of this study is that the selection of 16 queries reflects a limited scope, both in terms of breadth (focused on only 3 ophthalmic domains—myopia,

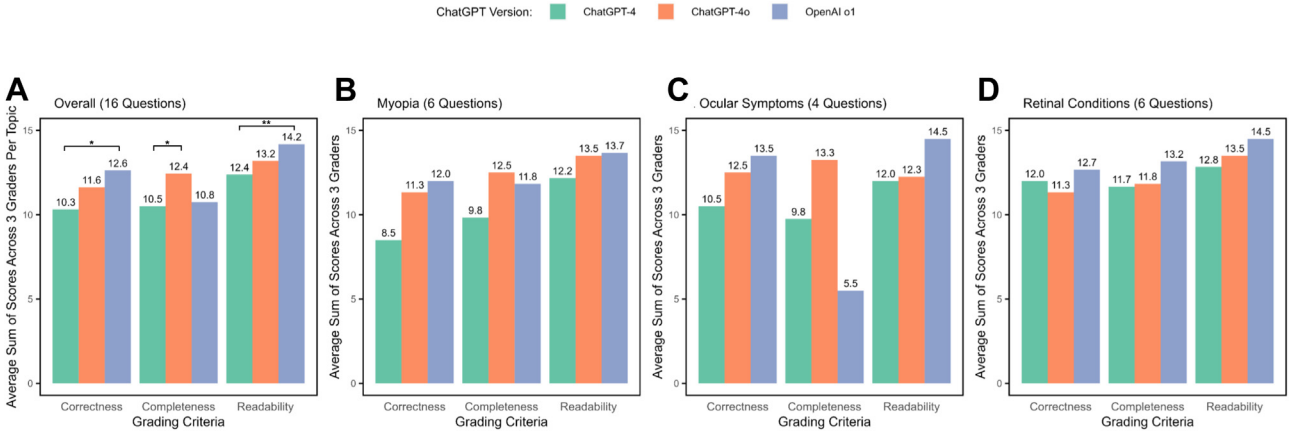

Figure 2. Bar charts comparing mean summed scores for correctness, completeness, and readability across ChatGPT-4, ChatGPT-4o, and OpenAI o1, evaluated on **(A)** overall performance across 16 eye care inquiries, **(B)** inquiries related to myopia, **(C)** inquiries concerning ocular symptoms, and **(D)** inquiries about retinal conditions. *Adjusted $P < 0.05$, **$P < 0.001$ for Dunn test conducted for multiple hypothesis comparisons. Statistical significance testing was not conducted in the topic-specific categories due to small sample sizes.

Table 3. Summed Scores for Correctness, Completeness, and Readability Attained for Each Question Item across ChatGPT Models

Subtopic	No	Question	Summed Scores* for Correctness			Summed Scores* for Completeness			Summed Scores* for Readability		
			ChatGPT-4	*ChatGPT-4o*	*OpenAI o1*	*ChatGPT-4*	*ChatGPT-4o*	*OpenAI o1*	*ChatGPT-4*	*ChatGPT-4o*	*OpenAI o1*
Myopia	1	What are the spectacles/contact lenses available to prevent myopia/myopia progression?	7	11	11	10	11	10	11	13	14
	2	What type of diet/vitamin supplementation can help to prevent myopia/myopia progression?	8	13	13	12	13	13	12	14	13
	3	My child has not developed myopia. Should he/she start using atropine?	9	14	9	9	14	8	14	14	11
	4	My child has myopia. What are the available treatments to slow down the progression of myopia?	7	11	11	9	14	12	12	14	16
	5	How do MiSight/MiYOSMART/Stellest/Abiliti/orthokeratology lenses help to prevent myopia/myopia progression?	14	13	13	13	14	14	13	15	14
	6	Are there any available vision therapy/eye exercises that can help to prevent myopia/myopia progression?	6	6	15	6	8	14	11	12	15
Ocular symptoms	1	Why am I experiencing sudden blurring of vision?	12	12	14	11	12	3	12	11	15
	2	Why am I having double vision?	11	13	14	11	14	3	12	13	15
	3	Why is there something blocking my vision?	9	12	14	8	13	3	12	12	15
	4	Why do I feel very sensitive to glare?	10	13	12	9	14	13	12	13	13
Retinal conditions	1	I have macular degeneration. Do I have to get an eye injection?	12	12	14	11	10	13	12	14	15
	2	How long does an eye injection last?	11	10	13	10	10	14	12	13	15
	3	Does macular degeneration occur in young people?	12	10	13	11	12	13	13	13	15
	4	Are floaters related to macular degeneration?	13	14	13	13	14	13	14	14	13
	5	What are the initial symptoms of diabetic retinopathy?	11	10	11	12	12	13	12	12	14
	6	What is central retinal vein occlusion?	13	12	12	13	13	13	14	15	15

*Summed across 3 ophthalmologist evaluators.

ocular symptoms, and retinal conditions) and depth (small number of questions). This may limit the generalizability of our findings, as it may not fully capture the range of eye care inquiries encountered in clinical practice. Thus, future research should consider a larger and more diverse set of queries, including additional ophthalmic subspecialties, to provide a more comprehensive assessment of model performance across a broader range of clinical scenarios.

Although generic proprietary LLMs show potential for ophthalmic use, their current inability to consistently deliver accurate and comprehensive responses underscores the need for ophthalmology-specific models. These generic LLMs, trained on broad proprietary datasets, can address basic eye care queries effectively but occasionally lack the precision required in clinical contexts. Developing ophthalmology-specific models, grounded in reliable ophthalmic data, could enhance accuracy and relevance, particularly for tasks such as patient education and triaging of ophthalmic conditions, which demand error-free performance.

The findings of this study indicate that o1 could serve as a valuable tool for addressing eye care inquiries, offering more accurate and clearer responses than ChatGPT-4. However, despite its promising reasoning capabilities, o1 did not show significant improvements over ChatGPT-4o across all evaluated metrics. Future studies should expand the range of eye care queries tested to comprehensively assess potential differences in model performance and utility. Moreover, limitations were identified in o1's ability to provide clinically comprehensive answers and maintain accuracy with the latest medical evidence. Further refinements are still necessary before o1 can be considered for clinical implementation in ophthalmology.

Footnotes and Disclosures

Originally received: November 30, 2024.
Final revision: February 1, 2025.
Accepted: February 14, 2025.
Available online: February 22, 2025. Manuscript no. XOPS-D-24-00529.

[1] Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore.

[2] Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore.

[3] Singapore Eye Research Institute, Singapore National Eye Centre, Singapore.

[4] Department of Ophthalmology, National University Hospital, Singapore.

[5] Institute for AI Industry Research, Tsinghua University, Beijing, China.

[6] Eye Academic Clinical Program (Eye ACP), Duke NUS Medical School, Singapore.

*K.P. and M.Z. contributed equally to this work as first authors. D.Z.C. and Y.-C.T. contributed equally to this work as last authors.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The authors have no proprietary or commercial interest in any materials discussed in this article.

This work is supported by the National University of Singapore and Tsinghua University's Joint Initiative Scientific Research Program.

Dr Yih-Chung Tham is supported by the National Medical Research Council of Singapore (NMRC/MOH/HCSAINV21nov-0001).

Support for Open Access publication was provided by the National University of Singapore and Tsinghua University's Joint Initiative Scientific Research Program.

HUMAN SUBJECTS: No human subjects were included in this study.

No animal subjects were included in this study.

Author Contributions:

Conception and design: Pushpanathan, Zou, Koh, Chen, Tham

Analysis and interpretation: Pushpanathan, Zou, Srinivasan, Ma, Chen, Tham

Data collection: Pushpanathan, Zou, Wong, Mangunkusumo, Thomas, Lai, Sun, Lam, Tan, Lin, Chen, Tham

Obtained funding: N/A

Overall responsibility: Pushpanathan, Zou, Srinivasan, Wong, Mangunkusumo, Thomas, Lai, Sun, Lam, Tan, Lin, Ma, Koh, Chen, Tham

Abbreviation and Acronym:
LLM = large language model.

Keywords:
Large language models, OpenAI o1, Myopia, Ocular symptoms, Retinal conditions.

Correspondence:
Yih-Chung Tham, PhD, Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Level 13, MD1 Tahir Foundation Building, 12 Science Drive 2, Singapore 117549, Singapore. E-mail: thamyc@nus.edu.sg.

References

1. Tamkin A, Brundage M, Clark J, Ganguli D. Understanding the capabilities, limitations, and societal impact of large language models. *ArXiv*. 2021. https://doi.org/10.48550/arXiv.2102.02503.
2. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620:172–180.
3. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med*. 2023;3:141.
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med*. 2023;29:1930–1940.
5. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. *Nat Commun*. 2024;15:2050.
6. Amin KS, Mayes LC, Khosla P, Doshi RH. Assessing the efficacy of large language models in health literacy: a comprehensive cross-sectional study. *Yale J Biol Med*. 2024;97:17–27.
7. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183:589–596.

8. Chen S, Kann BH, Foote MB, et al. Use of artificial intelligence chatbots for cancer treatment information. *JAMA Oncol*. 2023;9:1459−1462.
9. Bernstein IA, Zhang YV, Govil D, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA Netw Open*. 2023;6:e2330320.
10. 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.
11. Betzler BK, Chen H, Cheng CY, et al. Large language models and their impact in ophthalmology. *Lancet Digit Health*. 2023;5:e917−e924.
12. OpenAI. Introducing OpenAI o1-preview 2024. https://openai.com/index/introducing-openai-o1-preview/. Accessed September 17, 2024.
13. OpenAI. Hello GPT-4o 2024. https://openai.com/index/hello-gpt-4o/. Accessed May 13, 2024.
14. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine*. 2023;95:104770.
15. Pushpanathan K, Lim ZW, Er Yew SM, et al. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. *iScience*. 2023;26:108163.
16. Cheong KX, Zhang C, Tan TE, et al. Comparing generative and retrieval-based chatbots in answering patient questions regarding age-related macular degeneration and diabetic retinopathy. *Br J Ophthalmol*. 2024;108:1443−1449.
17. OpenAI. How do I use ChatGPT browse with bing to search the web? 2024. https://help.openai.com/en/articles/8077698-how-do-i-use-chatgpt-browse-with-bing-to-search-the-web. Accessed October 4, 2024.
18. Yam JC, Zhang XJ, Zhang Y, et al. Effect of low-concentration atropine eyedrops vs placebo on myopia incidence in children: the LAMP2 randomized clinical trial. *JAMA*. 2023;329:472−481.