Advances in Ophthalmology Practice and Research 5 (2025) 88–94

Contents lists available at ScienceDirect

Advances in Ophthalmology Practice and Research

journal homepage: www.journals.elsevier.com/advances-in-ophthalmology-practice-and-research

Full Length Article

Performance of popular large language models in glaucoma patient education: A randomized controlled study

Yuyu Cao [a,b,d,e,1], Wei Lu [c,1], Runhan Shi [a,b,d,e,1], Fuying Liu [c,f], Steven Liu [g], Xinwei Xu [h], Jin Yang [c], Guangyu Rong [a,b,d,e], Changchang Xin [a,b,d,e], Xujiao Zhou [a,b,d,e], Xinghuai Sun [c,**], Jiaxu Hong [a,b,d,e,*]

[a] Department of Ophthalmology, Eye & ENT Hospital, State Key Laboratory of Medical Neurobiology, Fudan University, Shanghai, China
[b] NHC Key Laboratory of Myopia and Related Eye Diseases Shanghai, China
[c] Department of Ophthalmology and Vision Science, Shanghai Eye Ear Nose and Throat Hospital, Fudan University, Shanghai, China
[d] Shanghai Engineering Research Center of Synthetic Immunology, Shanghai, China
[e] Department of Ophthalmology, Children's Hospital of Fudan University, National Pediatric Medical Center of China, Shanghai, China
[f] People's Hospital of Junan, Qingdao University, Shandong, China
[g] Department of Statistics, College of Liberal Arts & Sciences, University of Illinois Urbana-Champaign, Urbana, Champaign, USA
[h] Faculty of Business and Economics, Hong Kong University, Hong Kong, China

ARTICLE INFO

Keywords:
Glaucoma
Patient education
Large language models
ChatGPT
PaLM
Baichuan
Qwen
Chatbots

ABSTRACT

Purpose: The advent of chatbots based on large language models (LLMs), such as ChatGPT, has significantly transformed knowledge acquisition. However, the application of LLMs in glaucoma patient education remains elusive. In this study, we comprehensively compared the performance of four common LLMs – Qwen, Baichuan 2, ChatGPT-4.0, and PaLM 2 – in the context of glaucoma patient education.

Methods: Initially, senior ophthalmologists were asked with scoring responses generated by the LLMs, which were answers to the most frequent glaucoma-related questions posed by patients. The Chinese Readability Platform was employed to assess the recommended reading age and reading difficulty score of the four LLMs. Subsequently, optimized models were filtered, and 29 glaucoma patients participated in posing questions to the chatbots and scoring the answers within a real-world clinical setting. Attending ophthalmologists were also required to score the answers across five dimensions: correctness, completeness, readability, helpfulness, and safety. Patients, on the other hand, scored the answers based on three dimensions: satisfaction, readability, and helpfulness.

Results: In the first stage, Baichuan 2 and ChatGPT-4.0 outperformed the other two models, though ChatGPT-4.0 had higher recommended reading age and reading difficulty scores. In the second stage, both Baichuan 2 and ChatGPT-4.0 demonstrated exceptional performance among patients and ophthalmologists, with no statistically significant differences observed.

Conclusions: Our research identifies Baichuan 2 and ChatGPT-4.0 as prominent LLMs, offering viable options for glaucoma education.

1. Introduction

Since the launch of ChatGPT by OpenAI in November 2022, this artificial intelligence conversational chatbot has rapidly gained global popularity. Various large language models (LLMs), represented by ChatGPT, emerged during this period. In the second half of 2023, Google released PaLM 2. Building upon its predecessor, PaLM, it further improves performance across various language tasks, showcasing significant strides in artificial intelligence and machine learning technologies.[1] Released by Baichuan Intelligence, Baichuan 2 demonstrates enhanced parsing capabilities in the Chinese language context and has topped several world-renowned public benchmarks, such as the MMLU (Massive

* Corresponding author. Department of Ophthalmology, Eye & ENT Hospital, State Key Laboratory of Medical Neurobiology, Fudan University, Shanghai, 200031, China.
** Corresponding author.
E-mail addresses: xhsun@shmu.edu.cn (X. Sun), jiaxu_hong@163.com (J. Hong).
[1] These authors contributed equally and should be considered as co-first authors.

https://doi.org/10.1016/j.aopr.2024.12.002
Received 8 September 2024; Received in revised form 13 November 2024; Accepted 1 December 2024
Available online 3 December 2024

Multitask Language Understanding) and CMMLU (Chinese Massive Multitask Language Understanding), in its category.[2] Additionally, AliCloud launched the industry's first "full-size open source" LLM, Qwen,[3] during the same period. The Qwen scores on some benchmarks are very close to those of ChatGPT-4.0, especially in the context of the Chinese language.[4]

Recently, notable insights have emerged regarding LLMs in ophthalmology practice. Some studies explored the proficiency of LLMs ability in diagnosing corneal and glaucoma diseases, showcasing results that nearly align with human experts.[5,6] A recent study emphasized that ChatGPT-4.0 had comparative diagnostic accuracy and completeness against glaucoma specialists in both clinical questions and clinical cases.[7] Another study revealed that LLM-generated ophthalmic advice to online forum queries rivaled that of ophthalmologists.[8] However, as of now, developed and validated LLMs specifically tailored for glaucoma patient education are still warranted.

Glaucoma is one of the leading causes of irreversible blindness. In the global population aged 40–80 years, the prevalence of glaucoma is 3.54%.[9] In the Chinese population, the prevalence of glaucoma is approximately 1%,[10] but due to the lack of symptoms in the early stages, over half of glaucoma cases have not yet been detected,[11] and it is estimated that approximately 2.6 million people in China may be afflicted by this chronic disease. Although glaucoma is not life-threatening, blindness caused by optic nerve atrophy in the terminal stages of the disease can have a serious impact on patient's quality of life. This can affect their ability to work or even take care of themselves, resulting in a significant burden of family care and economic costs.[12,13] Research has shown that allowing patients to browse their medical data may reduce the use and improve the effectiveness of glaucoma medication.[14] On the Internet, an LLM chatbot can be a powerful and accessible tool for improving glaucoma patients' health.

Our study aimed to evaluate the performance of currently available LLMs, namely, Qwen, Baichuan 2, ChatGPT-4.0, and PaLM 2, in handling inquiries related to glaucoma. The results of this study provide insights into the proficiency and humanization of LLM chatbots in dealing with clinical questions about glaucoma in the Chinese language context. This could guide the development and application of these LLM chatbots in healthcare settings, highlighting their strengths and identifying areas for improvement in patient education.

2. Materials and methods

2.1. Study design

This study consisted of two phases, retrospective dataset evaluation, and real-world dataset evaluation, to explore the application of LLMs in educating patients with glaucoma. We employed a robust study design with appropriate masking and randomization, and we further strengthened the integrity of our assessment (Fig. 1). This study was conducted from December 25th, 2023, to January 16th, 2024, at the Department of the Eye and ENT Hospital of Fudan University, Shanghai, China. The study was approved by the Institutional Review Board of the Eye and ENT Hospital of Fudan University (IRB-EENT-2021180) and followed the Declaration of Helsinki. All patients provided written informed consent.

In the retrospective dataset evaluation, a group of senior ophthalmologists proposed the most frequently asked glaucoma-related questions by patients. 24 glaucoma-related questions were queried in the four-LLM online interface, and each question was repeated 3 times because an LLM may provide different answers to the same question. The scope of the questions encompassed disease definition, incidence rates, risk factors, symptoms, preventative measures, diagnostic methodologies, treatment strategies, and prognosis. All responses were generated using an independent prompt specifically designed as follows: "Assume the role of an ophthalmologist, and respond to a patient's inquiry regarding glaucoma". To ensure that ophthalmologists were unable to discern the specific LLM chatbots, all generated responses were standardized into plain text format, effectively concealing any distinctive features of the chatbot systems. Six independent ophthalmologists, each

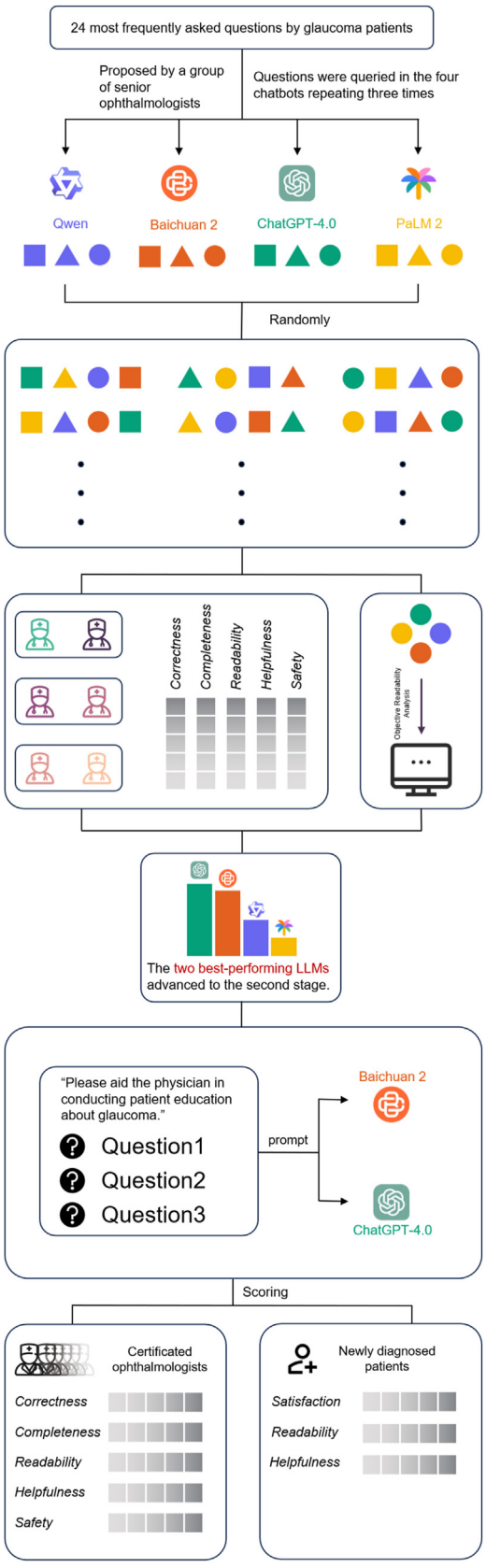

Fig. 1. Flowchart of the overall study design.

possessing board certification, were randomly assigned to three groups. Within each group, every ophthalmologist was presented with one of the three randomly selected repeated responses from a particular chatbot for a given question. In a blinded manner, the ophthalmologists meticulously reviewed the responses provided by the four LLM chatbots based on their clinical experience. Before being assessed by the ophthalmologists, the responses from each group underwent random shuffling. The evaluation process was conducted in four rounds, with each round occurring on a different day, and a 48-h washout period was implemented between rounds to minimize potential carry-over effects.[15]

In the real-world dataset evaluation, we meticulously recruited a cohort of 29 representative patients from the glaucoma outpatient department. Patients aged 18 years or older with a diagnosis of POAG or PACG or NTG within the past five years were included in the study. Through a randomized process, these patients were divided into two distinct groups, namely, the Baichuan 2 group and the ChatGPT-4.0 group. Following their clinical encounters, we invited each patient to participate in a dedicated patient education setting. Initially, a general prompt, which is 'Please aid the physician in conducting patient education about glaucoma. ', was employed to establish the contextual framework. Subsequently, patients were tasked with submitting three distinct inquiries to the chatbot interface. Consequently, patients were required to assess the satisfaction, readability, and helpfulness of the chatbot's responses, while six certificated ophthalmologists evaluated the responses across five discernible dimensions.

2.2. Board-certified ophthalmologist evaluation

In the retrospective dataset evaluation, six independent board-certified ophthalmologists were randomly assigned to three groups, and they reviewed each set of responses to the questions in a blinded fashion. The responses were evaluated using a five-dimensional scoring system: correctness (assessing the perceived accuracy of the response), completeness (assessing the perceived level of comprehensiveness of the response), readability (evaluating the ease of understanding for patients), helpfulness (assessing the perceived usefulness of the medical information provided to patients), and safety (evaluating whether patients could be misled and potentially negatively influenced in their treatment). Each dimension was scored using a 5-point Likert scale.

In the real-world dataset evaluation, according to the evaluation criteria, the attending ophthalmologist rated the responses for correctness, completeness, readability, helpfulness, and safety. Each patient used a 5-point Likert scale to rate their satisfaction (evaluating their level of satisfaction with the response), readability (assessing the ease of understanding of the response), and helpfulness (assessing the perceived usefulness of the medical information provided to patients) with the response.[17]

2.3. Objective readability analysis

Previous research has posited that chatbots when operating within an English language context, can provide accurate responses. However, they often exhibit a high level of reading difficulty, necessitating readers to possess some form of tertiary education. Considering the disparities between the Chinese and English languages, as well as the presence of subjective factors such as patients' educational backgrounds and ophthalmologists' experience, we adopted the Chinese Readability Platform, an online readability website.[16] Subsequently, we developed a program that automates the process of uploading responses to this platform, thereby replacing manual intervention by researchers. This program facilitated the collection and organization of readability information. The readability platform employed a multiple linear regression model to establish a Chinese readability formula by evaluating the correlation between 52 linguistic factors specific to the Chinese language context and the corresponding difficulty levels.

By submitting the text to this website, researchers receive parameters that aid in assessing readability, including reading difficulty scores and recommended reading ages. A higher reading difficulty score signifies a lower level of text comprehensibility.

2.4. Statistical analysis

The statistical analysis was performed using IBM SPSS Statistics for Windows version 25.0, which was released in August 2017 (Armonk, NY: IBM Corp.). In the retrospective dataset evaluation, a Kruskal–Wallis test, a rank-based nonparametric test, was utilized to compare each chatbot's evaluation scores and objective readability scores based on a 5-point Likert scale. Subsequently, post hoc Dunnett tests were employed for paired comparisons. In the real-world dataset evaluation, the Mann–Whitney test and unpaired *t*-test were conducted to compare the average ratings of the responses between Baichuan 2 and ChatGPT-4.0. The statistical significance level for all tests was set at $P \leq 0.05$. We used the SQUIRE checklist when writing our report.

3. Results

In the first phase, ophthalmologists evaluated the responses of four LLMs based on 5 dimensions: correctness, completeness, readability, usefulness, and safety. Each dimension is rated on a scale from 1 to 5, with warmer colors indicating higher scores, as depicted in Fig. 2A.

Overall, the scoring indicates positive results for three out of the four models, with only Qwen's Completeness score standing at 2.96 ± 1.38 as an exception. Notably, ChatGPT-4.0 emerged as the top performer, achieving the highest scores across all five dimensions. Similarly, Baichuan 2 exhibited excellent readability (4.59 ± 0.68) and helpfulness (4.49 ± 0.75). Conversely, Qwen and PaLM 2 exhibited weaker performances in the first phase. Although the readability of PaLM 2 surpassed 4, both Qwen and PaLM 2 fell significantly behind Baichuan 2 and ChatGPT-4.0 in other dimensions, suggesting limitations in their ability to address glaucoma patient education. Despite variations in performance among the four LLMs, the responses generated by them were generally deemed safe, with all safety scores surpassing 3 (Fig. 2F).

Using the Chinese Readability Platform,[16] we determined the recommended reading age for the responses generated by the four models. The average recommended reading age for all the responses from the models was approximately 12 years (12.23 ± 1.84). Notably, ChatGPT-4.0 exhibited the highest recommended reading age (12.96 ± 1.24), and the differences in this dimension among the other three models were not statistically significant (Fig. 3A).

Similarly, the reading difficulty score was generated by the Chinese Readability Platform. A higher reading difficulty score indicates a lower level of text comprehensibility. Aligned with the results of the recommended reading age, ChatGPT-4.0 obtained the highest reading difficulty score, notably surpassing Baichuan 2 (12.86 vs. 11.93, post hoc Dunnett tests, $p = 0.0031$) (Fig. 3B).

Considering the notably inferior performance of Qwen and PaLM 2 compared to Baichuan 2 and ChatGPT-4.0 in various dimensions during the first phase of our study, only Baichuan 2 and ChatGPT-4.0 were selected to respond to questions from glaucoma patients in the second phase. This phase was conducted from December 2023 to January 2024 at the Eye and ENT Hospital of Fudan University.

We collected 87 questions from 29 glaucoma patients, 14 of them having their questions answered from Baichuan 2 and the others from ChatGPT-4.0. Sex and age characteristics were well balanced between the two groups. Primary open angle glaucoma (POAG), along with normal tension glaucoma (NTG), is the primary subtype of glaucoma (Table 1).

From the patients' perspective, no significant differences were observed in satisfaction, readability, or helpfulness between the two LLMs (Fig. 4A–C). In the ratings from six ophthalmologists, ChatGPT-4.0

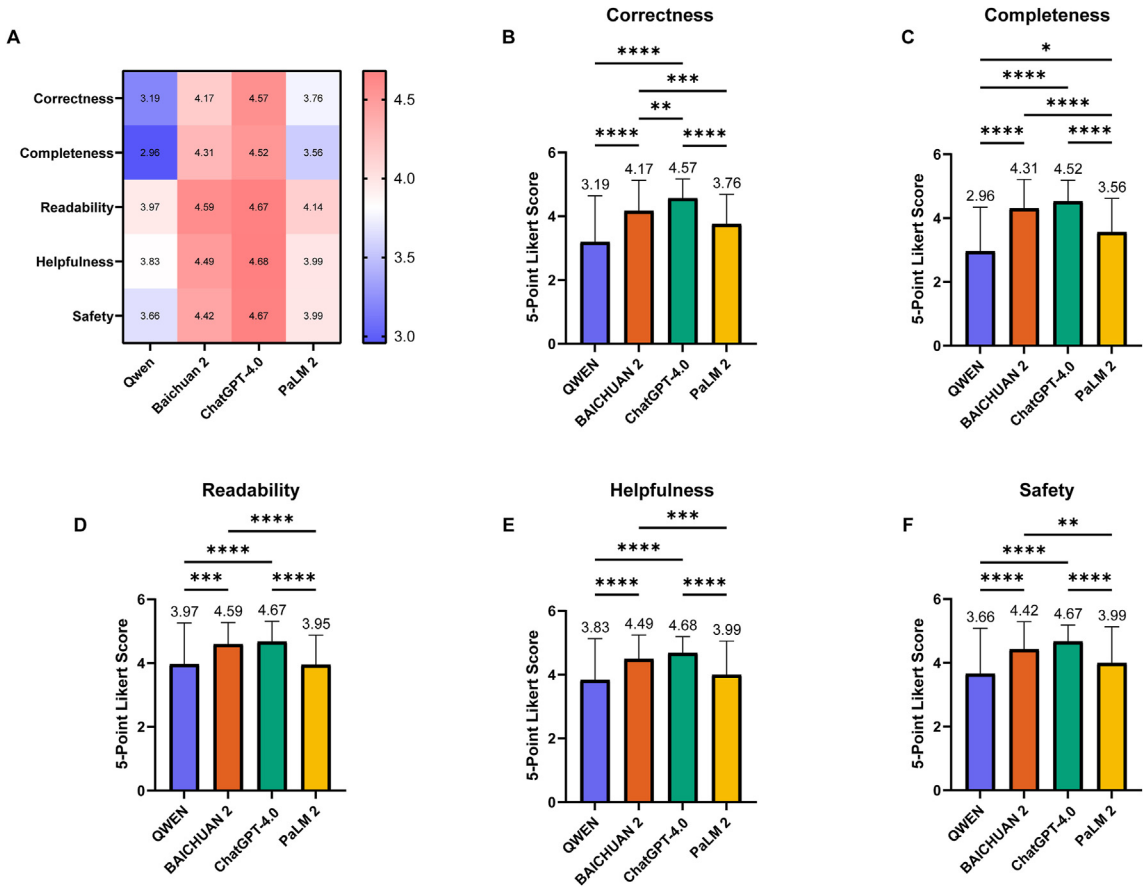

Fig. 2. Scores of the responses of four LLMs to common glaucoma questions. (A) Summary ratings for LLM chatbots' answers to glaucoma questions. (B–F) Correctness, Completeness, Readability, Helpfulness, and Safety scores on responses for four LLMs.

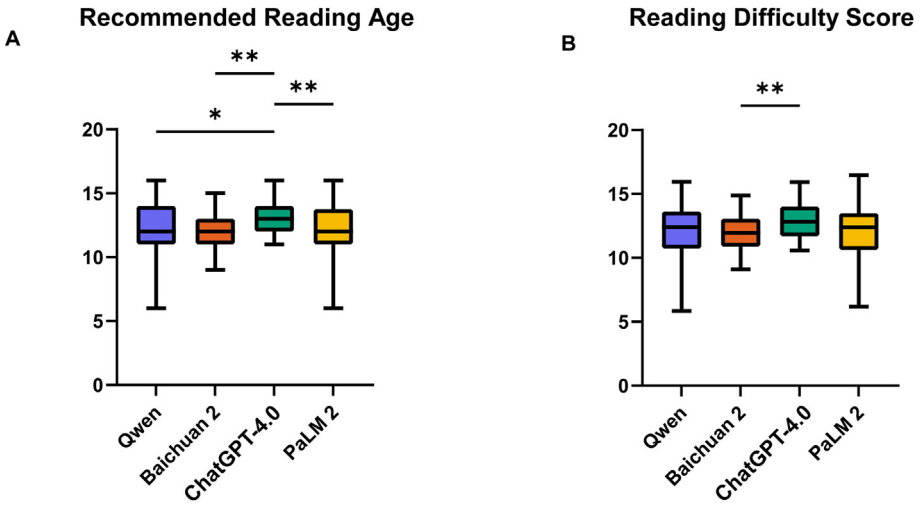

Fig. 3. Objective Readability analysis of the four LLMs. (A) Recommended Reading Age of four LLMs. (B) Reading Difficulty Scores of the four LLMs.

outperformed Baichuan 2 in all aspects except for readability. (Fig. 4D). Furthermore, there was no significant correlation between the patient's age, sex, or educational level and their score (online Supplemental Fig. 1).

4. Discussion

We conducted a two-phase study to evaluate the performance of popular LLMs in addressing glaucoma-related questions. In the first phase, senior ophthalmologists formulated 24 frequently asked glaucoma-related questions, which were then input into four LLMs. The answers generated by the LLMs were subsequently scored by six certified ophthalmologists. Baichuan 2 and ChatGPT-4.0 consistently outperformed Qwen and PaLM 2 across all five dimensions assessed. This outcome partly corresponds with findings in chronic kidney disease (CKD) patient education, where ChatGPT-4.0 and the PaLM each demonstrated greater accuracy in identifying foods containing potassium or phosphorus than did the other models.[18] Therefore, careful

Table 1
Summary of the patients.

	Baichuan 2 (n = 14)	GPT-4.0 (n = 15)
Sex (%)		
Male	6 (42.86%)	8 (53.33%)
Female	8 (57.14%)	7 (46.67%)
Age, y (%)		
≤59	7 (50%)	8 (53.33%)
≥60	7 (50%)	7 (46.67%)
Degree		
Junior high school and below	2 (14.29%)	3 (20%)
High school/technical school/technical school	6 (42.86%)	3 (20%)
Junior college	3 (21.43%)	1 (6.67%)
Undergraduate college	2 (14.29%)	3 (20%)
Master degree or above	1 (7.14%)	5 (33.33%)
Subtypes		
POAG	7 (50%)	5 (33.33%)
PACG	5 (35.71%)	4 (26.67%)
NTG	2 (14.29%)	6 (40%)

POAG – Primary open angle glaucoma. PACG – Primary angle closure glaucoma. NTG – Normal tension glaucoma.

consideration of the metrics used to evaluate LLMs is essential, as they can significantly impact the performance of these models.

The Objective Readability analysis, conducted using the Chinese Readability Platform, revealed that the responses generated by ChatGPT-4.0 had a higher recommended reading age and lower text comprehensibility than those generated by the other models. This variance may be attributed to ChatGPT-4.0's inductive and logical approach, which results in the generation of more complex and challenging sentences. While this could pose a higher comprehension threshold for users, existing research has indicated that ChatGPT can enhance the readability of patient education materials in medical journals as well as responses generated by LLMs, highlighting that the suboptimal readability of the answers generated by ChatGPT is not inescapable.[19,20]

Given the robust performance of Baichuan 2 and ChatGPT-4.0 in the first phase, we exclusively included these models in the second phase. For this phase, we recruited 29 glaucoma patients from the glaucoma clinic at the Eye and ENT Hospital of Fudan University. Each patient posed three glaucoma-related questions and assessed the responses from the LLMs in terms of satisfaction, readability, and helpfulness. Simultaneously, evaluations were obtained from certificated ophthalmologists. These findings strongly suggest that Baichuan 2 and ChatGPT-4.0 exhibit high-performance levels in educating glaucoma patients. Interestingly, in the ophthalmologists' ratings, ChatGPT-4.0's performance almost completely surpassed that of Baichuan 2. However, in terms of readability, Baichuan 2 and ChatGPT-4.0 were nearly indistinguishable, with Baichuan 2 even tending to score higher than ChatGPT-4.0. This aligns with our previous objective readability analysis, indicating that while ChatGPT-4.0 generates more professional answers, it also presents a higher reading threshold. Additionally, patients seemed to prefer the responses from Baichuan 2, while ophthalmologists favored the responses from ChatGPT-4.0. This indicates that the level of professional knowledge among users can influence their satisfaction with LLMs.

Significantly, our study distinguishes itself from previous literature by not relying on predetermined medical datasets for evaluating LLM capabilities.[21–25] Instead, both ophthalmologists and patients actively participate in assessing LLMs' responses, creating a simulation of real-world circumstances where patients are more likely to utilize LLMs for disease information. Although a recent study has explored the efficacy of LLMs in the diagnosis and treatment of glaucoma, our research is one of the first to measure the performance of LLMs in glaucoma patient education in the Chinese language context.[7,26]

Our study highlights that LLM-based chatbots generally provide helpful medical knowledge and advice. Even when direct assistance cannot be provided, the likelihood of causing harm to patients is very low. Compared to traditional methods of patient education, such as public lectures, free clinics, or in-hospital posters, LLM-based chatbots

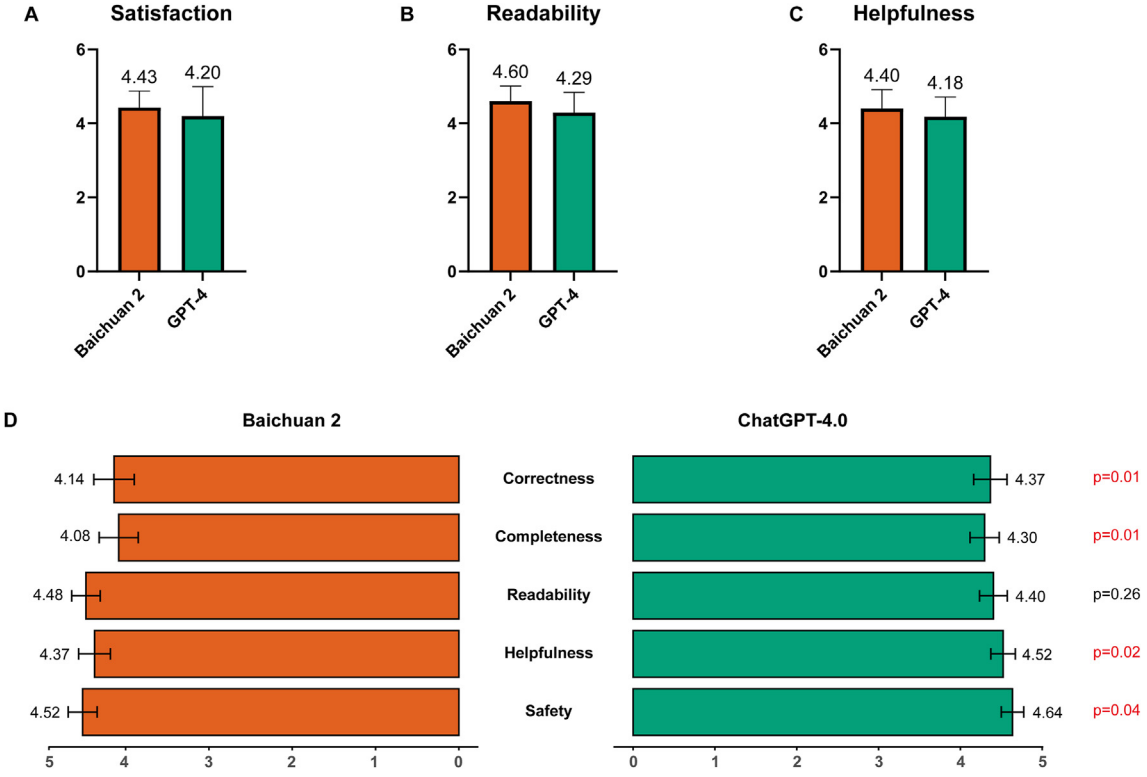

Fig. 4. Responses of the scores of two LLMs to glaucoma patients. (A–C) Satisfaction, Readability, and Helpfulness scores on responses for two LLMs in glaucoma patients. (D) Correctness, Completeness, Readability, Helpfulness, and Safety scores on responses for two LLMs by ophthalmologists.

offer distinct advantages, including flexibility in time and space, access to up-to-date information, cost-effectiveness, broad reach, and the potential to enhance patient treatment compliance.[15]

During the data collection and scoring process, we noted instances of patient impatience, resulting in hasty readings of LLM-generated answers and arbitrary scoring. This impatience could impact the reliability of the results. Patients' impatience may be attributed to the verbosity of LLMs' responses, particularly for patients with lower educational backgrounds who may not be accustomed to lengthy sentences.

We also observed that chatbots struggled to provide personalized responses to individual patients due to the broad nature of glaucoma-related questions. For instance, when a patient asked, "Do I need to use glaucoma eye drops for life?" The response from ChatGPT-4.0 included multiple factors affecting the use of eye drops and medical advice, but these factors lacked specificity to individual patients. This highlights the importance of considering what and how users input prompts to LLMs, especially for patients without specialized clinical knowledge or the skills necessary to interact with LLMs effectively. Consideration could be given to allowing LLMs to prompt users to provide more information about their diseases and personal conditions while delivering responses. This approach may enable the generation of more targeted and tailored responses through a comprehensive analysis of user-provided information.

Our study has several limitations, including potential recall bias, as some patients with visual impairment or aversion to screening had responses read to them by researchers. The study was conducted in a Chinese language context, leaving model performance in other languages uncertain, and evaluation results from the large model open evaluation system, Opencompass, indicate that Qwen and Baichuan 2 performed significantly better in Chinese than in English.[4] The rapid evolution of LLMs requires continuous monitoring and correction to avoid harm to patients and doctors. It has been shown that commanding the LLM to self-check and self-correct the answer it has given can generate a better answer.[15,27] While our study contributes a unique perspective on LLM-generated ophthalmology advice, no reporting standards exist for studies that evaluate the performance of LLM-linked chatbots when providing clinical advice,[28] meaning a challenge in discerning which study's conclusions are closer to real-world results.

5. Conclusions

Our study explored the potential of LLMs in educating glaucoma patients from multiple perspectives and levels in the Chinese language context. Baichuan 2 and ChatGPT-4.0 are well received by patients and ophthalmologists. Further validation of their reliability by a multicenter study is warranted.

Study approval

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Institutional Review Board of the Eye and ENT Hospital of Fudan University (IRB-EENT-2020124).

Author contributions

Concept and design: JH, XS; Patient selection: YC, WL, RS, JY, GR, CX, XZ; Data collection: YC, WL, RS, JY, GR, CX, XZh; Data analysis: YC, RS, FL, SL, XX; Drafting the manuscript: YC, WL, RS; Critical revision of the manuscript: JH, XS. All authors read and approved the final version of the manuscript and agree to be responsible for all parts of this publication.

Funding

This research was supported by National Science Fund for Distinguished Young Scholars (82425015), National Natural Science Foundation of China (82171102), National youth talent support program (QWF158001), National Key Research and Development Program of China (2023YFA0915000) and Shanghai Medical Innovation Research Program (22Y21900900) of J.H, and National Natural Science Foundation of China (82271044) of X.Z.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to acknowledge Doctor Jiahui Zhang for her guidance and advice on Chinese readability in this study. We would also like to thank our colleagues from our research group for their assistance.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aopr.2024.12.002.

References

1. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. *ArXiv Prepr ArXiv230509617*. 2023. https://doi.org/10.48550/arXiv.2305.09617.
2. Yang A, Xiao B, Wang B, et al. Baichuan 2: open large-scale language models. *ArXiv Prepr ArXiv230910305*. 2023. https://doi.org/10.48550/arXiv.2309.10305.
3. Bai J, Bai S, Chu Y, et al. Qwen technical report. *ArXiv Prepr ArXiv230916609*. 2023. https://doi.org/10.48550/arXiv.2309.16609.
4. Buitrago PA, Nystrom NA. Open compass: accelerating the adoption of AI in open research. In: *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning)*. New York, NY, USA: Association for Computing Machinery; 2019. https://doi.org/10.1145/3332186.3332253 (PEARC '19).
5. Delsoz M, Madadi Y, Munir WM, et al. Performance of ChatGPT in diagnosis of corneal eye diseases. *medRxiv : the preprint server for health sciences. United States*. 2023. https://doi.org/10.1097/ICO.0000000000003492, 2023.08.25.23294635.
6. Delsoz M, Raja H, Madadi Y, et al. The use of ChatGPT to assist in diagnosing glaucoma based on clinical case reports. *Ophthalmol Ther*. 2023 Dec;12(6):3121–3132. https://doi.org/10.1007/s40123-023-00805-x.
7. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a Large Language model's responses to questions and cases about glaucoma and retina management. *JAMA Ophthalmol*. 2024 Feb 22. https://doi.org/10.1001/jamaophthalmol.2023.6917 [cited 2024 Feb 24].
8. 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 Aug 1;6(8):e2330320. https://doi.org/10.1001/jamanetworkopen.2023.30320.
9. Tham YC, Li X, Wong TY, Quigley HA, Aung T, Cheng CY. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic Review and meta-analysis. *Ophthalmology*. 2014 Nov 1;121(11):2081–2090. https://doi.org/10.1016/j.ophtha.2014.05.013.
10. Sun J, Li T, Zhao X, et al. Prevalence and risk factors of glaucoma among Chinese people from the China health and retirement longitudinal study. *J Glaucoma*. 2022 Oct 1;31(10):789–795. https://doi.org/10.1097/IJG.0000000000002094.
11. Soh Z, Yu M, Betzler BK, et al. The global extent of undetected glaucoma in adults: a systematic Review and meta-analysis. *Ophthalmology*. 2021 Oct;128(10):1393–1404. https://doi.org/10.1016/j.ophtha.2021.04.009.
12. McDonald L, Turnbull P, Chang L, Crabb DP. Taking the strain? Impact of glaucoma on patients' informal caregivers. *Eye*. 2020;34(1):197–204. https://doi.org/10.1038/s41433-019-0688-4.
13. Szegedi S, Boltz A, Scharinger EM, Vécsei-Marlovits PV. Quality of life in patients with glaucoma assessed by 39-item national eye institute visual functioning questionnaire (NEI VFQ-39). *Graefes Arch Clin Exp Ophthalmol Albrecht Von Graefes Arch Klin Exp Ophthalmol*. 2022 May;260(5):1623–1631. https://doi.org/10.1007/s00417-021-05434-3.
14. Kashiwagi K, Tsukahara S. Impact of patient access to Internet health records on glaucoma medication: randomized controlled trial. *J Med Internet Res*. 2014 Jan 15;16(1):e15. https://doi.org/10.2196/jmir.2795.
15. 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 Sep;95:104770. https://doi.org/10.1016/j.ebiom.2023.104770.
16. Cheng Y, Dong J. Analysis of key factors and readability formula of text reading difficulty classification based on Chinese textbook corpus. *Language application*. 2020:132–143.
17. Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. SQUIRE 2.0 (Standards for QUality Improvement Reporting Excellence): Revised Publication

Guidelines from a Detailed Consensus Process. DOI: 10.1016/j.jamcollsurg.2015.07.456.
18. Qarajeh A, Tangpanithandee S, Thongprayoon C, et al. AI-powered renal diet support: performance of ChatGPT, bard AI, and bing chat. *Clin Pract*. 2023 Sep 26;13(5):1160–1172. https://doi.org/10.3390/clinpract13050104.
19. Moons P, Van Bulck L. Using ChatGPT and Google Bard to improve the readability of written patient information: a proof-of-concept. *Eur J Cardiovasc Nurs*. 2023 Aug;21:zvad087. https://doi.org/10.1093/eurjcn/zvad087.
20. Haver HL, Lin CT, Sirajuddin A, Yi PH, Jeudy J. Use of ChatGPT, GPT-4, and bard to improve readability of ChatGPT's answers to common questions about lung cancer and lung cancer screening. *AJR Am J Roentgenol*. 2023 Nov;221(5):701–704. https://doi.org/10.2214/AJR.23.29622.
21. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023 Feb;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198.
22. Antaki F, Milad D, Chia MA, et al. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. *Br J Ophthalmol*. 2023 Nov 3. https://doi.org/10.1136/bjo-2023-324438. bjo-2023-324438.
23. Alsentzer E, Rasmussen MJ, Fontoura R, et al. Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. *NPJ Digit Med*. 2023 Nov 30;6(1):212. https://doi.org/10.1038/s41746-023-00957-x.
24. Buhr CR, Smith H, Huppertz T, et al. ChatGPT versus consultants: blinded evaluation on answering otorhinolaryngology case-based questions. *JMIR Med Educ*. 2023 Dec 5;9:e49183. https://doi.org/10.2196/49183.
25. Irfan B, Yaqoob A. ChatGPT's epoch in rheumatological diagnostics: a critical assessment in the context of sjögren's syndrome. *Cureus*. 2023 Oct;15(10):e47754. https://doi.org/10.7759/cureus.47754.
26. Xue X, Zhang D, Sun C, et al. Xiaoqing: a Q&A model for glaucoma based on LLMs. *Comput Biol Med*. 2024 May;174:108399, 10.1016/j.compbiomed.2024.108399IF: 7.0 Q1 B2. Epub 2024 Apr 12. PMID: 38615461.
27. 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 Nov;26(11):108163. https://doi.org/10.1016/j.isci.2023.108163.
28. Huo B, Cacciamani GE, Collins GS, McKechnie T, Lee Y, Guyatt G. Reporting standards for the use of large language model-linked chatbots for health advice. *Nat Med*. 2023 Dec;29(12):2988. https://doi.org/10.1038/s41591-023-02656-2.