Readability and Appropriateness of Responses Generated by ChatGPT 3.5, ChatGPT 4.0, Gemini, and Microsoft Copilot for FAQs in Refractive Surgery

Fahri Onur Aydın, Burakhan Kürşat Aksoy, Ali Ceylan, Yusuf Berk Akbaş, Serhat Ermiş, Burçin Kepez Yıldız, Yusuf Yıldırım

University of Health Sciences Türkiye, Başakşehir Çam and Sakura City Hospital, Clinic of Ophthalmology, İstanbul, Türkiye

Abstract

Objectives: To assess the appropriateness and readability of large language model (LLM) chatbots' answers to frequently asked questions about refractive surgery.

Materials and Methods: Four commonly used LLM chatbots were asked 40 questions frequently asked by patients about refractive surgery. The appropriateness of the answers was evaluated by 2 experienced refractive surgeons. Readability was evaluated with 5 different indexes.

Results: Based on the responses generated by the LLM chatbots, 45% (n=18) of the answers given by ChatGPT 3.5 were correct, while this rate was 52.5% (n=21) for ChatGPT 4.0, 87.5% (n=35) for Gemini, and 60% (n=24) for Copilot. In terms of readability, it was observed that all LLM chatbots were very difficult to read and required a university degree.

Conclusion: These LLM chatbots, which are finding a place in our daily lives, can occasionally provide inappropriate answers. Although all were difficult to read, Gemini was the most successful LLM chatbot in terms of generating appropriate answers and was relatively better in terms of readability.

Keywords: Artificial intelligence, chatbots, refractive surgery FAQs, ChatGPT, Gemini, Copilot

Cite this article as: Aydın FO, Aksoy BK, Ceylan A, Akbaş YB, Ermiş S, Kepez Yıldız B, Yıldırım Y. Readability and Appropriateness of Responses Generated by ChatGPT 3.5, ChatGPT 4.0, Gemini, and Microsoft Copilot for FAQs in Refractive Surgery. Turk J Ophthalmol. 2024;54:313-317

Address for Correspondence: Fahri Onur Aydın, University of Health Sciences Türkiye, Başakşehir Çam and Sakura City Hospital, Clinic of Ophthalmology, İstanbul, Türkiye
E-mail: onuraydin90@hotmail.com **ORCID-ID:** orcid.org/0000-0001-9414-6788

Received: 16.07.2024 **Accepted:** 11.11.2024

DOI: 10.4274/tjo.galenos.2024.28234

Introduction

The rapid integration of artificial intelligence (AI) into healthcare has transformed patient engagement and information dissemination. As AI models increasingly become a primary source of medical information, it is essential to evaluate the feasibility and accuracy of their responses to medical queries.[1,2] The rise of conversational robots, driven by advancements in natural language processing, marks a promising new era in the healthcare industry. These robots show remarkable potential in various medical fields, including disease prevention, diagnosis, treatment, monitoring, and patient support.[3]

Large language model (LLM) chatbots, such as OpenAI's ChatGPT, Google's Gemini, and Microsoft's Copilot, represent a significant leap forward in AI technology. These models are designed to generate human-like responses to a variety of text-based queries, leveraging extensive training data and sophisticated algorithms.[4] The evolution of LLM chatbots, characterized by self-supervised learning and training on vast textual data, has enabled them to produce responses that closely mimic human interactions. Their ability to provide detailed and relevant information makes them particularly valuable for medical applications.[5,6]

In the field of ophthalmology, especially in refractive surgery, patients often turn to the internet to obtain information about their conditions and treatment options. The quality and readability of this information are crucial, as they directly impact patient comprehension and decision-making. Despite the potential benefits of LLM chatbots in providing medical advice, their effectiveness in delivering accurate and understandable information still requires a thorough assessment.

This study aimed to explore the strengths and limitations of different LLM chatbots in providing reliable and accessible information about refractive surgery. By evaluating the relevance and readability of their responses, this research seeks to enhance

AI-driven patient education, thereby ensuring that patients receive accurate and comprehensible information to make informed decisions about their eye health.

Materials and Methods

Approval from the ethics committee was not required since no patients were involved in our study.

This study was designed to investigate the appropriateness and readability of the information provided by LLM chatbots. Four newly developed and frequently used LLM chatbots were selected: ChatGPT 3.5, ChatGPT 4.0, Google Gemini, and Microsoft Copilot. Refractive surgeons were encouraged to compile a list of 40 questions about refractive surgery that patients frequently ask either through the patient portal or in the clinic. These questions were then answered by the LLM chatbots on July 3, 2024. The answers were evaluated for appropriateness and adequacy by two experienced refractive surgeons (Y.Y., B.K.Y.). The answers were categorized as "appropriate", "incomplete", and "inappropriate". An appropriate response was defined as a correct answer that was similar to the recommendations that the reviewer would give patients. An inappropriate response was either inaccurate or differed from the reviewer's recommendation in a clinical setting. An incomplete response was relevant and accurate but did not provide enough information.

To assess the ease of reading each answer for the average person, we entered the answers into an online readability application called Readable (https://app.readable.com/text/).[7] The readability and understandability criteria and standardization we used in the study were based on English. In our study, we formulated the questions in English and received answers in English. Five different indices were used to evaluate the readability of each response: the Gunning Fog Index, Coleman-Liau Index, Flesch Reading Ease Score, Flesch-Kincaid Grade Level, and Simple Measure of Gobbledygook (SMOG) Index.[8] The mathematical formulae used in Flesch reading tests are based on word complexity and sentence length. The Flesch Reading Ease score is a numerical value between 1 and 100. Higher numbers indicate more readability, and a score between 70 and 80 corresponds to an 8th-grade level.[7] The Gunning Fog Index evaluates the frequency of polysyllabic words along with the average sentence length.[9] This index score, which ranges from 0 to 20, rates simplicity and clarity.[7] The Coleman-Liau Index helps assess medical data and is typically used in conjunction with other indices.[1] It focuses on the mean length of sentences and the mean number of letters per hundred words.[9] The SMOG Index uses the frequency of polysyllabic words in a sample of sentences.[9] Although widely used, SMOG is most frequently applied in healthcare.[10] The results of the latter three indices correspond to the grade level at which a student must be studying in the United States in order to comprehend the written material. Thus, texts with lower Gunning Fog, Coleman-Liau, and SMOG index values should be easier to read and understand.[11]

Statistical Analysis

Statistical analysis was performed using the SPSS program (IBM SPSS Statistics, version 25; IBM Corp., Armonk, NY, USA). Descriptive analysis and normality distribution test (Shapiro-Wilk) were performed. Considering the abnormal distribution of the data, a non-parametric Kruskal-Wallis test and Bonferroni correction were performed to compare mean scores across the four LLM chatbots. An adjusted p value less than 0.05 was considered statistically significant.

Results

Appropriateness

Based on the responses generated by the LLM chatbots, 45% (n=18) of the answers given by ChatGPT 3.5 were correct, while 52.5% (n=21) of ChatGPT 4.0, 87.5% (n=35) of Gemini, and 60% (n=24) of Copilot answers were correct. ChatGPT 3.5, ChatGPT 4.0, and Copilot gave inappropriate answers to one question each, while Gemini did not give inappropriate answers to any question (Figure 1).

The LLM chatbots showed a statistically significant difference when compared in terms of appropriateness (p=0.001). When subgroup analysis was performed, this difference was observed between Gemini and ChatGPT 3.5 and 4.0 (p=0.001, p=0.008 respectively) (Table 1).

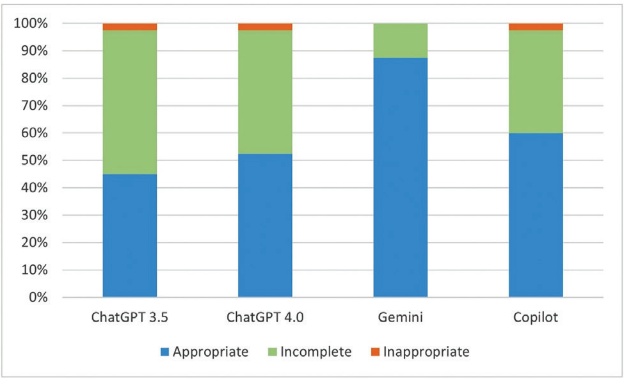

Figure 1. Consensus-based accuracy ratings of large language model chatbot responses to questions about refractive surgery, as determined by two experienced refractive surgeons

Table 1. Overview of the appropriateness and length of large language model chatbots' responses to questions about refractive surgery

	ChatGPT 3.5	ChatGPT 4.0	Gemini	Copilot	p value
Appropriateness	2.42±0.54	2.50±0.55	2.87±0.33	2.57±0.54	0.001
Word count	21.15±4.40	21.67±6.24	318.62±73.98	103.90±46.44	<0.001
Character count	115.00±24.33	118.65±32.97	1767.02±450.00	587.15±260.76	<0.001

Readability

Readability indices are summarized in Table 2. A significant difference among the LLM chatbots was observed when compared according to Flesch-Kincaid Grade Level (p=0.003). Pairwise evaluations revealed this difference to be between ChatGPT 3.5 and Gemini and between ChatGPT 3.5 and Copilot, with ChatGPT 3.5 having significantly higher values (p=0.017 and p=0.008, respectively; Figure 2a). No significant difference was observed between the other chatbots. There were no significant differences among the chatbots in terms of Flesch Reading Ease scores (p=0.534; Figure 2b) or Coleman-Liau score (p=0.867; Figure 2c). When the SMOG index was compared, a significant difference was observed between the chatbots (p=0.012). This was found to be a result of a significantly lower SMOG value for Copilot compared to ChatGPT 3.5 (Figure 2d). A significant difference was again observed between the groups when Gunning Fog scores were evaluated (p=0.001). Pairwise comparisons showed that Copilot had a significantly lower score than both ChatGPT 3.5 and ChatGPT 4.0 (p=0.003 and 0.021, respectively) (Figure 2e).

Table 2. Readability indices for large language model chatbots' responses to frequently asked questions about refractive surgery

	ChatGPT 3.5	ChatGPT 4.0	Gemini	Copilot	p value
Coleman-Liau	14.86±3.90	14.99±5.11	14.60±1.58	14.88±2.51	0.867
Flesch Reading Ease score	30.97±22.49	31.79±26.72	37. 39±7.98	32.76±12.88	0.534
Flesch-Kincaid Grade Level	13.95±3.38	13.49±3.92	11.76±1.35	11.71±1.89	0.003
SMOG Index	15.03±2.80	14.50±3.07	13.93±1.35	13.34±1.28	0.012
Gunning Fog	16.30±3.96	15.74±4.73	14.09±1.65	13.48±2.08	0.001

SMOG: Simple Measure of Gobbledygook

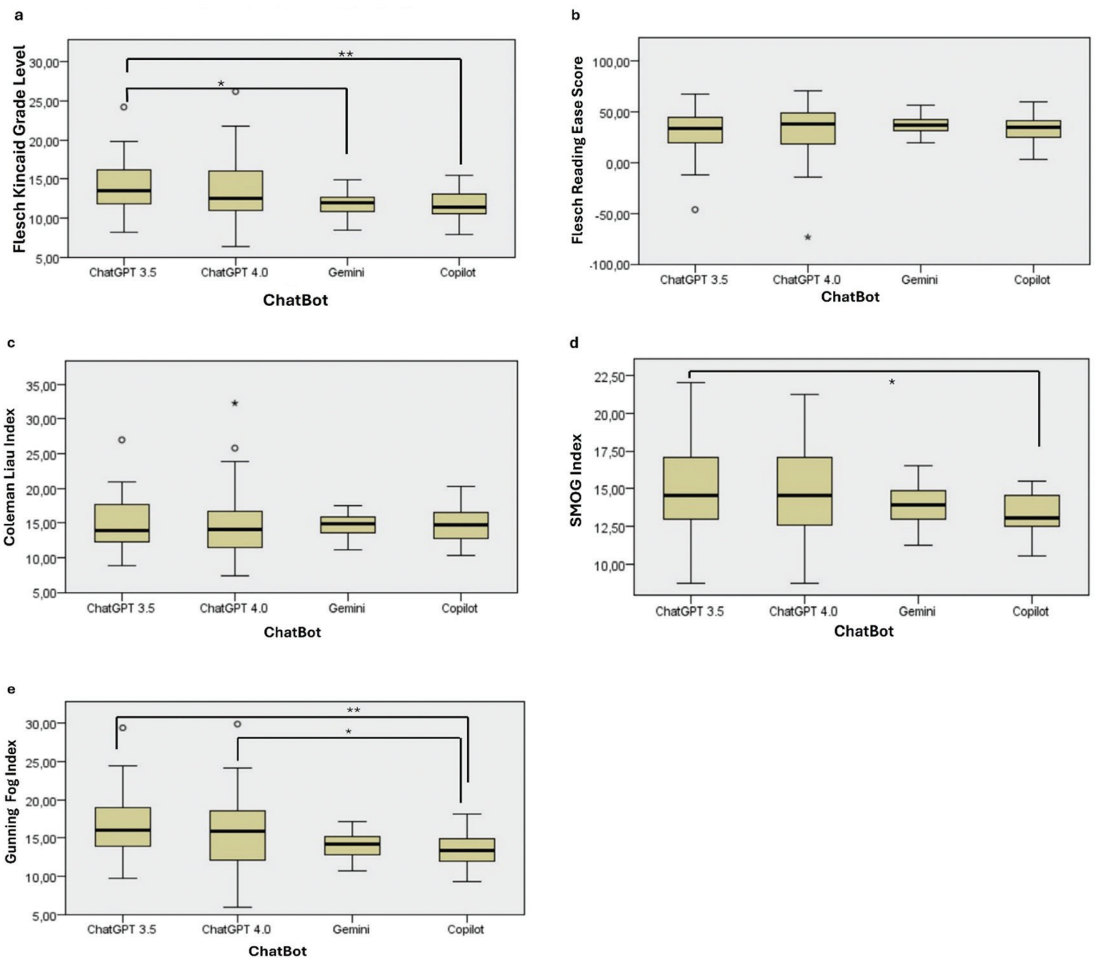

Figure 2. The scores of large language model chatbots in terms of readability shown on a boxplot. a) Flesch-Kincaid Grade Level, b) Flesch Reading Ease score, c) Coleman-Liau Index, d) Simple Measure of Gobbledygook (SMOG) Index, e) Gunning Fog Index

A comparison of word and character counts showed that Gemini had significantly higher values than the other LLM chatbots ($p<0.001$ for both). Word and character counts were significantly higher for Gemini compared to Copilot ($p=0.001$ for both) and both ChatGPT 3.5 and 4.0 ($p<0.001$ for all). The ChatGPT versions had comparable word and character counts (Table 1).

Discussion

The use of AI is becoming increasingly widespread worldwide. With its increasing use, many new AI models are being developed. These include language models trained to use learned data to browse the internet and produce immediate responses in chatbot conversations.[12] This article presents an in-depth analysis of how this variation affects LLM chatbot performance and response quality, highlighting that differences between the responses of different LLM chatbots are mainly due to differences in the algorithms used.

Today, many people use LLM chatbots for various purposes. One of them is to get answers to their questions in the field of health. However, using AI to get health-related information can cause several problems. These include obtaining outdated or inaccurate information and misunderstanding correct information that is presented in a complex way. Therefore, it is very important that this information is both accurate and understandable by everyone.

In our study, when the appropriateness of the chatbots' responses was evaluated, it was observed that Gemini answered the questions correctly at a significantly higher rate than the other LLM chatbots. In contrast to our findings, Tepe and Emekli[13] reported in a study comparing ChatGPT 4.0, Gemini, and Copilot that ChatGPT 4.0 gave the most appropriate answers to questions about breast imaging. In another study, Lee et al.[14] compared Gemini and ChatGPT 3.5 as sources for hypertension education and determined that they provided similar results.

In our study, five different recognized readability indices were used to provide comprehensive results. According to these indices, the responses generated by the LLM chatbots had low readability scores. Flesch Reading Ease scores ranged from 30 to 50, with Gemini having the highest score (i.e., the most readable answers). This suggests that the texts could be understood by university students and the level of difficulty was suitable for only 33% of adults.[15] In terms of Flesch-Kincaid Grade Level, ChatGPT responses were found to be suitable for people in grade 14 and above, while Gemini and Copilot were appropriate for those in grade 12 and above, suggesting that Gemini and Copilot had slightly better readability.[16] The Gunning Fog Index also indicated a university level for all of the LLM chatbots. However, it was observed that ChatGPT responses were at the level of senior undergraduate students, while Gemini and Copilot were at the freshman level. The Coleman-Liau Index was similar for all LLM chatbots, indicating an undergraduate level that was difficult to read.[17] Similarly, SMOG Index values for all LLM chatbots showed their responses were at the undergraduate level and difficult to read for the general majority.[10]

In a study conducted with ChatGPT 4.0, the results of readability analyses were similar to those in our study, indicating an undergraduate or graduate level that was fairly difficult to read.[18] In the study conducted by Tepe and Emekli[13], comparison of ChatGPT 4.0, Gemini, and Copilot in terms of readability revealed that ChatGPT 4.0 was the most difficult and Gemini was relatively easier, but all had low readability.

When the number of words and characters were evaluated, it was observed that both ChatGPT versions used a significantly lower number of words and characters than the other LLM chatbots. Gemini used the highest number of words and characters. Despite being significantly longer, Gemini responses showed better readability and accuracy.

Although similar methodology has been used in other studies on LLM chatbots in the literature, a more holistic evaluation may be possible if a patient cohort is used. This idea may guide future research.

As the results show, LLM chatbots may provide incomplete or occasionally incorrect information. In addition, even if the information they provide is correct, there is also the possibility of misleading patients due to its relatively low readability. This poses a potential risk for patients. To reduce these possibilities, new LLM chatbots developed in collaboration with healthcare professionals specifically for health-related information may be beneficial in improving accuracy and accessibility.

Study Limitations

This study has several limitations. First, the search was limited to 40 questions, which may limit the generalizability of the findings. In addition, the formulation of inputs when interacting with LLM chatbots can significantly affect the quality and nature of the responses produced. The repeatability of LLM chatbots is also questionable. In this study, each question was sent to the LLM chatbots only once. Furthermore, when assessing the readability of the answers, the absence of real patients as evaluators is another limitation of the study.

Conclusion

In conclusion, we observed that Gemini was better than other LLM chatbots in giving appropriate answers to questions about refractive surgery. In terms of readability, we found that all chatbot responses were difficult to read, but Gemini and Copilot were relatively more readable. As a result, when the responses of the LLM chatbots were compared, it was seen that Gemini was the best in terms of both relevance and readability, while ChatGPT 3.5 was the worst. It is worth reminding our patients that these LLM chatbots can give inappropriate answers, albeit rarely.

Ethics

Ethics Committee Approval: Not required.

Informed Consent: Not required.

Declarations

Authorship Contributions

Concept: F.O.A., S.E., Y.B.A., B.K.Y., Design: F.O.A., A.C., Y.B.A., Y.Y., Data Collection or Processing: B.K.A., A.C., S.E., Analysis or Interpretation: A.C., B.K.Y., Y.Y., Literature Search: F.O.A., B.K.A., B.K.Y., Writing: F.O.A., B.K.A., A.C., Y.B.A., S.E., B.K.Y., Y.Y.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Xu L, Sanders L, Li K, Chow JC. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR Cancer. 2021;7:27850.
2. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, Rajpurkar P. Foundation models for generalist medical artificial intelligence. Nature. 2023;616:259-265.
3. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, Rizzo C. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health. 2023;11:1166120.
4. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388:1233-1239.
5. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, Cielo D, Oyelese AA, Doberstein CE, Telfeian AE, Gokaslan ZL, Asaad WF. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery. 2023;93:1090-1098.
6. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, Shue A, Chou JC, Schehlein E, Christopher KL, Groth SL, Ludwig C, Wang SY. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open. 2023;6:2330320.
7. Readability is an essential content marketing tool. Readable. Available from: https://readable.com/readability/#goodscore. Accessed April 26, 2023.
8. Patel AJ, Kloosterboer A, Yannuzzi NA, Venkateswaran N, Sridhar J. Evaluation of the content, quality, and readability of patient accessible online resources regarding cataracts. Semin Ophthalmol. 2021;36:384-391.
9. Basch CH, Mohlman J, Hillyer GC, Garcia P. Public health communication in time of crisis: readability of on-line COVID-19 information. Disaster Med Public Health Prep. 2020;14:635-637.
10. Hedman AS. Using the SMOG formula to revise a health-related document. Am J Health Educ. 2008;39:61-64.
11. Robinson E, McMenemy D. 'To be understood as to understand': a readability analysis of public library acceptable use policies. J Librarians Inf Sci. 2020;52:713-725.
12. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:0000198.
13. Tepe M, Emekli E. Assessing the responses of large language models (ChatGPT-4, Gemini, and Microsoft Copilot) to frequently asked questions in breast imaging: a study on readability and accuracy. Cureus. 2024;16:59960.
14. Lee TJ, Campbell DJ, Patel S, Hossain A, Radfar N, Siddiqui E, Gardin JM. Unlocking health literacy: the ultimate guide to hypertension education from ChatGPT versus Google Gemini. Cureus. 2024;16:59898.
15. DuBay WH. The principles of readability. 2014. Available from: https://files.eric.ed.gov/fulltext/ED490073.pdf
16. Flesch R. How To Write Plain English: A Book For La Wyers And Consumers. 2014. Available from: https://books.google.com.tr/books/about/How_to_Write_Plain_English.html?id=-kpZAAAAMAAJ&redir_esc=y
17. Coleman M, Liau TL. A computer readability formula designed for machine scoring. J Appl Psychol. 1975;60:283-284.
18. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, Mansour HA, Abishek RM, Xu D, Sridhar J, Yonekawa Y, Kuriyan AE. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. Ophthalmol Retina. 2023;7:862-868.