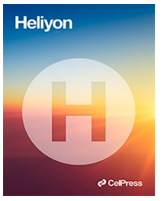
Heliyon 10 (2024) e34391

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Benchmarking four large language models' performance of addressing Chinese patients' inquiries about dry eye disease: A two-phase study

Runhan Shi [a,b,c,d], Steven Liu [e], Xinwei Xu [f], Zhengqiang Ye [a], Jin Yang [a], Qihua Le [a], Jini Qiu [a], Lijia Tian [a], Anji Wei [a], Kun Shan [a], Chen Zhao [a], Xinghuai Sun [a], Xingtao Zhou [a], Jiaxu Hong [a,b,c,d,*]

[a] Department of Ophthalmology and Vision Science, State Key Laboratory of Molecular Engineering of Polymerse, Fudan University, Shanghai, 200031, China
[b] NHC Key laboratory of molecular engineering of polymers, Fudan University, Shanghai, 200031, China
[c] Shanghai Engineering Research Center of Synthetic Immunology, Shanghai, 200032, China
[d] Department of Ophthalmology, Children's Hospital of Fudan University, National Pediatric Medical Center of China, Shanghai, China
[e] Department of Statistics, College of Liberal Arts & Sciences, University of Illinois Urbana-Champaign, Illinois, USA
[f] Faculty of Business and Economics, Hong Kong University, Hong Kong Special Administrative Region, China

ARTICLE INFO

Keywords:
Large language model
Ophthalmology
Dry eye disease
Patient education
Real world interview

ABSTRACT

Purpose: To evaluate the performance of four large language models (LLMs)—GPT-4, PaLM 2, Qwen, and Baichuan 2—in generating responses to inquiries from Chinese patients about dry eye disease (DED).

Design: Two-phase study, including a cross-sectional test in the first phase and a real-world clinical assessment in the second phase.

Subjects: Eight board-certified ophthalmologists and 46 patients with DED.

Methods: The chatbots' responses to Chinese patients' inquiries about DED were assessed by the evaluation. In the first phase, six senior ophthalmologists subjectively rated the chatbots' responses using a 5-point Likert scale across five domains: correctness, completeness, readability, helpfulness, and safety. Objective readability analysis was performed using a Chinese readability analysis platform. In the second phase, 46 representative patients with DED asked the two language models (GPT-4 and Baichuan 2) that performed best in the in the first phase questions and then rated the answers for satisfaction and readability. Two senior ophthalmologists then assessed the responses across the five domains.

Main outcome measures: Subjective scores for the five domains and objective readability scores in the first phase. The patient satisfaction, readability scores, and subjective scores for the five-domains in the second phase.

Results: In the first phase, GPT-4 exhibited superior performance across the five domains (correctness: 4.47; completeness: 4.39; readability: 4.47; helpfulness: 4.49; safety: 4.47, $p < 0.05$). However, the readability analysis revealed that GPT-4's responses were highly complex, with an average score of 12.86 ($p < 0.05$) compared to scores of 10.87, 11.53, and 11.26 for Qwen, Baichuan 2, and PaLM 2, respectively. In the second phase, as shown by the scores for the five

* Corresponding author. Department of Ophthalmology and Vision Science, State Key Laboratory of Molecular Engineering of Polymerse, Fudan University, Shanghai, 200031, China.
E-mail address: jiaxu.hong@fdeent.org (J. Hong).

https://doi.org/10.1016/j.heliyon.2024.e34391
Received 9 January 2024; Received in revised form 8 July 2024; Accepted 9 July 2024
Available online 14 July 2024

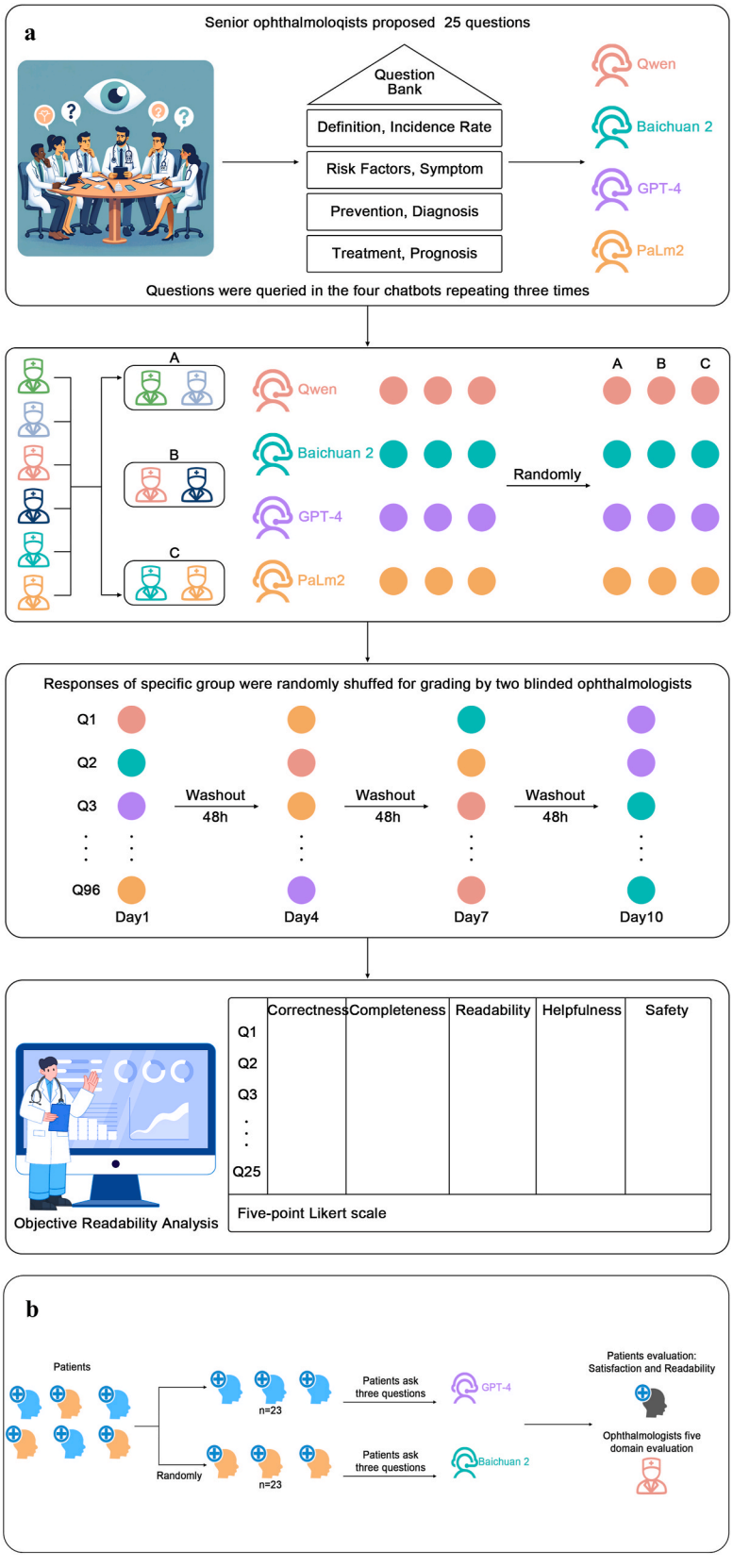


Fig. 1. Flowchart of overall study design.

domains, both GPT-4 and Baichuan 2 were adept in answering questions posed by patients with DED. However, the completeness of Baichuan 2's responses was relatively poor (4.04 vs. 4.48 for GPT-4, $p < 0.05$). Nevertheless, Baichuan 2's recommendations more comprehensible than those of GPT-4 (patient readability: 3.91 vs. 4.61, $p < 0.05$; ophthalmologist readability: 2.67 vs. 4.33). *Conclusions:* The findings underscore the potential of LLMs, particularly that of GPT-4 and Baichuan 2, in delivering accurate and comprehensive responses to questions from Chinese patients about DED.

1. Introduction

A large language model refers to a machine learning model characterized by a massive parameter count and intricate computational architecture.These models are trained on large volumes of textual data, enabling them to perform a broad range of tasks, such as text summarization, translation, and sentiment analysis [1,2]. In November 2022, OpenAI introduced its inaugural LLM, known as ChatGPT (Chat Generative Pre-trained Transformer), which marked a significant leap in natural language processing [3]. This system is fundamentally built upon by the OpenAI GPT-3 language model, the third iteration of generatively pretrained models, endowing it with the ability to comprehend, respond to, and generate text [4]. After the release of ChatGPT, there has been a proliferation of various types of LLMs. Their capacity to process large quantities of textual data both rapidly and accurately make them a promising tool for patient education and healthcare [5–7].

Dry eye disease (DED) is one of the most common eye diseases worldwide, with a prevalence rate ranging from 5 % to 50 % [8]. Patient education and healthcare play vital roles in the management and treatment outcomes of DED. In China, with a large patient population and limited medical resources, providing individually tailored education and support to every patient is challenging. Therefore, the internet has become a significant channel for Chinese patients to access medical information [9,10]. Traditionally, patients have relied on search engines to retrieve the information they needed, requiring them to sift through vast amounts of data to find relevant answers. The emergence of generative Artificial Intelligence(AI) technologies, like ChatGPT, offers the potential for a faster and simpler approach to this task. In terms of ophthalmology, AI technologies have the potential to augment the work of ophthalmologists by providing patients with preliminary information about common eye health concerns [11]. However, LLMs are not specifically designed for the medical field, and their training process involves self-supervised methods using diverse internet texts, rather than specially curated datasets [1]. This could give rise to limitations, including potential inaccuracies, biases, or responses that are difficult to comprehend. Moreover, despite the persuasive nature of their responses, these may not always be factually accurate. Therefore, the reliability and accuracy of chatbots in answering medical questions posed by patients need to be comprehensively evaluated [12].

A number of recent studies have explored the performance of LLMs in the field of ophthalmology [13–18]. In one study, GPT-4 achieved a high accuracy rate (84.6 %) in answering queries related to common vitreoretinal surgeries, although the readability (The level of difficulty in reading and comprehending a text for non-medical professionals) of its responses was suboptimal [18]. In another study that assessed three LLMs (ChatGPT-3.5, ChatGPT-4.0, and Google Bard) in the context of myopia care, ChatGPT-4.0 demonstrated superior accuracy [13]. However, only a few studies have evaluated the performance of LLM-generated responses in the context of educating Chinese patients about DED [13–19]. Furthermore, no studies on the performance of chatbots in the field of patient education have incorporated real-world patient interactions with chatbots.

In this study, we aimed to assess and compare the performance of four publicly available LLMs for education of Chinese patients on DED in a real-world setting: GPT-4, PaLM2, Qwen, and Baichuan 2.

2. Methods

2.1. Study design

This study consisted of two phases: a retrospective cross-sectional study for test and a real-world study for validation (Fig. 1). The study took place between 9 October and November 2,2023 in the Department of Ophthalmology, Eye, Ear, Nose, and Throat Hospital of Fudan University, Shanghai, China. The Institutional Review Board of the Eye and ENT Hospital of Fudan University (IRB-EENT-2020124) approved the study, and the study followed the tenets of the Declaration of Helsinki. All patients provided written informed consent.

Regarding the relationship between the two phases of our research, it is necessary to provide an important clarification. The first

Table 1
Large Language Models (LLMs) used in this study.

LLMs	Version	Company
GPT-4	4.0	OpenAI (USA)
PaLM 2	2.0	Google (USA)
Qwen	14B	Alibaba (China)
Baichuan 2	53B	Baichuan (China)

phase involves the preliminary screening and selection of four chatbots, with the results obtained solely used to determine the candidate chatbots for further evaluation in the second phase. Apart from the screening process, the implementation and outcomes of the second phase—focused on detailed performance evaluation and user satisfaction—are entirely independent and unaffected by the specific results of the first phase.

In the first phase of the study, senior ophthalmologists (N = 6) proposed the most frequently asked questions by patients about dry eye disease (Table S1). Twenty-five questions about DED were inputted into the online interfaces of the four LLMs (Table 1), with each question repeated three times to account for potential variations in the LLMs' responses [18]. The scope of the questions encompassed disease definition, incidence rate, risk factors, symptoms, preventative measures, diagnostic methodologies, treatment strategies, and prognosis. All responses were generated using an independent prompt, as follows: "Assume the role of an ophthalmologist and respond to a patient's inquiry about a specific anterior segment eye disorder."

To prevent the ophthalmologists from identifying the specific LLM chatbot, all generated responses were converted into plain text format, effectively concealing any distinctive features of the chatbot systems. Six independent board-certified ophthalmologists were randomly assigned to three groups (Fig. 1a). Within each group, every ophthalmologist was presented with one of three randomly selected repeated responses from a particular chatbot to a given question. The ophthalmologists, in a blinded manner, then reviewed the responses of the four LLM chatbots based on their clinical experience. Prior to the ophthalmologists assessing the responses, the responses from chatbots (Table S2) underwent random shuffling. The response evaluation process was conducted in four rounds, with each round occurring on a different day and a 48-h washout period used between rounds to minimize potential carry-over effects [13]. Recognizing the potential bias in medical professionals' evaluations of readability, a Chinese readability platform was developed to assess the reading difficulty of the responses generated by the LLM chatbots objectively.

In the second phase of the study, a cohort of 46 representative patients was recruited from the ophthalmology outpatient department. The patients were randomized to one of two LLM groups: a GPT-4 group or a Baichuan 2 group that performed best in the in the first phase of the study. We invited each patient to take part in a dedicated patient education session. Prior to the patient-chatbot interaction, a general prompt was input to establish the contextual framework: "Please aid the ophthalmologist in conducting patient education on dry eye disease." Subsequently, the patients asked the chatbots three different questions about DED. The patients then assessed the satisfaction and readability of the chatbots' responses, and two ophthalmologists evaluated the responses across five domains (Table S4).

2.2. Study population

Eight ophthalmologists met the following inclusion criteria: (1) serving as chief physicians specializing in DED with a minimum of 5 years' clinical experience; (2) being native speakers of Mandarin and holding a Level II Grade A certificate or higher in the Mandarin Proficiency Test for China (MPTC); (3) possessing experience in evaluating the quality of medical information materials, such as patient educational materials and health consultation texts.

Forty-six patients met the following inclusion criteria: (1) > 18-year-old and <75-year-old; (2) native language is Mandarin; (3) newly diagnosed as DED in two weeks.

The exclusion criteria included: (1) patients who could not cooperate with the process; (2)history of cognitive impairment (such as Alzheimer's disease, severe mental disorders, etc.); (3) history of severe visual impairment (such as blindness)

2.3. Dry eye question bank

The development of the question bank was initiated by engaging six experienced dry eye experts (X. Zhou, X. Sun, C. Zhao, H. Le, A. Wei, and L. Tian), who were each tasked with recalling the most frequently asked questions by their patients suffering from dry eye, as well as the key health education. Subsequently, these questions and were systematically compiled and organized by Steven Liu, who harmonized the inputs from all participating experts into a cohesive pool of candidate questions. This step involved eliminating duplicates, grouping similar inquiries, and refining the language to maintain consistency and clarity. To finalize the selection of the 25 most pertinent questions, a single-blind methodology was employed. The consolidated list was then presented anonymously to the same six experts, who independently ranked the questions based on their perceived importance and relevance to patients with dry eye. This process mitigated potential biases and ensured that the chosen questions represented a consensual, expert-driven prioritization of the topics deemed most crucial for patient education and engagement.

2.4. Board-certified ophthalmologists' evaluation of five domains

In the first phase of the study, six independent board-certified ophthalmologists (X. Zhou, X. Sun, C. Zhao, H. Le, A. Wei, and L. Tian) were randomly assigned to three groups (Group A, Group B and Group C), and they independently reviewed each group (Group A, Group B and Group C) of responses to the questions in a blinded fashion. Ophthalmologists in each group (e.g., Group A) are assigned to review and provide feedback on the texts from the matching response group (also labeled Group A). The responses were evaluated in five domains: correctness (assessing the perceived accuracy of the response), completeness (assessing the perceived level of comprehensiveness of the response), readability (evaluating the ease of understanding of the response for patients), helpfulness (assessing the perceived usefulness of the response for patients), and safety (evaluating the potential of the response to mislead patients or potentially negatively influence in their treatment). Each domain was assessed using a 5-point Likert scale (Table S2).

In the real-world assessment, two ophthalmologists (J. Qiu and K. Shan) assessed the responses for correctness, completeness,

readability, usefulness, and safety based on the evaluation criteria. Each patient utilized a 5-point Likert scale to evaluate their satisfaction with the response (assessing their satisfaction level) and the readability (assessing the ease of understanding).

In instances where the opinions of two ophthalmologists were discordant, i.e., when the scores for a particular item differed by more than 2 points, Dr. Hong (Director, Dry Eye Center, Eye and Ear Nose Hospital, Fudan University, China) would provide a third-party rating based on the same criteria, and an average value would be computed. Moreover, discussion among the ophthalmologists was permitted until a unanimous scoring opinion was reached.

2.5. Objective readability analysis

Previous research has suggested that chatbots, when operating within an English language context, possess the capability to provide accurate responses. However, the responses are often complex, requiring readers to possess some form of tertiary education. To take account of differences between Chinese and English languages, as well as the presence of subjective factors, such as patients' educational backgrounds and physicians' experience, we used an online website, the Chinese readability platform (http://120.27.70.114:8000/analysis_a), for assessing readability [20]. Subsequently, we developed a program that automated 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 employs 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 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 indicated a lower level of text comprehensibility.

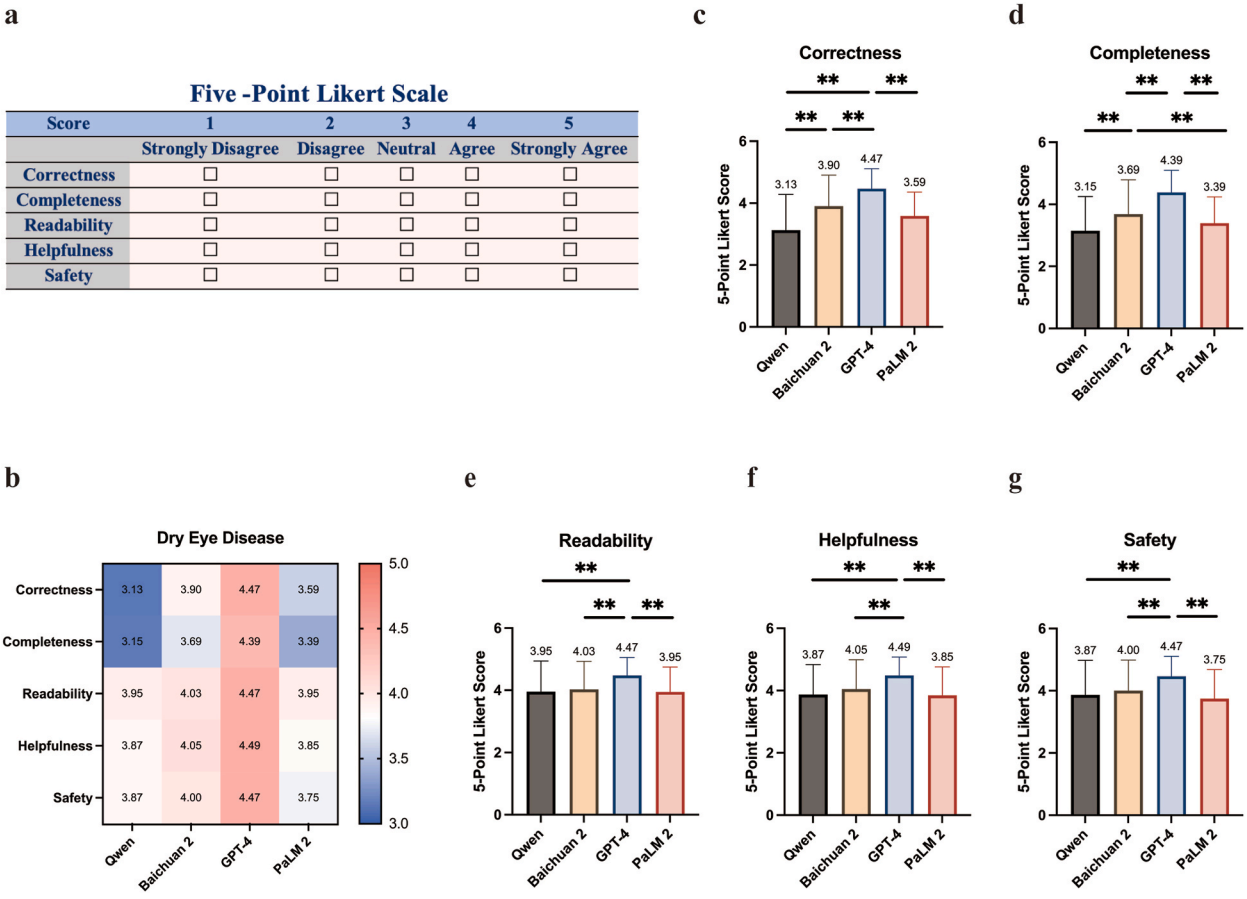


Fig. 2. Evaluation of the four LLM chatbot responses utilizing a 5-point Likert scale in the first phase of the study. (a) The scheme used to evaluate each response. (b)–(g) Average scores for the five domains (correctness, completeness, helpfulness, and safety) of the four LLM chatbot responses in the first phase of the study. Friedman's test and post hoc Dunnett tests were used to assess the statistical significance of the differences observed. Data are expressed as mean ± standard deviation. $*p < 0.05$ $**p < 0.01$.

2.6. Statistical analysis

The statistical analysis was performed using IBM SPSS Statistics for Mac version 25.0, which was released in August 2017 (IBM Corp., Armonk, NY). In first phase of the study, a Friedman test, a rank-based nonparametric test, was utilized to compare the evaluation scores and objective readability scores of each chatbot based on a 5-point Likert scale. Subsequently, post hoc Dunnett tests were employed for paired comparisons. In the second phase of the study, a two-tailed *t*-test was conducted to compare the average ratings of the responses of GPT-4 and Baichuan 2. The statistical significance level for all tests was set at $p \leq 0.05$.

3. Results

3.1. Retrospective cross-sectional study for test in the first phase

Fig. 2 displays the average scores across the five domains for the responses of the LLM chatbots when addressing DED-related questions. Among the four LLM chatbots evaluated, GPT-4 exhibited the highest level of proficiency in answering DED-related queries (correctness: 4.47; completeness: 4.39; readability: 4.47; helpfulness: 4.49; safety: 4.47, post hoc Dunnett test, $p < 0.05$) (Fig. 2). Table S2 provides a comprehensive breakdown of the scores for the five domains for each LLM chatbot's responses to individual questions.

Fig. 3b displays the passage statistics for every response generated by the LLM chatbots. Among the responses, those of GPT-4 had the highest reading difficulty (reading difficulty score: 12.86, post hoc Dunnett test, $p < 0.05$), indicating a higher recommended reading age (12.88 years old) (Fig. 3c and d). Thus, a higher level of education was required to understand the information (i.e., the chatbot's response) (Fig. 3e). Table S3 contains detailed information on the accumulated reading difficulty scores of each LLM chatbot's responses to individual questions.

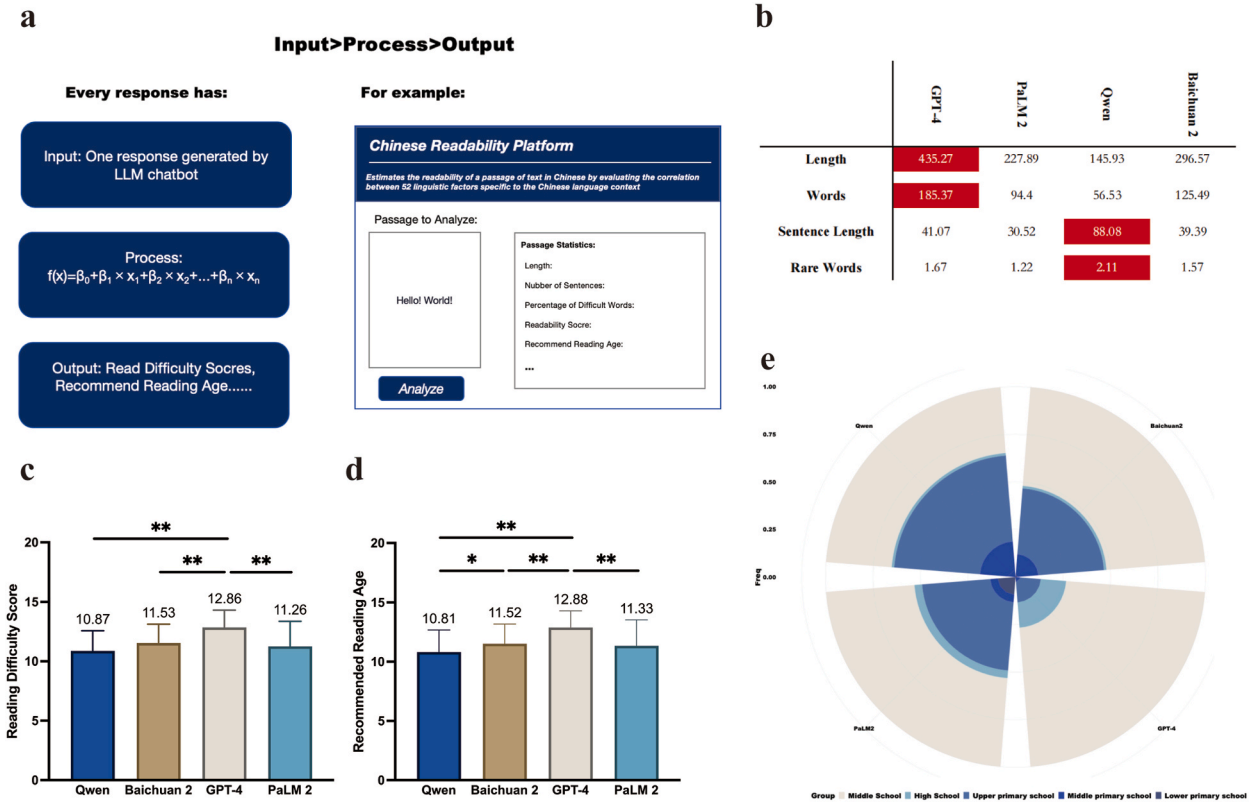


Fig. 3. Objective readability analysis of the four LLM chatbot responses in the first phase of the study. **(a)**The scheme of the Chinese readability analysis platform; **(b)** Passage statistics providing descriptive information about the responses generated by the four chatbots; **(c)** Reading difficulty scores representing the level of comprehension difficulty of the responses generated by the four chatbots; **(d)** Recommended reading age, indicating the age group for which the responses are suitable; **(e)** Literacy level corresponding to the passage generated by the four chatbots, depicting the readability level. The radius of each sector represents the frequency of occurrence. Friedman's test and post hoc Dunnett tests were used to assess the statistical significance of the differences observed. Data are expressed as mean ± standard deviation. *$p < 0.05$ **$p < 0.01$.

3.2. Real-world study for validation in the second phase

In the second phase of the study, 46 representative patients with DED posed questions to the two language models (GPT-4 and Baichuan 2) that had performed best in the first phase of the study. In terms of patient satisfaction and readability scores, the statistical analysis revealed a significant difference between GPT-4 and Baichuan 2 (satisfaction: 4.33 vs. 2.67, $p < 0.05$; readability: 2.67 vs. 4.67, $p < 0.05$) (Fig. 4b). The results of the ophthalmologists' assessments of the responses of the GPT-4 and Baichuan 2 chatbots were as follows for the five domains (correctness: 4.00 vs. 3.33; completeness: 4.33 vs. 3.33, $p < 0.05$; readability: 2.67 vs. 4.33, $p < 0.01$; helpfulness: 3.67 vs. 3.00; safety: 3.67 vs. 3.33). The readability scores for Baichuan 2 were higher than those for GPT-4 (Fig. 4c and d).

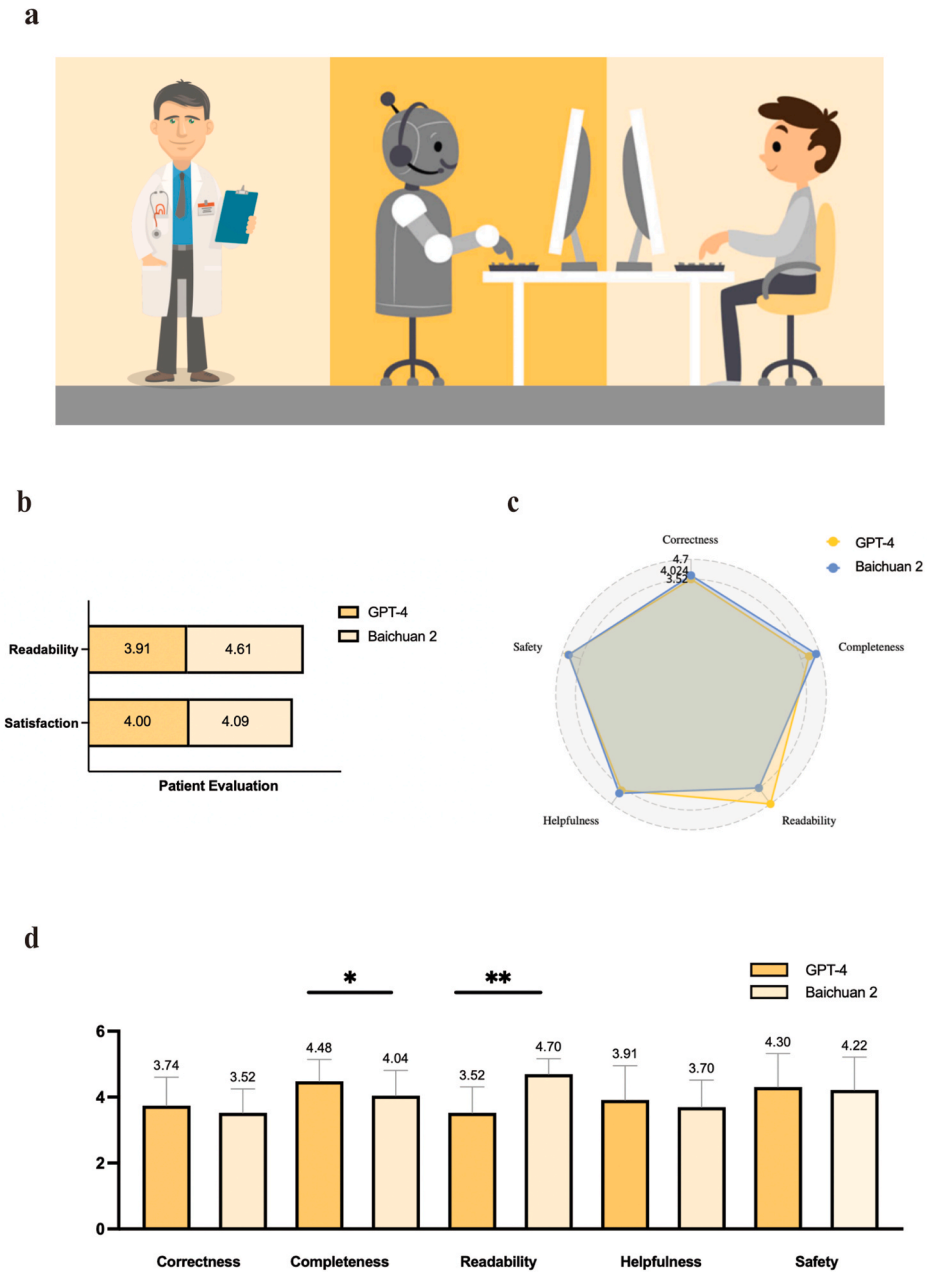


Fig. 4. Evaluation of GPT-4 and Baichuan 2 addressing DED patients queries in real-world assessment. **(a)** Human-chatbot interface; **(b)** Patient satisfaction with the responses of the four LLM chatbots and readability of the responses of the four chatbots in the real-world evaluation; **(c)** and **(d)** Average scores for the five domains (correctness, completeness, readability, helpfulness, and safety) of the responses of GPT-4 and Baichuan 2 in the real-world evaluation. A two-tailed *t* test was employed to assess the statistical significance of the differences observed. Data are expressed as mean ± standard deviation. *$p < 0.05$ **$p < 0.01$.

4. Discussion

In this study, we conducted a two-phase evaluation of the responses of GPT-4, PaLM 2, Qwen, and Baichuan 2 to frequent DED-related queries of Chinese patients. In the first phase of the study, we established a question bank consisting of 25 questions concerning DED in the context of patient education. Six ophthalmologists conducted evaluations in five domains. Across multiple domains (correctness, completeness, readability, helpfulness, and safety), when responding to patient education questions, GPT-4 performed best, followed by Baichuan 2. Overall, the evaluations of the ophthalmologists indicated that all four LLM chatbots have the potential to effectively address patient education needs concerning common ophthalmic conditions, with accurate and comprehensive responses. Recognizing that patients' understanding of medical information differs from that of physicians, we assessed the objective readability of the chatbots' responses. Our findings revealed that the responses generated by GPT-4 had the highest reading difficulty scores, making them more challenging to comprehend.

To assess the practical value of the chatbots in a real-world setting, we assessed the responses of the two chatbots that performed best to questions about DED posed by 46 patients. Based on the patients' assessments of satisfaction and readability, in addition to an assessment by two ophthalmologists of five domains, both GPT-4 and Baichuan 2 gave responses that were considered satisfactory, accurate, helpful, and safe, effectively addressing personalized DED education of Chinese patients. However, the responses of Baichuan 2 were relatively more comprehensible than those of GPT-4. This could be attributed to the utilization of a larger Chinese language corpus in Baichuan 2's training dataset [21,22]. To the best of our knowledge, our study is the first to explore the feasibility of LLM chatbots in dry eye education of Chinese patients in a real-world setting. Previous studies have predominantly relied on standardized testing methods or simulated patient-chatbot dialogues [14,15]. In contrast, we utilized LLM chatbots in real clinical settings, thereby exploring their practical applications in healthcare.

Chatbots often include disclaimers in their responses to patient queries. Examples include the following: "It is important to emphasize that the recommendations are intended solely for informational purposes. In the presence of corresponding symptoms, it is strongly advised that patients seek timely consultation with a healthcare professional." The chatbots in the present study performed similarly when queried about nonpharmacological treatment options for dry eye syndrome, with all four chatbots emphasizing in their responses that patients should follow measures under the guidance of a healthcare professional or seek prompt medical attention. The ability of the chatbots to mitigate risk or harmful behaviors by directing patients to seek professional advice meant the chatbots achieved higher safety scores in the evaluation of the five domains.

The accessibility of patient education materials is contingent upon their ability to facilitate comprehension among patients, while ensuring that the text's readability aligns with patients' reading and writing proficiency levels. In the objective analysis of readability, the responses generated by the four LLMs exhibited a moderate level of readability difficulty in Chinese, necessitating readers to possess a foundational knowledge equivalent to that of the ninth grade in junior high school. This contrasts with earlier research conducted by Momenaei et al. on chatbots' responses to questions posed in English. In their study, the educational material generated by GPT-4 exhibited a level of complexity that exceeded the grasp of the general public and required a collegiate or university education for comprehension [18]. Considering these disparities, we consulted with Dr. Zhang, a Ph.D. in linguistics, who suggested that the differences in findings between the two studies could be attributed to the use of different evaluation systems for assessing readability. Furthermore, despite the moderate readability difficulty scores observed in the responses generated by the four LLMs, it is important to acknowledge that less than 40 % of the Chinese population has an educational background equivalent to junior high school or below. Thus, these chatbot-generated Chinese texts remain inherently challenging to comprehend. This finding highlights the importance of ophthalmologists considering a patient's nationality, primary language, and cultural proficiency when selecting the most suitable chatbot for patient education.

In China, the prevalence of DED is high, and there is a limited availability of medical resources. The exceptional performance of specific LLM chatbots, such as GPT-4 and Baichuan 2, presents an opportunity to disseminate healthcare information and alleviate the burden on ophthalmic professionals. Several studies have evaluated the effectiveness of LLM chatbots, including ChatGPT, in providing assistance for medical queries at the primary care level [23,24]. First, large language models have a critical flaw, that is, hallucination, which refers to generating nonfactual statements. This flaw could potentially lead patients to make incorrect medical decisions. This raises the question of who bears responsibility for adverse outcomes stemming from ChatGPT's advice. Second, there is currently a scarcity of medical-specific LLMs, and the integration of LLMs to assist patient education in medical institutions is still an unresolved matter. Furthermore, other issues, such as patient accessibility, the conflict between readability and educational context, overreliance on LLMs, user privacy, and data protection all demand careful consideration.

In practice, despite LLM chatbots demonstrating remarkable zero/few-shot inference performance in most natural language processing tasks, their current limitations confine them to discrete text-based interactions. In clinical settings, patients often require ophthalmologists to address queries pertaining to the interpretation of visual data, including diverse modalities, such as optical coherence tomography and fundus photographs. As a result, the multimodal nature of clinical data goes beyond textual information, necessitating the development of multimodal LLMs that align more effectively with the practical clinical landscape. This represents a significant avenue for future advancement [25,26]. Moreover, there is a stark difference between the ways in which physicians and patients communicate and patients and chatbots communicate. Although LLMs offers the convenience of expediency, LLMs frequently fail to provide contextually tailored information that accommodates the unique circumstances of individual patients [27]. Furthermore, patients frequently introduce uncertainty into inquiries/dialogue due to the ways in which they prompt chatbots and the ways in which they process medical information. The risk of incorrect medical decision making by patients increases in the nonprofessional and psychologically atypical context of patient–chatbot interactions [28,29].

Several limitations should be noted in the current study. First, our test question bank encompassed only the 25 most common

questions from Chinese patients with DED, whereas DED comprises a wide range of conditions. Second, only a small number of ophthalmologists evaluated the responses generated by LLM chatbots. It is possible that other ophthalmologists may hold different opinions regarding the accuracy and utility of the recommendations generated by the LLM chatbots evaluated in the present study. Thus, the design and conduct of large-scale, multi-center, prospective studies are essential to rigorously assess the accuracy, generalizability, and clinical utility of LLMs. We emphasize the importance of such investigations in establishing standardized protocols, ensuring data quality, and incorporating diverse populations to enhance the external validity of the findings. Third, our test questions in the first phase of the study adopted a short-answer format, which was more aligned with medical dialogue scenarios compared to the multiple-choice questions that are prevalent in medical test datasets. However, in real-world settings, patients usually prefer to acquire information through continuous dialogues rather than isolated inquiries. Furthermore, existing testing protocols mainly rely on single-task systems, which inherently lack adequate expressive and interactive capabilities. As a result, a disparity exists between current evaluation procedures and the expected performance within the actual clinical workflow. Finally, the real-world dataset included only 46 patients, and we restricted the number of questions asked to three questions per patient. In doctor–patient relationships, conversations typically take place without such limitations and in the context of patients' medical histories. To improve the generalizability and novelty of our research findings, we are planning further research with a larger clinical cohort.

5. Conclusions

Our two-phase evaluation study offers empirical evidence of the potential of using LLMs, particularly GPT-4 and Baichuan 2, to meet the educational needs of Chinese patients with DED. The information provided by the LLM chatbots is relatively accurate, complete, helpful, and safe. However, it is crucial to acknowledge that the introduction of any new technology brings both opportunities and risks. Therefore, it is imperative to continue to focus on the application of LLMs in the field of DED and ophthalmology, exploring strategies and conducting assessments to refine and establish the effectiveness of these tools. Such work will be critical in advancing the utilization of LLMs and maximizing their benefits in the context of personalized patient education in ophthalmology.

Financial support

This work was supported by the National Natural Science Foundation of China (81970766 and 82171102), the National Key Research and Development Program of China (2023YFA0915000), and the Shanghai Medical Innovation Research Program (22Y21900900).

Data availability statement

Data included in article and supplementary material.

Code availability

The algorithms utilized in the automated analysis platform for Chinese readability were tailored specifically to our developmental milieu, serving primarily the functions of data ingress/egress and computational parallelization across multiple systems.

Ethics approval and consent to participate

The study was approved by the Institutional Review Board of the Eye and ENT Hospital of Fudan University and followed the tenet of the Declaration of Helsinki. All patients gave written informed consent (IRB-EENT-2020124).

Availability of data and materials

All the data are included herein (main text and supplementary section).

Funding

This work was supported by the National Natural Science Foundation of China (82171102, 81970766, 82271044), the National Key Research and Development Program of China (2023YFA0915000), the Shanghai Medical Innovation Research Program (22Y21900900) and the Shanghai Key Clinical Research Program (SHDC2020CR3052B).

CRediT authorship contribution statement

Runhan Shi: Writing – review & editing, Writing – original draft, Software, Investigation, Formal analysis. **Steven Liu:** Software, Investigation. **Xinwei Xu:** Project administration, Methodology, Investigation. **Zhengqiang Ye:** Investigation. **Jin Yang:** Investigation. **Qihua Le:** Investigation. **Jini Qiu:** Investigation. **Lijia Tian:** Investigation. **Anji Wei:** Investigation. **Kun Shan:** Investigation. **Chen Zhao:** Writing – review & editing, Validation, Supervision. **Xinghuai Sun:** Writing – review & editing, Validation, Supervision. **Xingtao Zhou:** Writing – review & editing, Validation, Supervision. **Jiaxu Hong:** Writing – review & editing, Validation, Supervision,

Funding acquisition.

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.

Acknowledgments

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 the 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.heliyon.2024.e34391.

References

[1] A.J. Thirunavukarasu, D.S.J. Ting, K. Elangovan, L. Gutierrez, T.F. Tan, D.S.W. Ting, Large language models in medicine, Nature Med 29 (8) (2023) 1930–1940.
[2] W.X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.Y. Nie, J.R. Wen, A survey of large Language Models, arXiv:2303.18223 [cs.CL] (2023).
[3] OpenAI, ChatGPT: optimizing Language Models for dialogue, Retrieved April 25, 2023, from, https://openai.com/blog/chatgpt/, 2022.
[4] T. Susnjak, ChatGPT: the End of Online Exam Integrity? arXiv Preprint arXiv:2212.09292, 2022.
[5] C.E. Haupt, M. Marks, AI-Generated medical advice-GPT and beyond, JAMA 329 (16) (2023) 1349–1350.
[6] L. De Angelis, F. Baglivo, G. Arzilli, et al., ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health, Front. Public Health 11 (2023).
[7] R. Gozalo-Brizuela, E.C. Garrido-Merchan, ChatGPT is not all you need. A State of the Art Review of large Generative AI models, arXiv preprint arXiv: 2301.04655 (2023).
[8] F. Stapleton, M. Alves, V.Y. Bunya, et al., TFOS DEWS II epidemiology report, Ocul. Surf. 15 (3) (2017) 334–365.
[9] X. Wang, H.M. Sanders, Y. Liu, et al., ChatGPT: promise and challenges for deployment in low- and middle-income countries, Lancet Reg Health West Pac 41 (2023) 100905.
[10] R. Calixte, A. Rivera, O. Oridota, W. Beauchamp, M. Camacho-Rivera, Social and demographic patterns of health-related internet use among adults in the United States: a secondary data analysis of the health information national trends survey, Int. J. Environ. Res. Publ. Health 17 (18) (2020) 6856.
[11] I.A. Bernstein, Y.V. Zhang, D. Govil, et al., Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions, JAMA Netw. Open 6 (8) (2023) e2330320, https://doi.org/10.1001/jamanetworkopen.2023.30320. Published 2023 Aug 1.
[12] Y. Shen, L. Heacock, J. Elias, et al., ChatGPT and other large Language Models are double-edged swords, Radiology 307 (2) (2023) e230163.
[13] Z.W. Lim, K. Pushpanathan, S.M.E. Yew, et al., Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard, EBioMedicine 95 (2023) 104770.
[14] M.L.R. Rasmussen, A.C. Larsen, Y. Subhi, I. Potapenko, Artificial intelligence-based ChatGPT chatbot responses for patient and parent questions on vernal keratoconjunctivitis, Graefes Arch. Clin. Exp. Ophthalmol. 261 (10) (2023) 3041–3043.
[15] S. Singh, A. Djalilian, M.J. Ali, ChatGPT and ophthalmology: exploring its potential with discharge summaries and operative notes, Semin. Ophthalmol. 38 (5) (2023) 503–507.
[16] A. Mihalache, M.M. Popovic, R.H. Muni, Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment, JAMA Ophthalmol 141 (6) (2023) 589–597.
[17] L.Z. Cai, A. Shaheen, A. Jin, et al., Performance of generative large Language Models on ophthalmology board-style questions, Am. J. Ophthalmol. 254 (2023) 141–149.
[18] B. Momenaei, T. Wakabayashi, A. Shahlaee, et al., Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases, Ophthalmol Retina 7 (10) (2023) 862–868.
[19] Z. Ying, Y. Fan, J. Lu, P. Wang, L. Zou, Q. Tang, Y. Chen, X. Li, Y. Chen, Exploration of ChatGPT application in diabetes education: a multi-dataset, multi-reviewer study, medRxiv (September 27, 2023). Published online.
[20] Y. Cheng, D.K. Xu, J. Dong, Key factors analysis and readability formula research based on Chinese textbook corpus text reading difficulty grading, Language and Text Application (1) (2020) 132–143 [in Chinese].
[21] L. Ouyang, J. Wu, X. Jiang, et al., Training Language Models to Follow Instructions with Human Feedback. arXiv Preprint arXiv:2203.02155, 2022.
[22] A. Yang, B. Xiao, B. Wang, et al., Baichuan 2: Open Large-Scale Language Models, 2023 arXiv preprint arXiv:2309.10305.
[23] M. Balas, E.B. Ing, Conversational AI models for ophthalmic diagnosis: comparison of ChatGPT and the Isabel Pro differential diagnosis generator, JFO Open Ophthalmology. 1 (2023).
[24] K. Pushpanathan, Z.W. Lim, S.M. Er Yew, et al., Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries, iScience 26 (11) (2023) 108163.
[25] Y. Zhang, H. Jiang, Y. Miura, C.D. Manning, C.P. Langlotz, Contrastive learning of medical visual representations from paired images and text, arXiv preprint arXiv:2010.00747 (2022).
[26] H.Y. Zhou, Y. Yu, C. Wang, et al., A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics, Nat. Biomed. Eng. 7 (6) (2023) 743–755.
[27] S.Y. Lu, Risk Identification of Online Medical Information Acquisition and Research on Network Governance Strategies [Dissertation], Zhejiang University, 2018 [Chinese].
[28] W.X. Zhao, K. Zhou, J. Li, et al., A Survey of Large Language Models, 2023 arXiv preprint arXiv:2303.18223.
[29] A.D. Saenz, Z. Harned, O. Banerjee, M.D. Abràmoff, P. Rajpurkar, Autonomous AI systems in the face of liability, regulations and costs, NPJ Digit Med 6 (1) (2023) 185. Published 2023 Oct 6.