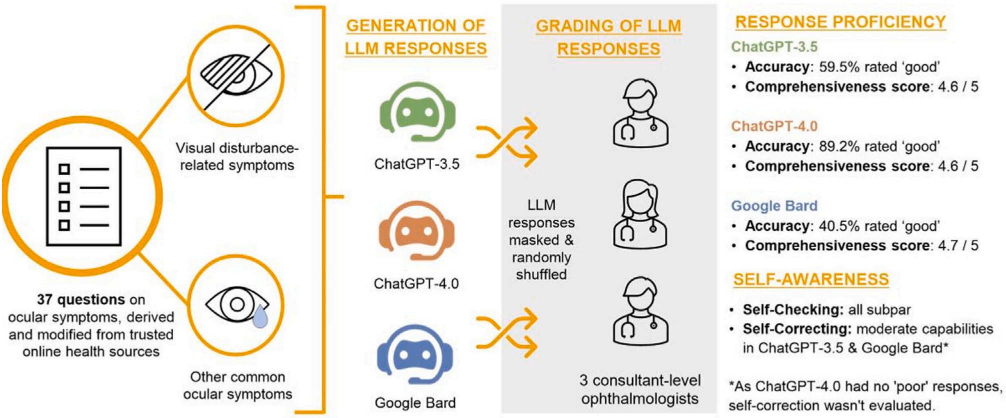
iScience

CellPress
OPEN ACCESS

Article

Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries


37 questions on ocular symptoms, derived and modified from trusted online health sources
Visual disturbance-related symptoms
Other common ocular symptoms
GENERATION OF LLM RESPONSES
ChatGPT-3.5
ChatGPT-4.0
Google Bard
GRADING OF LLM RESPONSES
LLM responses masked & randomly shuffled
3 consultant-level ophthalmologists
RESPONSE PROFICIENCY
ChatGPT-3.5
• Accuracy: 59.5% rated 'good'
• Comprehensiveness score: 4.6 / 5
ChatGPT-4.0
• Accuracy: 89.2% rated 'good'
• Comprehensiveness score: 4.6 / 5
Google Bard
• Accuracy: 40.5% rated 'good'
• Comprehensiveness score: 4.7 / 5
SELF-AWARENESS
• Self-Checking: all subpar
• Self-Correcting: moderate capabilities in ChatGPT-3.5 & Google Bard*
*As ChatGPT-4.0 had no 'poor' responses, self-correction wasn't evaluated.

Krithi Pushpanathan, Zhi Wei Lim, Samantha Min Er Yew, ..., Marcus Chun Jin Tan, Victor Teck Chang Koh, Yih-Chung Tham

thamyc@nus.edu.sg

Highlights

ChatGPT-4.0 has higher accuracy in addressing ocular symptom queries

ChatGPT-3.5, ChatGPT-4.0, and Google Bard's responses are equally comprehensive

All three chatbots exhibited differing self-checking abilities, none being adequate

ChatGPT-3.5 and Google Bard demonstrated moderate self-correcting capabilities

Pushpanathan et al., iScience *26*, 108163
November 17, 2023 © 2023 The Authors.
https://doi.org/10.1016/j.isci.2023.108163

iScience

CellPress OPEN ACCESS

Article

Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries

Krithi Pushpanathan,[1,2,8] Zhi Wei Lim,[1,8] Samantha Min Er Yew,[1,2] David Ziyou Chen,[1,2,3] Hazel Anne Hui'En Lin,[1,2,3] Jocelyn Hui Lin Goh,[4] Wendy Meihua Wong,[1,2,3] Xiaofei Wang,[5,6] Marcus Chun Jin Tan,[1,2,3] Victor Teck Chang Koh,[1,2,3] and Yih-Chung Tham[1,2,4,7,9,*]

SUMMARY

In light of growing interest in using emerging large language models (LLMs) for self-diagnosis, we systematically assessed the performance of ChatGPT-3.5, ChatGPT-4.0, and Google Bard in delivering proficient responses to 37 common inquiries regarding ocular symptoms. Responses were masked, randomly shuffled, and then graded by three consultant-level ophthalmologists for accuracy (poor, borderline, good) and comprehensiveness. Additionally, we evaluated the self-awareness capabilities (ability to self-check and self-correct) of the LLM-Chatbots. 89.2% of ChatGPT-4.0 responses were 'good'-rated, outperforming ChatGPT-3.5 (59.5%) and Google Bard (40.5%) significantly (all $p < 0.001$). All three LLM-Chatbots showed optimal mean comprehensiveness scores as well (ranging from 4.6 to 4.7 out of 5). However, they exhibited subpar to moderate self-awareness capabilities. Our study underscores the potential of ChatGPT-4.0 in delivering accurate and comprehensive responses to ocular symptom inquiries. Future rigorous validation of their performance is crucial to ensure their reliability and appropriateness for actual clinical use.

INTRODUCTION

The advent of publicly available large language models (LLMs) marked a transformative shift in medicine, ushering in a new era of possibilities. Leveraging deep-learning techniques and vast repositories of data from diverse sources, LLMs excel in generating contextually relevant responses across diverse prompts.[1,2] The user-friendly interfaces of LLMs have propelled their popularity and facilitated their extensive adoption across diverse healthcare settings.[3] Their versatility extends to various clinical applications,[4] including supporting clinical decision-making,[5–7] generating medical documentation,[8–10] and assisting in diagnosis.[11–16]

Since the inception of the internet, the reliance on online sources for self-triage and diagnosis has been pervasive.[17–20] However, the integration of LLMs introduces a new dimension to this practice, potentially simplifying and enhancing the accessibility of such information. Indeed, despite ChatGPT, an LLM, only becoming publicly accessible in November 2022, a survey revealed that 78% of respondents were inclined to employ it for self-diagnosis purposes.[21] This trend is likely to endure in the field of ophthalmology, as patients increasingly seek information about ocular symptoms through online platforms.

Remarkably, early research evaluating LLMs' test-taking abilities has demonstrated promising results across various examinations of diverse complexities. This ranges from entry-level standardized medical admissions tests like MCAT and BMAT,[22,23] general medical examinations such as USMLE,[24] to ophthalmology specialist licensing examinations such as OKAP and FRCOphth examination.[25,26] Nonetheless, the suitability of responses generated by LLM-Chatbots in addressing inquiries related to ocular symptoms remains uncertain. While Tsui et al. have explored this area, they focused on a single LLM-Chatbot (ChatGPT-3.5), and a restricted set of 10 questions, without extensively examining potential misinformation conveyed by the LLM-Chatbot.[27] Notably, Chatbots have limited ability to critically appraise the veracity and reliability of extracted information and may thus generate a misleading response. Unfortunately, the often-sophisticated response can give an illusion of accuracy, with the reader imbibing the flawed content.[28–30]

[1] Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[2] Centre for Innovation and Precision Eye Health & Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[3] Department of Ophthalmology, National University Hospital, Singapore, Singapore
[4] Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore
[5] Key Laboratory for Biomechanics and Mechanobiology of Ministry of Education, Beijing, China
[6] Advanced Innovation Centre for Biomedical Engineering, School of Biological Science and Medical Engineering, Beihang University, Beijing, China
[7] Ophthalmology and Visual Sciences Academic Clinical Programme (Eye ACP), Duke NUS Medical School, Singapore, Singapore
[8] These authors contributed equally to this work
[9] Lead contact
*Correspondence: thamyc@nus.edu.sg

https://doi.org/10.1016/j.isci.2023.108163

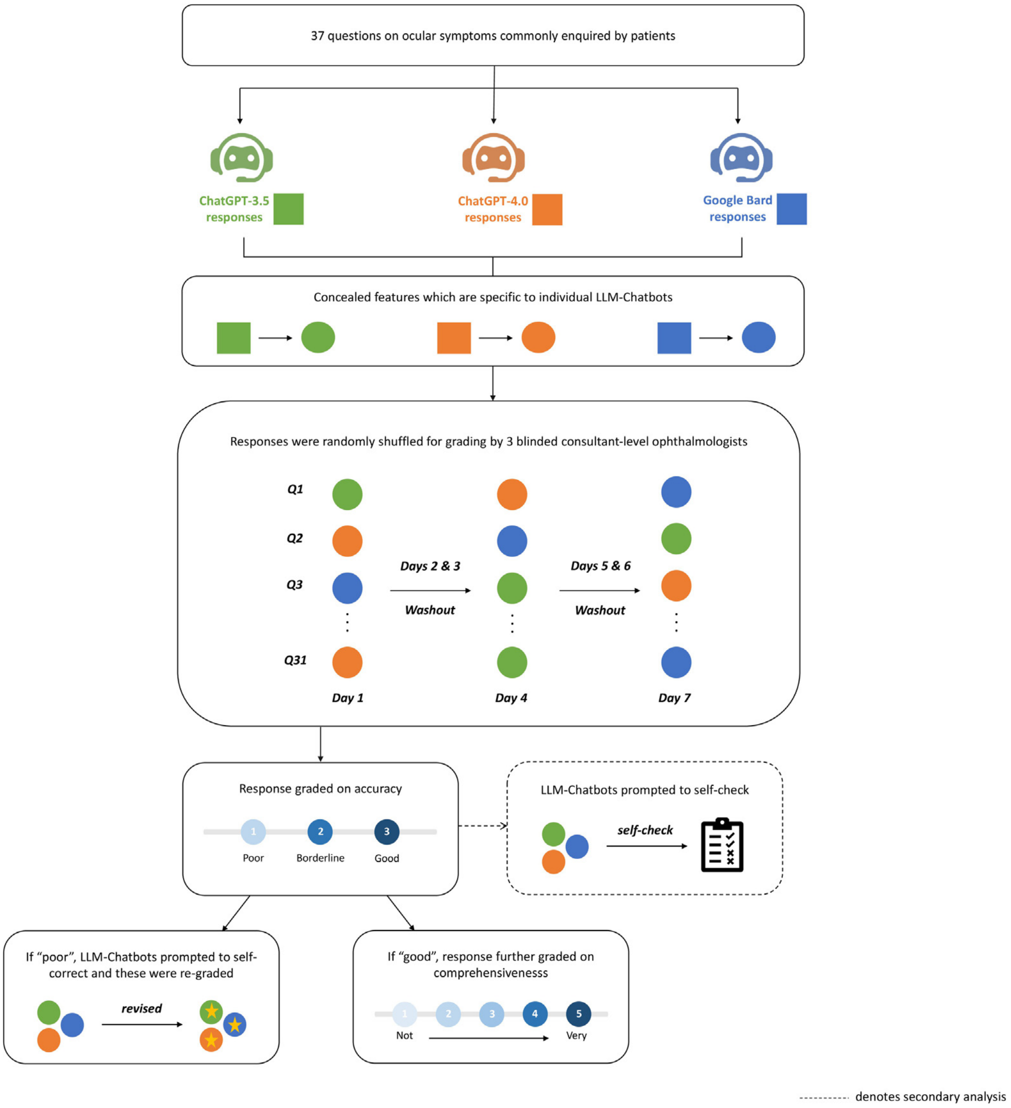

37 questions on ocular symptoms commonly enquired by patients
ChatGPT-3.5 responses
ChatGPT-4.0 responses
Google Bard responses
Concealed features which are specific to individual LLM-Chatbots
Responses were randomly shuffled for grading by 3 blinded consultant-level ophthalmologists
Q1
Q2
Q3
Q31
Days 2 & 3
Washout
Days 5 & 6
Washout
Day 1
Day 4
Day 7
Response graded on accuracy
1
2
3
Poor
Borderline
Good
LLM-Chatbots prompted to self-check
self-check
If "poor", LLM-Chatbots prompted to self-correct and these were re-graded
revised
If "good", response further graded on comprehensivenesss
1
2
3
4
5
Not
Very
denotes secondary analysis

Figure 1. Flowchart of overall study design

In this study, we aimed to evaluate the proficiency of three prominent and publicly available LLMs in addressing queries related to ocular symptoms. Our focus was on OpenAI's ChatGPT-3.5 and GPT-4.0, as well as Google's Bard. We pursued three primary areas in this evaluation. First, we assessed the accuracy and comprehensiveness of their response. Second, we explored their self-awareness by evaluating their self-checking and self-correcting capabilities. Lastly, we identified any instances of misinformation conveyed by these LLM-Chatbots. Findings from this study may provide invaluable insights into the efficacy and potential pitfalls of utilizing LLM-Chatbots to address common inquiries about ocular symptoms.

Methodology

Study design

The study took place between May 11th 2023 and July 8th 2023 at the Ophthalmology Department at National University Hospital, National University Health System (NUHS), Singapore.

Figure 1 illustrates the study design. To comprehensively evaluate the capabilities of LLM-Chatbots across a diverse spectrum of ocular symptoms, a team of experienced ophthalmologists (DZC, MCJT, HAHL) and clinical optometrists (SMEY, YCT) collaborated to curate 37 questions on ocular symptoms. After gathering common queries from reputable online health information sites (e.g., National Eye Institute,[31] American Academy of Ophthalmology[32]), the expert panel further refined these questions, selecting those commonly encountered in a clinical setting, ensuring the relevance and practicality of the inquiries. These questions were categorized into two broad categories: visual disturbance-related symptoms (including acute vision loss, chronic vision loss, diplopia, metamorphopsia, photopsia, myodesopsia, visual field defect, photophobia, and glare sensitivity/haloes), and other common ocular symptoms (including red eye, painful eye, pruritus, pressure sensation, epiphora, ptosis, blepharospasm, foreign body sensation, trauma, and asthenopia)

The queries were then individually posed to ChatGPT-3.5, ChatGPT-4.0 and Google Bard between May 18 to July 4, 2023. To mitigate potential memory bias from the LLM-Chatbots, the conversation was reset after each prompt. Any attributes specific to the LLM-Chatbots' answering format were then removed from the final generated responses (Tables S1, S2, and S3).

Subsequently, the anonymized responses from the three Chatbots were randomly shuffled and then presented as three 'randomized sets' to each of the three assigned ophthalmologists (each possessing ≥8 years of clinical ophthalmology experience), for independent grading. The grading was conducted over three separate rounds, with each round dedicated to one set. To minimize recency bias, a 48-h wash-out interval was implemented between the assessment of each round.

Accuracy evaluation

The three assigned graders independently assessed the accuracy of every response generated by the respective LLM-Chatbots. We employed a three-tier grading system as follows: (1) 'Poor' for responses that contained significant inaccuracies capable of misleading patients and potentially causing harm; (2) 'Borderline' for responses that might have had factual errors but were unlikely to mislead or harm patients; (3) 'Good' for responses that were free of errors. Final accuracy ratings for each LLM-Chatbot response were determined using a majority consensus approach amongst the three graders. When consensus was not reached, with each grader providing a different rating, we adopted a rigorous approach by assigning the lowest score (i.e., 'poor') to the LLM-Chatbot response.

The total accuracy score (continuous measure) for each LLM-Chatbot response was determined by summing the scores assigned by the three graders (Tables S4 and S5).

Comprehensiveness evaluation

We conducted an additional assessment on LLM-Chatbot responses that received a 'good' rating through majority consensus to evaluate their degree of comprehensiveness. This involved the use of a five-point scale, encompassing the following categories: (1) 'Not comprehensive' for responses lacking substantial details; (2) 'Slightly comprehensive' for responses with minimal but essential information; (3) 'Moderately comprehensive' for responses that presented a reasonable level of detail; (4) 'Comprehensive' for responses addressing most critical elements; (5) 'Very comprehensive' for responses presenting thorough and detailed information. The overall mean comprehensiveness score was determined by averaging the scores given by each grader across the total number of 'good' rated responses.

Evaluation of self-awareness levels in LLM-Chatbots

The term "self-awareness" in this context refers to the ability of these LLM models to self-check and self-correct their responses. To evaluate self-checking, we prompted the LLM-Chatbots with this question – "Could you kindly check if your answer is correct?" after each response. This prompt was deliberately designed to be generic and not directly point out possible errors, so as to provide a more accurate assessment of the LLM-Chatbots' capacity to self-check.

On the other hand, to assess the LLM-Chatbots' self-correcting capabilities, responses initially rated as 'poor' were further prompted with this line, "That does not seem quite right. Could you kindly review?". The regenerated responses upon self-correction, were re-assessed by the three graders one week after the initial grading. To minimize bias in grading for this segment, the graders were not informed that these responses had been self-corrected and remained blinded to the original 'poor'-rated responses.

Detailed qualitative analysis of poorly-rated LLM-Chatbot responses

To further shed light on the potential limitations and risks of relying on LLM-Chatbots for answers regarding ocular symptoms, a designated expert (DZC) further identified and emphasized erroneous or inaccurate sentences present in 'poor'-rated responses. Furthermore, explanations were also provided to elucidate the nature of these inaccuracies.

Statistical analysis

Statistical analyses were performed using R (Version 4.1.1, R Foundation, Vienna, Austria). To assess differences in character count, word count, total accuracy scores, and comprehensiveness scores among responses from the three LLM-Chatbots, we performed the Kruskal-Wallis rank-sum test and Dunn's multiple comparison post-hoc test (the data did not meet parametric assumptions). To compare the proportions of 'good', 'borderline', and 'poor' ratings among the LLM-Chatbots, we performed the two-tailed Pearson's $\chi 2$ test.

p-values were adjusted using the Bonferroni correction method to account for multiple hypothesis tests. Statistical significance was considered at $p < 0.05$.

RESULTS

Response length evaluation

Table 1 summarizes the response lengths of the LLM-Chatbots to the 37 questions related to ocular symptoms. The mean ± standard deviation (SD) of the word count was 184.0 ± 32.4 for ChatGPT-3.5, 234.8 ± 38.3 for ChatGPT-4.0, and 298.1 ± 98.9 for Google Bard. Similarly,

Table 1. Overview of response length from LLM-Chatbots to ocular symptoms queries

	Response length (words)			Response length (characters)		
LLM	Mean (SD)	Minimum	Maximum	Mean (SD)	Minimum	Maximum
ChatGPT-3.5	184.0 (32.4)	121	251	990.4 (171.5)	667	1383
ChatGPT-4.0	234.8 (38.3)	151	320	1305.9 (197.5)	965	1753
Google Bard	298.1 (98.9)	91	510	1533.8 (513.4)	489	2644

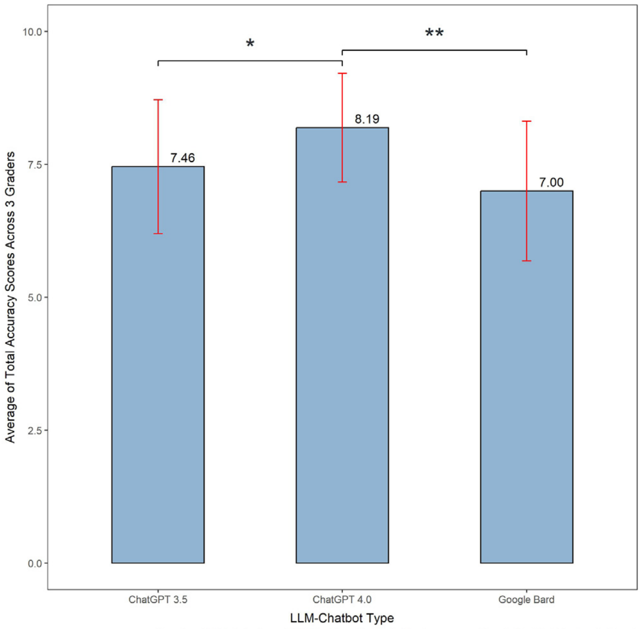
the mean ± standard deviation (SD) of the character count was 990.4 ± 171.5 for ChatGPT-3.5, 1305.9 ± 197.5 for ChatGPT-4.0, and 1533.8 ± 513.4 for Google Bard.

Accuracy evaluation

Figure 2 presents the average total accuracy scores of LLM-Chatbots' responses to questions related to ocular symptoms, evaluated by the three ophthalmologists. ChatGPT-4.0 exhibited a notably higher average total accuracy score of 8.19 ± 1.02, surpassing both ChatGPT-3.5 (7.46 ± 1.26; Dunn's post-hoc test, $p = 0.019$) and Google Bard (7.00 ± 1.31; Dunn's post-hoc test, $p < 0.001$). The total accuracy score and details for each LLM-Chatbot's response can be found in Tables S4 and S5.

Figure 3 illustrates the experts' consensus-based accuracy ratings of the responses by the three LLM-Chatbots. ChatGPT-4.0's responses achieved 89.2% of 'good' ratings, significantly higher than the proportion of 'good' ratings in ChatGPT-3.5's (59.5%, Pearson's chi-squared test, $p < 0.001$) and Google Bard's responses (40.5%, Pearson's chi-squared test, $p < 0.001$). Notably, ChatGPT-4.0 did not receive any 'poor' rating in any of its responses. When comparing ChatGPT-3.5 and Google Bard, ChatGPT-3.5 had lower proportions of 'borderline' (35.1% vs. 45.9%) and 'poor' (5.4% vs. 13.5%) responses compared to Google Bard (Pearson's chi-squared test, $p = 0.01$). Table S4 presents a detailed evaluation of each LLM-Chatbot's response to questions pertaining to visual disturbance symptoms, while Table S5 shows their performance in addressing other ocular symptoms.

Table 2 provides a detailed sub-analysis of the experts' consensus-based accuracy ratings across the two main question categories, visual and other ocular symptoms. Notably, ChatGPT-4.0 demonstrated superior performance compared to its counterparts in questions related to visual disturbance symptoms, receiving 77.8% 'good' ratings in this category. In comparison, ChatGPT-3.5 achieved 50.0%, and Google Bard achieved 22.2% (Pearson's chi-squared test, all $p < 0.001$). Interestingly, all three LLM-Chatbots displayed better performance when addressing questions related to other ocular symptoms. ChatGPT-4.0 achieved a perfect 100% 'good' ratings in this category, outperforming ChatGPT-3.5 and Google Bard, which received 68.4% (Pearson's chi-squared test, $p < 0.001$) and 57.9% (Pearson's chi-squared test, $p < 0.001$) of 'good' ratings respectively.



* denotes p=0.019; ** denotes p<0.001 (Dunn's post-hoc test). Error bars represent the standard deviation of each data set.

Figure 2. Average Total Accuracy Scores of LLM-Chatbot Responses to Ocular Symptoms-Related Questions, as Assessed by Three Consultant-Level Ophthalmologists

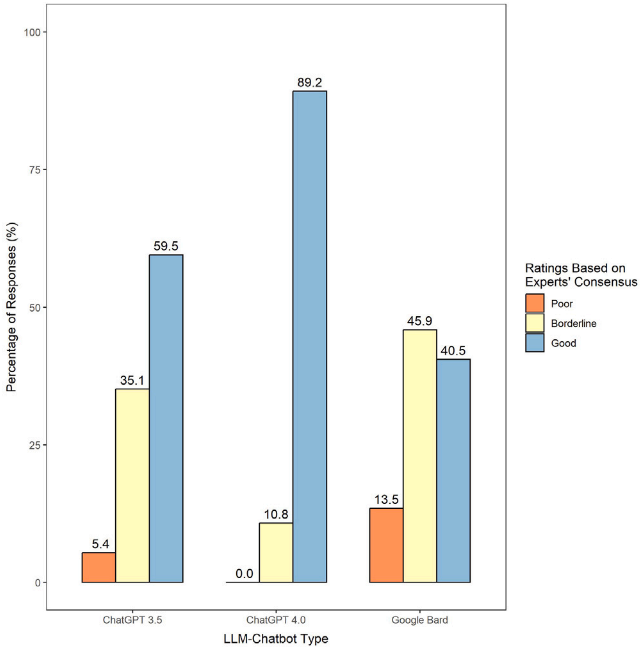


Figure 3. Consensus-Based Accuracy Ratings of LLM-Chatbot Responses to Ocular Symptoms-Related Questions, as Determined by Three Consultant-Level Ophthalmologists

Comprehensiveness evaluation

Table 3 provides a summary of the comprehensiveness scores for 'good' rated responses. Remarkably, all three LLM-Chatbots exhibited exemplary performance in this regard. ChatGPT-3.5 achieved an impressive overall mean comprehensive score of 4.6, ChatGPT-4.0 scored 4.6, and Google Bard obtained a score of 4.7, out of a maximum possible rating of 5. Additionally, when comparing the comprehensiveness scores of the three LLM-Chatbots based on a common set of questions (Table 4), similar performance was observed, with no statistically significant difference detected across the three LLM-Chatbots (Kruskal-Wallis rank-sum test, p = 0.703).

Evaluation of self-awareness levels in LLM-Chatbots

Tables S6–S8 present representative examples of each grade ('good', 'borderline', and 'poor') for the self-checking abilities of ChatGPT-3.5, ChatGPT-4.0, and Google Bard, respectively. ChatGPT-3.5, irrespective of the accuracy grades of the original responses, refrained from either revising its initial responses or reaffirming its correctness. Instead, it issued a general disclaimer, emphasizing the need for additional personal medical information to generate more accurate responses when prompted to self-check. Interestingly, this disclaimer was not originally present in its initial responses. On the other hand, both ChatGPT-4.0 and Google Bard consistently asserted the accuracy of their original responses, even when they were deemed as 'poor' or 'borderline' by our expert graders.

Tables 5 and 6 demonstrate the response adjustments made by the LLM-Chatbots on original 'poor'-rated responses when prompted for self-correction. Both ChatGPT-3.5 and Google Bard demonstrated improvements through self-correction prompts. ChatGPT-3.5 enhanced 2 (out of 2, 100%) of its initial responses, while Google Bard improved 4 responses (out of 5, 80%) after self-correction. However, most of these

Table 2. Consensus-based accuracy ratings of LLM-Chatbot responses to ocular symptoms-related questions

	No. of	ChatGPT-3.5, n (%)			ChatGPT-4.0, n (%)			Google bard, n (%)		
Domain[a]	Questions	Poor	Borderline	Good	Poor	Borderline	Good	Poor	Borderline	Good
Visual Disturbance Symptoms[b]	18	1 (5.6)	8 (44.4)	9 (50.0)	0 (0)	4 (22.2)	14 (77.8)	3 (16.7)	11 (61.1)	4 (22.2)
Other Ocular Symptoms[c]	19	1 (5.3)	5 (26.3)	13 (68.4)	0 (0)	0 (0)	19 (100)	2 (10.5)	6 (31.6)	11 (57.9)

[a] Based on majority consensus across the three graders.
[b] Includes acute vision loss, chronic vision loss, diplopia, metamorphopsia, photopsia, myodesopsia, visual field defect, photophobia, and glare sensitivity/haloes.
[c] Includes red eye, painful eye, pruritus, pressure sensation, epiphora, ptosis, blepharospasm, foreign body sensation, trauma, and asthenopia

Table 3. Comprehensiveness assessment for all LLM-Chatbot responses that received a 'good' accuracy rating

	Response comprehensiveness		
LLM[a]	n	Mean (SD)	Median
ChatGPT-3.5	22	4.6 (0.3)	4.5
ChatGPT-4.0	33	4.6 (0.4)	4.7
Google Bard	15	4.7 (0.2)	4.7

[a]Based on majority consensus across the three graders.

improvements were a marginal progression from a 'poor' rating to 'borderline'; in each LLM-Chatbot, only one response transitioned from 'poor' to 'good'. As none of ChatGPT-4.0's original responses were rated 'poor', its self-correction capabilities were not assessed. Tables S9 and S10 provide comprehensive details on the original responses and their respective self-corrected versions for ChatGPT-3.5 and Google Bard.

Analysis of misinformation in poorly-rated LLM-Chatbot responses

Tables S11 and S12 show examples of misinformation conveyed by the LLM-Chatbots, specifically focusing on the 'poor'-rated responses by ChatGPT-3.5 and Google Bard. The segments highlighted in red within these tables denote areas where our expert graders identified inaccuracies. Drawing insights from two experienced ophthalmologists (DZC, HAHL), further explanations for these identified errors are also included.

DISCUSSION

Our study provides a rigorous evaluation of ChatGPT-3.5, ChatGPT-4.0, and Google Bard in addressing patient queries about ocular symptoms. Through the curation of frequently posed questions by patients, we effectively simulated a series of scenarios where individuals may seek medical information and advice from these models. This approach diverges from previous literature that predominantly focused on the performance of LLM-Chatbots in the context of medical licensing or academic examinations.[25,26,33] Notably, our findings highlighted the potential of LLM-Chatbots, particularly ChatGPT-4.0, in delivering accurate and comprehensive responses to queries related to both visual disturbance and other types of ocular symptoms. However, our investigation also revealed subpar self-checking and moderate self-correction performances of the LLM-Chatbots, underscoring the need for further improvements in these areas. To ensure the validity of our findings, we adopted a robust research framework that included stringent masking procedures, random shuffling of generated responses during grading, and grounding our ground truth in the expert evaluations of three consultant-level ophthalmologists. Overall, our study offers unique insights into the potential utility and limitations of LLM-Chatbots in addressing common ocular symptom queries.

Of the three LLM-Chatbots assessed, ChatGPT-4.0 demonstrated superior performance in addressing queries associated with ocular symptoms. It achieved the highest average accuracy score and garnered significantly greater proportions of 'good' ratings as compared to its counterparts (Figures 2 and 3). Our findings corroborate previous studies which highlighted ChatGPT-4.0's superior performance over ChatGPT-3.5 and Google Bard in providing consistently accurate and contextually relevant answers within the medical domain.[25,34] ChatGPT-4.0's exceptional performance likely arises from its unique attributes, including an extensive parameter set and enhanced proficiency in handling intricate prompts.[35–37] With a substantial user base contributing to reinforcement learning, and the utilization of more recent reinforcement learning data, the tendency to generate false information compared to its predecessor, ChatGPT-3.5, is greatly reduced.[35–37] Notably, however, all three LLM-Chatbots exhibited comparable expertise in delivering comprehensive responses (Tables 3 and 4). Table S13 exemplifies a scenario where all three LLM-Chatbots scored perfectly for comprehensiveness when responding to the query " What should I do if I am experiencing double vision?", further attesting to the LLM-Chatbots' potential abilities to offer detailed information that addresses all aspects of the inquiry.

Table 4. Comprehensiveness assessment for common questions answered by the three LLM-Chatbots, with responses that received a 'good' accuracy rating

	Response Comprehensiveness		
LLM[a]	n	Mean (SD)	Median
ChatGPT-3.5	11	4.7 (0.3)	4.7
ChatGPT-4.0	11	4.7 (0.3)	4.7
Google Bard	11	4.7 (0.2)	4.7

[a]Based on majority consensus across the three graders.

Table 5. Demonstration of ChatGPT-3.5's ability to self-correct when prompted

		Summed score		Consensus-based rating	
Topic	Question	Initial	Self-corrected	Initial	Self-corrected
Metamorphopsia	1. Why do straight lines appear wavy to me?	4	6	Poor	Borderline
Red Eye	1. Why is my eye red?	6	8	Poor[a]	Good

[a]Where consensus on final accuracy rating was not reached (i.e., each grader provided a different rating), the lowest score ('poor') was assigned.

Interestingly, all three LLM-Chatbots exhibited better performance when addressing queries pertaining to 'other ocular symptoms', compared to those concerning visual disturbance symptoms (Table 2). This performance disparity could be attributed to the fact that many eye diseases share similar visual disturbance symptoms (e.g., acute or chronic vision loss), making these symptoms less disease-specific and more complex to address accurately. On the other hand, 'other ocular symptoms' (e.g., red eye, painful eye) are typically more specific to certain eye conditions, and therefore, may be less ambiguous. For instance, red eye is generally associated with diseases located in the anterior part of the eye, whereas visual loss can result from conditions affecting both the anterior and posterior parts of the eye. Consequently, the complexity and need for additional details to accurately address queries related to visual disturbance symptoms may have challenged the LLM-Chatbots' ability to deliver precise responses, leading to their relatively poorer performance in this category of questions.

The LLM-Chatbots exhibited subpar to moderate self-awareness capabilities. When prompted for self-checking, ChatGPT-3.5 typically responded by acknowledging its inability to verify accuracy due to the unavailability of "personal medical information", rather than making iterative improvements to its responses. While the disclaimer is modest, it effectively also highlights the lack of fact-checking, offering minimal assistance for users seeking to verify the information (see examples in Table S6). Meanwhile, both ChatGPT-4.0 and Google Bard consistently assert the accuracy of their information, even when responses were deemed inaccurate by our expert graders. This suggests a potential risk of 'convincingly' providing incorrect information (see examples in Tables S7 and S8). Our findings are consistent with previous studies which documented similar instances of LLM-Chatbots lacking effective internal verification capabilities.[38,39]

When evaluating self-correcting capabilities, ChatGPT-3.5 and Google Bard exhibited a moderate level of proficiency. However, among the 6 (out of 7) improved responses across both models, 4 showed only marginal improvement, transitioning from 'poor' to 'borderline'. This raises concerns about the persistence of misinformation, even when potential errors were already highlighted and prompted (Tables S9 and S10). While continuous refinement and user feedback may lead to improvement in the self-correcting capabilities of LLM-Chatbots over time, it is crucial to remain vigilant as these models might inadvertently incorporate and propagate errors or biases that were inherent in their training data.[40,41]

During our study, we closely tracked significant updates to Google Bard and assessed their impact on its performance. To this end, we compared the 'poor'-rated responses generated by Google Bard at the outset of our study and two months later (July 13, 2023).[42] All these original 'poor' responses improved marginally to 'borderline' (Table S14), indicating that the responses still harbor some form of errors despite the recent model and system update.

We conducted a detailed qualitative assessment on responses that were initially rated as "poor", carried out by two experienced ophthalmologists (DZC and HAHL). We identified the erroneous sections within each response and provided further explanations. Overall, the "poor" ratings were primarily attributed to two key factors: inaccuracies in the provided responses and omission of crucial information. Despite these shortcomings, it is noteworthy that both ChatGPT-3.5 and Google Bard consistently advised users to "see an eye doctor", demonstrating a level of caution in their responses. (Tables S11 and S12). We present two representative examples that showcase incomplete answers and inaccurate information. When ChatGPT-3.5 was asked, "Why is my eye red?", it offered a structured list of eye conditions in a seemingly confident manner but omitted sight-threatening causes such as keratitis and acute angle closure glaucoma, which necessitate prompt attention from an ophthalmologist (Table S11). ChatGPT-3.5 also overlooked other potential causes of red eye such as sub-conjunctival hemorrhage,

Table 6. Demonstration of Google Bard's ability to self-correct when prompted

		Summed score		Consensus-based rating	
Topic	Question	Initial	Self-corrected	Initial	Self-corrected
Diplopia	1. Why am I having double vision?	5	7	Poor	Borderline
Visual Field Defect	1. Why is there something blocking my vision?	6	6	Poor[a]	Borderline
	2. What should I do if I notice that something is blocking my vision?	6	6	Poor[a]	Borderline
Painful Eye	2. What should I do if I am having a painful eye?	6	8	Poor[a]	Good
Pressure Sensation	2. What should I do if I am feeling pressure in my eye?	6	6	Poor[a]	Poor[a]

[a]Where consensus on final accuracy rating was not reached (i.e., each grader provided a different rating), the lowest score ('poor') was assigned.

blepharitis, uveitis, and scleritis. In another instance, when Google Bard was queried, ''Why am I having double vision?'', the Chatbot adeptly categorised potential causes into five primary sectors (eye muscle, brain problems, eye diseases, medications, trauma). However, it failed to mention another common cause – uncorrected refractive error and astigmatism (Table S12). Moreover, Google Bard incorrectly suggested that cataracts, glaucoma, and retinal detachment cause diplopia by "damaging the nerves controlling eye movement." This misinformation could cause undue anxiety to patients. Importantly, some of these misrepresentations and omission of important information may delay timely intervention. Furthermore, through this qualitative assessment, we also noticed that LLM-Chatbots are generally prompt-dependent.[43] They appeared to face challenges with vague or ambiguous prompts (e.g., ''Why is my eye red?''), leading to responses that were not thorough and occasionally deviated from the desired information. Therefore, to mitigate the risk of misinformation or hallucinations, specially tailored prompts may be used to guide the LLM-Chatbots in their responses or to halt them from answering complex questions that are beyond their current capabilities.

Our findings underscore the potential value of LLM-Chatbots, particularly ChatGPT-4.0, for answering queries and disseminating ocular symptom-related information. In the dynamic global healthcare landscape, healthcare providers are grappling with multifaceted challenges. The pressures on healthcare professionals have reached unprecedented levels due to the ongoing global health crises, amplified patient loads, and increasing diversity in patient populations.[44,45] In this regard, LLMs can potentially alleviate burdens by enabling remote and real-time patient triage,[46] reducing wait times,[47] and improving access to care, particularly for individuals in remote or underserved regions.[47,48] They can also optimize consultations by handling administrative tasks, allowing healthcare professionals to focus on direct patient care.[49,50] Given their multilingual capabilities and cultural adaptability, LLMs may prove to be useful in advancing health access and equity.[2] The availability of Application Programming Interface (API) access[51] further accelerates the adoption of LLMs by enabling the integration of their sophisticated natural language processing into various online platforms. However, it should be noted that LLM-Chatbots still exhibit limitations in their medical acumen when compared to expert physicians. In some instances, these LLM-Chatbots conveyed inaccurate information, posing risks of misdiagnoses and suboptimal triage. Additionally, patient reservations about privacy, and potential algorithmic issues that could amplify social, racial, or cultural disparities need to be addressed comprehensively.[52] Until these challenges are adequately tackled, the clinical deployment of LLM-Chatbots carries inherent risks.

Our study's strengths lie in its rigorous design, incorporating several safeguards such as concealing LLM-specific traits in generated responses, randomizing the order of responses during grading, and implementing wash-out periods between grading sessions. These methodological considerations were implemented to mitigate grader biases and enhance the credibility of our findings.

Limitations of the study

This study has several limitations. Firstly, subjectivity may arise among individual graders when assigning accuracy and comprehensiveness scores to LLM-Chatbot responses. However, we addressed this concern by enlisting three highly experienced consultant-level ophthalmologists (with at least 8 years of clinical expertise) and employing a majority consensus-based rating approach. Second, each symptom-specific domain consisted of a maximum of only two questions, which made the analysis of domain-specific LLM-Chatbot performance less meaningful. It would have been insightful to assess whether the LLM-Chatbots demonstrated higher proficiency in certain domains compared to others. Finally, given the constant updates to LLM-Chatbots, the results reported here should be interpreted with caution. The time-sensitive nature of these models demands robust evaluation and careful interpretation.

In conclusion, our thorough and rigorous evaluation offers unique insights into the varying degrees of accuracy and capabilities among three popular LLM-Chatbots. Our findings highlight the superior performance of ChatGPT-4.0 in addressing a broad spectrum of common ocular symptom-related queries. Continuous refinement of these LLM-Chatbots, coupled with rigorous validation and evaluation, remain crucial to ensure their reliability and appropriateness before they can be adopted for mainstream clinical use.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Accuracy evaluation
  - ○ Comprehensiveness evaluation
  - ○ Evaluation of self-awareness levels in LLM-Chatbots
  - ○ Detailed qualitative analysis of poorly-rated LLM-Chatbot responses
  - ○ Statistical analysis
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108163.

ACKNOWLEDGMENTS

Role of funders: The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

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

Funding: National Medical Research Council of Singapore.

AUTHOR CONTRIBUTIONS

Conceptualization, K.P., Z.W.L., S.M.E.Y., and Y-C.T.; Methodology, K.P., Z.W.L., S.M.E.Y., X.W., and Y-C.T.; Validation, K.P., S.M.E.Y., and Y-C.T.; Formal Analysis, K.P.; Investigation, K.P., Z.W.L., S.M.E.Y., and Y-C.T.; Data Curation, K.P., Z.W.L., S.M.E.Y., and Y-C.T.; Writing – Original Draft, K.P., Z.W.L., S.M.E.Y., and Y-C.T.; Writing – Review and Editing, K.P., Z.W.L., S.M.E.Y., D.Z.C., H.A.H.L., J.H.L.G., W.M.W., X.W., M.C.J.T., V.T.C.K., and Y-C.T.; Visualization, K.P.; Supervision, Y-C.T.; Funding Acquisition, Y-C.T.

DECLARATION OF INTERESTS

All authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT to edit and proofread the manuscript for improved readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Received: August 7, 2023
Revised: September 19, 2023
Accepted: October 5, 2023
Published: October 10, 2023

REFERENCES

1. De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G.P., Ferragina, P., Tozzi, A.E., and Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front. Public Health *11*, 1166120. https://doi.org/10.3389/fpubh.2023.1166120.
2. Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare *11*, 887. https://doi.org/10.3390/healthcare11060887.
3. Sedaghat, S. (2023). Success Through Simplicity: What Other Artificial Intelligence Applications in Medicine Should Learn from History and ChatGPT. Ann. Biomed. Eng. https://doi.org/10.1007/s10439-023-03287-x.
4. Sedaghat, S. (2023). Early applications of ChatGPT in medical practice, education and research. Clin. Med. *23*, 278–279. https://doi.org/10.7861/clinmed.2023-0078.
5. Kao, H.J., Chien, T.W., Wang, W.C., Chou, W., and Chow, J.C. (2023). Assessing ChatGPT's capacity for clinical decision support in pediatrics: A comparative study with pediatricians using KIDMAP of Rasch analysis. Medicine (Baltim.) *102*, e34068. https://doi.org/10.1097/md.0000000000034068.
6. Chen, J.H., Dhaliwal, G., and Yang, D. (2022). Decoding Artificial Intelligence to Achieve Diagnostic Excellence. JAMA *328*, 709–710. https://doi.org/10.1001/jama.2022.13735.
7. Haemmerli, J., Sveikata, L., Nouri, A., May, A., Egervari, K., Freyschlag, C., Lobrinus, J.A., Migliorini, D., Momjian, S., Sanda, N., et al. (2023). ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of a doctor in the tumour board? BMJ Health Care Inform. *30*, e100775. https://doi.org/10.1136/bmjhci-2023-100775.
8. Patel, S.B., and Lam, K. (2023). ChatGPT: the future of discharge summaries? Lancet. Digit. Health *5*, e107–e108. https://doi.org/10.1016/s2589-7500(23)00021-3.
9. Ting, D.S.J., Tan, T.F., and Ting, D.S.W. (2023). ChatGPT in ophthalmology: the dawn of a new era? Eye. https://doi.org/10.1038/s41433-023-02619-4.
10. Puthenpura, V., Nadkarni, S., Diluna, M., Hieftje, K., and Marks, A. (2023). Personality Changes and Staring Spells in a 12-Year-Old Child: A Case Report Incorporating ChatGPT, a Natural Language Processing Tool Driven by Artificial Intelligence (AI). Cureus *15*, e36408. https://doi.org/10.7759/cureus.36408.
11. Bilal, M., Jamil, Y., Rana, D., and Shah, H.H. (2023). Enhancing Awareness and Self-diagnosis of Obstructive Sleep Apnea Using AI-Powered Chatbots: The Role of ChatGPT in Revolutionizing Healthcare. Ann. Biomed. Eng. https://doi.org/10.1007/s10439-023-03298-8.
12. Hirosawa, T., Harada, Y., Yokose, M., Sakamoto, T., Kawamura, R., and Shimizu, T. (2023). Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. Int. J. Environ. Res. Publ. Health *20*, 3378. https://doi.org/10.3390/ijerph20043378.
13. Lahat, A., Shachar, E., Avidan, B., Glicksberg, B., and Klang, E. (2023). Evaluating the Utility of a Large Language Model in Answering Common Patients' Gastrointestinal Health-Related Questions: Are We There Yet? Diagnostics *13*, 1950. https://doi.org/10.3390/diagnostics13111950.
14. Cao, J.J., Kwon, D.H., Ghaziani, T.T., Kwo, P., Tse, G., Kesselman, A., Kamaya, A., and Tse, J.R. (2023). Accuracy of Information Provided by ChatGPT Regarding Liver Cancer Surveillance and Diagnosis. AJR Am. J. Roentgenol. *221*, 556–559. https://doi.org/10.2214/ajr.23.29493.
15. Rahsepar, A.A., Tavakoli, N., Kim, G.H.J., Hassani, C., Abtin, F., and Bedayat, A. (2023).

How AI Responds to Common Lung Cancer Questions (ChatGPT vs Google Bard).
16. Cheng, K., Li, Z., Guo, Q., Sun, Z., Wu, H., and Li, C. (2023). Emergency surgery in the era of artificial intelligence: ChatGPT could be the doctor's right-hand man. Int. J. Surg. *109*, 1816–1818. https://doi.org/10.1097/js9.0000000000000410.
17. Hochberg, I., Allon, R., and Yom-Tov, E. (2020). Assessment of the Frequency of Online Searches for Symptoms Before Diagnosis: Analysis of Archival Data. J. Med. Internet Res. *22*, e15065. https://doi.org/10.2196/15065.
18. Aoun, L., Lakkis, N., and Antoun, J. (2020). Prevalence and Outcomes of Web-Based Health Information Seeking for Acute Symptoms: Cross-Sectional Study. J. Med. Internet Res. *22*, e15148. https://doi.org/10.2196/15148.
19. Kwakernaak, J., Eekhof, J.A.H., De Waal, M.W.M., Barenbrug, E.A.M., and Chavannes, N.H. (2019). Patients' Use of the Internet to Find Reliable Medical Information About Minor Ailments: Vignette-Based Experimental Study. J. Med. Internet Res. *21*, e12278. https://doi.org/10.2196/12278.
20. Bujnowska-Fedak, M.M., and Węgierek, P. (2020). The Impact of Online Health Information on Patient Health Behaviours and Making Decisions Concerning Health. Int. J. Environ. Res. Publ. Health *17*, 880. https://doi.org/10.3390/ijerph17030880.
21. Shahsavar, Y., and Choudhury, A. (2023). User Intentions to Use ChatGPT for Self-Diagnosis and Health-Related Purposes: Cross-sectional Survey Study. JMIR Hum. Factors *10*, e47564. https://doi.org/10.2196/47564.
22. Bommineni, V., Bhagwagar, S., Balcarcel, D., Davatzikos, C., and Boyer, D. (2023). Performance of ChatGPT on the MCAT: The Road to Personalized and Equitable Premedical Learning. Preprint at medRxiv. https://doi.org/10.1101/2023.03.05.23286533.
23. Giannos, P., and Delardas, O. (2023). Performance of ChatGPT on UK Standardized Admission Tests: Insights From the BMAT, TMUA, LNAT, and TSA Examinations. JMIR Med. Educ. *9*, e47737. https://doi.org/10.2196/47737.
24. Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., and Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit. Health *2*, e0000198. https://doi.org/10.1371/journal.pdig.0000198.
25. Raimondi, R., Tzoumas, N., Salisbury, T., Di Simplicio, S., Romano, M.R., North East Trainee Research in Ophthalmology Network NETRiON, Chawla, H., Chen, Y., Connolly, S., El Omda, S., et al. (2023). Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. Eye. https://doi.org/10.1038/s41433-023-02563-3.
26. Antaki, F., Touma, S., Milad, D., El-Khoury, J., and Duval, R. (2023). Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. Ophthalmol. Sci. *3*, 100324. https://doi.org/10.1016/j.xops.2023.100324.
27. Tsui, J.C., Wong, M.B., Kim, B.J., Maguire, A.M., Scoles, D., Vanderbeek, B.L., and Brucker, A.J. (2023). Appropriateness of ophthalmic symptoms triage by a popular online artificial intelligence chatbot. Eye. https://doi.org/10.1038/s41433-023-02556-2.
28. Mello, M.M., and Guha, N. (2023). ChatGPT and Physicians' Malpractice Risk. JAMA Health Forum *4*, e231938. https://doi.org/10.1001/jamahealthforum.2023.1938.
29. Shen, Y., Heacock, L., Elias, J., Hentel, K.D., Reig, B., Shih, G., and Moy, L. (2023). ChatGPT and Other Large Language Models Are Double-edged Swords. Radiology *307*, e230163. https://doi.org/10.1148/radiol.230163.
30. Au Yeung, J., Kraljevic, Z., Luintel, A., Balston, A., Idowu, E., Dobson, R.J., and Teo, J.T. (2023). AI chatbots not yet ready for clinical use. Front. Digit. Health *5*, 1161098. https://doi.org/10.3389/fdgth.2023.1161098.
31. National Eye Institute; National Institutes of Health. Eye Conditions and Diseases. https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases.
32. American Academy of Ophthalmology. Eye Symptoms. https://www.aao.org/eye-health/symptoms-list.
33. Mihalache, A., Popovic, M.M., and Muni, R.H. (2023). Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. JAMA Ophthalmol. *141*, 589–597. https://doi.org/10.1001/jamaophthalmol.2023.1144.
34. Ali, R., Tang, O.Y., Connolly, I.D., Fridley, J.S., Shin, J.H., Zadnik Sullivan, P.L., Cielo, D., Oyelese, A.A., Doberstein, C.E., Telfeian, A.E., et al. (2023). Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. Neurosurgery. Publish Ahead of Print. https://doi.org/10.1227/neu.0000000000002551.
35. Hu, K. (2023). ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note (Reuters). https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.
36. OpenAI (2023). GPT-4 Technical Report.
37. Meskó, B., and Topol, E.J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit. Med. *6*, 120. https://doi.org/10.1038/s41746-023-00873-0.
38. Alberts, I.L., Mercolli, L., Pyka, T., Prenosil, G., Shi, K., Rominger, A., and Afshar-Oromieh, A. (2023). Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be? Eur. J. Nucl. Med. Mol. Imag. *50*, 1549–1552. https://doi.org/10.1007/s00259-023-06172-w.
39. Alkaissi, H., and Mcfarlane, S.I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus *15*, e35179. https://doi.org/10.7759/cureus.35179.
40. Meyer, J.G., Urbanowicz, R.J., Martin, P.C.N., O'Connor, K., Li, R., Peng, P.-C., Bright, T.J., Tatonetti, N., Won, K.J., Gonzalez-Hernandez, G., and Moore, J.H. (2023). ChatGPT and large language models in academia: opportunities and challenges. BioData Min. *16*, 20. https://doi.org/10.1186/s13040-023-00339-9.
41. Harrer, S. (2023). Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine *90*, 104512. https://doi.org/10.1016/j.ebiom.2023.104512.
42. Google (2023). Bard's Latest Update: More Features, Languages and Countries (Google).
43. Zhou, Y., Muresanu, A.I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. (2023). Large Language Models Are Human-Level Prompt Engineers.
44. Al Shamsi, H., Almutairi, A.G., Al Mashrafi, S., and Al Kalbani, T. (2020). Implications of Language Barriers for Healthcare: A Systematic Review. Oman Med. J. *35*, e122. https://doi.org/10.5001/omj.2020.40.
45. Lucero-Prisno, D.E., Shomuyiwa, D.O., Kouwenhoven, M.B.N., Dorji, T., Odey, G.O., Miranda, A.V., Ogunkola, I.O., Adebisi, Y.A., Huang, J., Xu, L., et al. (2023). Top 10 public health challenges to track in 2023: Shifting focus beyond a global pandemic. Public Health Challenges *2*. https://doi.org/10.1002/puh2.86.
46. Wang, J., Zhang, G., Wang, W., Zhang, K., and Sheng, Y. (2021). Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT. J. Cloud Comput. *10*, 4. https://doi.org/10.1186/s13677-020-00218-2.
47. Srivastav, S., Chandrakar, R., Gupta, S., Babhulkar, V., Agrawal, S., Jaiswal, A., Prasad, R., and Wanjari, M.B. (2023). ChatGPT in Radiology: The Advantages and Limitations of Artificial Intelligence for Medical Imaging Diagnosis. Cureus *15*, e41435. https://doi.org/10.7759/cureus.41435.
48. Sharma, S., Pajai, S., Prasad, R., Wanjari, M.B., Munjewar, P.K., Sharma, R., and Pathade, A. (2023). A Critical Review of ChatGPT as a Potential Substitute for Diabetes Educators. Cureus *15*, e38380. https://doi.org/10.7759/cureus.38380.
49. Zheng, Y., Wang, L., Feng, B., Zhao, A., and Wu, Y. (2023). Innovating Healthcare: The Role of ChatGPT in Streamlining Hospital Workflow in the Future. Ann. Biomed. Eng. https://doi.org/10.1007/s10439-023-03323-w.
50. Loh, E. (2023). ChatGPT and generative AI chatbots: challenges and opportunities for science, medicine and medical leaders. BMJ Lead. https://doi.org/10.1136/leader-2023-000797.
51. Niszczota, P., and Rybicka, I. (2023). The credibility of dietary advice formulated by ChatGPT: Robo-diets for people with food allergies. Nutrition *112*, 112076. https://doi.org/10.1016/j.nut.2023.112076.
52. Li, H., Moon, J.T., Purkayastha, S., Celi, L.A., Trivedi, H., and Gichoya, J.W. (2023). Ethics of large language models in medicine and medical research. Lancet. Digit. Health *5*, e333–e335. https://doi.org/10.1016/s2589-7500(23)00083-3.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
R (Version 4.1.1)	https://www.r-project.org/	RRID: SCR_001905
RStudio	https://www.rstudio.com/	RRID:SCR_000432
ggplot2 (package)	https://cran.r-project.org/web/packages/ggplot2/index.html	RRID:SCR_014601
FSA (package)	https://cran.r-project.org/web/packages/FSA/index.html	NA
chisq.posthoc.test	https://cran.r-project.org/web/packages/chisq.posthoc.test/readme/README.html	NA

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Dr. Yih-Chung Tham (thamyc@nus.edu.sg).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- We have ensured that all the essential data necessary for replicating our results is included in our supplementary file. The only exception is the raw scores assigned by individual graders, which can be provided by the lead contact upon request.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Not applicable.

METHOD DETAILS

The study took place between May 11th 2023 and July 8th 2023 at the Ophthalmology Department at National University Hospital, National University Health System (NUHS), Singapore.

Figure 1 illustrates the study design. To comprehensively evaluate the capabilities of LLM-Chatbots across a diverse spectrum of ocular symptoms, a team of experienced ophthalmologists (DZC, MCJT, HAHL) and clinical optometrists (SMEY, YCT) collaborated to curate 37 questions on ocular symptoms. After gathering common queries from reputable online health information sites (e.g., National Eye Institute,[31] American Academy of Ophthalmology[32]), the expert panel further refined these questions, selecting those commonly encountered in a clinical setting, ensuring the relevance and practicality of the inquiries. These questions were categorised into two broad categories: visual disturbance-related symptoms (including acute vision loss, chronic vision loss, diplopia, metamorphopsia, photopsia, myodesopsia, visual field defect, photophobia, and glare sensitivity/haloes), and other common ocular symptoms (including red eye, painful eye, pruritus, pressure sensation, epiphora, ptosis, blepharospasm, foreign body sensation, trauma, and asthenopia).

The queries were then individually posed to ChatGPT-3.5, ChatGPT-4.0 and Google Bard between May 18th to July 4th 2023. To mitigate potential memory bias from the LLM-Chatbots, the conversation was reset after each prompt. Any attributes specific to the LLM-Chatbots' answering format were then removed from the final generated responses (Tables S1, S2, and S3).

Subsequently, the anonymised responses from the three Chatbots were randomly shuffled and then presented as three 'randomized sets' to each of the three assigned ophthalmologists (each possessing $\geq$8 years of clinical ophthalmology experience), for independent grading. The grading was conducted over three separate rounds, with each round dedicated to one set. To minimise recency bias, a 48-h wash-out interval was implemented between the assessment of each round.

QUANTIFICATION AND STATISTICAL ANALYSIS

Accuracy evaluation

The three assigned graders independently assessed the accuracy of every response generated by the respective LLM-Chatbots. We employed a three-tier grading system as follows: 1) 'Poor' for responses that contained significant inaccuracies capable of misleading patients and potentially causing harm; 2) 'Borderline' for responses that might have had factual errors but were unlikely to mislead or harm patients; 3) 'Good' for responses that were free of errors. Final accuracy ratings for each LLM-Chatbot response were determined using a majority consensus approach amongst the three graders. When consensus was not reached, with each grader providing a different rating, we adopted a rigorous approach by assigning the lowest score (i.e., 'poor') to the LLM-Chatbot response.

The total accuracy score (continuous measure) for each LLM-Chatbot response was determined by summing the scores assigned by the three graders (Tables S4 and S5).

Comprehensiveness evaluation

We conducted an additional assessment on LLM-Chatbot responses that received a 'good' rating through majority consensus to evaluate their degree of comprehensiveness. This involved the use of a five-point scale, encompassing the following categories: 1) 'Not comprehensive' for responses lacking substantial details; 2) 'Slightly comprehensive' for responses with minimal but essential information; 3) 'Moderately comprehensive' for responses that presented a reasonable level of detail; 4) 'Comprehensive' for responses addressing most critical elements; 5) 'Very comprehensive' for responses presenting thorough and detailed information. The overall mean comprehensiveness score was determined by averaging the scores given by each grader across the total number of 'good' rated responses.

Evaluation of self-awareness levels in LLM-Chatbots

The term "self-awareness" in this context refers to the ability of these LLM models to self-check and self-correct their responses. To evaluate self-checking, we prompted the LLM-Chatbots with this question - "Could you kindly check if your answer is correct?" after each response. This prompt was deliberately designed to be generic and not directly point out possible errors, so as to provide a more accurate assessment of the LLM-Chatbots' capacity to self-check.

On the other hand, to assess the LLM-Chatbots' self-correcting capabilities, responses initially rated as 'poor' were further prompted with this line, "That does not seem quite right. Could you kindly review?". The regenerated responses upon self-correction, were re-assessed by the three graders one week after the initial grading. To minimise bias in grading for this segment, the graders were not informed that these responses had been self-corrected and remained blinded to the original 'poor'-rated responses.

Detailed qualitative analysis of poorly-rated LLM-Chatbot responses

To further shed light on the potential limitations and risks of relying on LLM-Chatbots for answers regarding ocular symptoms, a designated expert (DZC) further identified and emphasised erroneous or inaccurate sentences present in 'poor'-rated responses. Furthermore, explanations were also provided to elucidate the nature of these inaccuracies.

Statistical analysis

Statistical analyses were performed using R (Version 4.1.1, R Foundation, Vienna, Austria). To assess differences in character count, word count, total accuracy scores, and comprehensiveness scores among responses from the three LLM-Chatbots, we performed the Kruskal-Wallis rank-sum test and Dunn's multiple comparison post-hoc test (the data did not meet parametric assumptions). To compare the proportions of 'good', 'borderline', and 'poor' ratings among the LLM-Chatbots, we performed the two-tailed Pearson's $\chi 2$ test.

p-values were adjusted using the Bonferroni correction method to account for multiple hypothesis tests. Statistical significance was considered at $p < 0.05$.

ADDITIONAL RESOURCES

Not applicable.