EUROPEAN EYE RESEARCH

DOI: 10.14744/eer.2025.44154
Eur Eye Res 2025;5(2):103–110

ORIGINAL ARTICLE

Evaluating the accuracy, readability, and relevance of answers generated by large language models for frequently asked questions about cataract and cataract surgery

Ayse Bozkurt Oflaz, Sule Acar Duyan
Department of Ophthalmology, Selcuk University, Konya, Türkiye

Abstract

Purpose: To evaluate the accuracy, relevance, and readability of large language models (LLMs) such as ChatGPT-3.5, ChatGPT-4o, Gemini, and Copilot in answering frequently asked questions (FAQs) about cataract and cataract surgery.

Methods: Ten FAQs about cataract and cataract surgery were answered by LLMs. The respondents scored the answers for accuracy and readability. Two experienced cataract surgeons assessed the accuracy of the answers. Flesch reading ease score, flesch-kincaid grade level, gunning fog index, Coleman-Liau Index and simple measure of gobbledygook index were used for readability.

Results: According to expert assessment, the rates of "correct and complete" answers were: ChatGPT-3.5 (81%), ChatGPT-4o (100%), Gemini (98%), and Copilot (54%), with a statistically significant difference among the models ($P < 0.0001$). Post hoc comparisons showed that ChatGPT-4o and Gemini outperformed ChatGPT-3.5 ($P = 0.0005$ and $P = 0.0079$, respectively). Significant differences were also found in word and sentence counts across models ($P < 0.0001$). No statistically significant differences were observed in readability scores.

Conclusion: ChatGPT-4o and Gemini provided more accurate responses. However, no significant difference was observed in readability across models, emphasizing the need for algorithmic improvements to enhance comprehensibility in artificial intelligence-generated patient education content.

Keywords: Artificial intelligence; cataract; chatbots; ChatGPT; copilot; gemini.

Cataracts are one of the most common causes of vision loss worldwide and a leading cause of preventable blindness. Cataract surgery is recognized as one of the most common surgical procedures in ophthalmology and is usually performed with high success rates.[1] Patients undergoing cataract surgery often have questions about the causes of cataracts, the necessity of surgery, the surgical procedure, and the recovery period. Providing accurate and understandable information is essential to patient education.[2]

In recent years, artificial intelligence (AI)-driven tools, such as large language models (LLMs), have been used to extract medical information. These models can answer medical questions and facilitate patient education thanks to their natural language processing capabilities.[3] However, there is a lack of data on the appropriateness and comprehensibility of the responses generated by these technologies.[4] Both accuracy and comprehensibility of answers are essential factors for patients. Conveying medical information using complex language and long

Cite this article as: Bozkurt Oflaz A, Acar Duyan S. Evaluating the accuracy, readability, and relevance of answers generated by large language models for frequently asked questions about cataract and cataract surgery. Eur Eye Res 2025;5(2):103–110.

Correspondence: Ayse Bozkurt Oflaz, M.D. Department of Ophthalmology, Selcuk University, Konya, Türkiye
E-mail: draysebozkurtoflaz@yahoo.com

Submitted Date: 31.01.2025 **Revised Date:** 28.05.2025 **Accepted Date:** 06.06.2025 **Available Online Date:** 26.08.2025

sentences may prevent patients from understanding and lead to misinterpretations.[5] Evaluating the relevance of the outputs generated by large multimodal models (LMMs) facilitates an understanding of the reliability of these models in terms of medical accuracy and proficiency.[6] The extant literature offers a paucity of studies on the readability of frequently asked questions (FAQs) about cataracts. Nevertheless, it is essential to acknowledge the importance of patient comprehension of these documents in the context of health literacy. Such content must be comprehensible and readily accessible to patients, ensuring they can locate the requisite information and make informed decisions.

While LLMs promise to improve access to medical information and enhance patient education, their use is not without risks. The generative nature of these models may lead to the dissemination of inaccurate or misleading information, which could have serious implications for patient safety. Moreover, the absence of source transparency, the potential for biased outputs, and unresolved ethical and legal questions regarding accountability further complicate their integration into clinical settings.[7] These concerns underscore the importance of systematic evaluation of LLM-generated medical content in terms of accuracy and readability and in relation to safety, reliability, and ethical standards.

This study aimed to analyze LMMs' responses to FAQs concerning cataracts and cataract surgery, focusing on relevance and readability.

Materials and Methods

Since no patient data were used in our study, ethics committee approval was not required. In the study, we focused on questions frequently asked by patients about cataract and analyzed the performance of AI -based LMMs in answering these questions.

A search was conducted on Google, utilizing the terms "cataract" and "cataract surgery" as search queries. The FAQs from the "People Also Ask" (PAA) section for each query were meticulously documented. Initially, four questions were displayed, and the list was expanded with additional questions opened by clicking on any of these questions. This approach enabled the identification of the top 10 most relevant and FAQs, as provided by Google, for each search term.[5] Consequently, the subjects that users were most curious about – cataract and cataract surgery – were obtained systematically. A total of 10 questions were selected and categorized into the following categories: definition (2 questions), diagnosis (1 question), need for surgery (1 question), details of surgery (3 questions), and postoperative care (3 questions) (Table 1). It has been established that the outcomes of Google searches may be contingent upon a multitude of variables, including user-specific historical data, geographical location, and device characteristics. In light of the potential limitations imposed by these algorithmic personalization effects on the objectivity and reproducibility of study findings, a series of precautionary measures was implemented during the data collection process. Searches were conducted using incognito mode in an environment where cookies, browser history, and session data were thoroughly cleared. To ensure the absence of any influence from Google accounts or other location-determining software, a strict protocol was followed: No Google account was logged in, and no virtual private network, fixed IP address, or other location-determining software was used. All queries were executed on a single device and browser, thereby minimizing the impact of environmental variables. This methodological approach was undertaken to mitigate potential biases introduced by the algorithmic processes and to enhance the methodological reliability of the study's findings. All Google searches were conducted using English keywords, and the browser language was set to English. All searches were conducted on the same day and

Table 1. Questions about cataract and cataract surgery asked to chatbots

Cataract definition
What is cataract? Why does it happen?
What are the symptoms of cataract? How do I know if I have such a problem with my eyes?
Diagnosis of the disease
How to recognise a cataract? Which tests are performed during the examination?
Need for Surgery
Does every cataract require surgery? Are there other treatment methods?
Surgery
How is cataract surgery performed? What will I feel during the operation?
How are the lenses worn during the operation selected? Can a special selection be made for me?
Is cataract surgery risky? Will I have any problems?
Postoperative Period
When can I see clearly after the operation?
How can I protect my eyes after the operation? What should I pay attention to?
Will I get cataract again after the operation?

within a time frame of approximately 1 h, thereby limiting the effect of time-dependent variability in Google's "People Also Ask" (PAA) feature.

The language models used in the study were ChatGPT-3.5, ChatGPT-4o, Google Gemini, and Microsoft Copilot. The questions were posed to each model in the same format, and the answers given by the models were recorded as they were without modification. No intervention was made to the content of the responses, whereas the data were obtained.

All questions were posed to each chatbot 5 times on the same day, with the history cleared before each session. The responses were evaluated based on two main criteria: Appropriateness and readability. Two experienced cataract surgeons (A.B.O. and S.A.D.) conducted the relevance assessment, categorizing responses into three groups: "correct and sufficient," "correct but incomplete," and "incorrect".[8] A response was deemed "correct and sufficient" if it was similar to the recommendations that an assessor would provide to patients. An "incorrect" response was defined as a response that contained erroneous information or differed from the assessor's recommendation in the clinical setting. Responses that, while correct, lacked sufficient detail were categorized as "correct but incomplete." The accuracy assessment of responses was based on the current Preferred Practice Pattern® guidelines published by the American Academy of Ophthalmology.[9] After two expert evaluators independently classified the responses, the evaluation consistency was analyzed with Cohen's Kappa statistics.

For measuring readability, an online tool called Readable (https://app.readable.com/text/) was utilized, which assessed each chatbot's response using various parameters, including the Flesch ease of reading score, Flesch-Kincaid grade level (FKGL), Gunning Fog index, Coleman-Liau Index, and the simple measure of gobbledygook (SMOG) index. The Flesch reading ease score utilizes a mathematical formula that takes into account word length and sentence length, yielding a score that ranges from 1 to 100. A higher score indicates greater readability, for example, a score of 70–80 corresponds to a child's reading level, while a score of 30–50 signifies college-level readability. FKGL measures text comprehensibility according to the U.S. educational system. The resulting number indicates the minimum grade level required to understand the text, with higher scores implying greater difficulty and lower scores suggesting easier readability. The gunning fog ındex is calculated based on the average sentence length and the number of complex words in the text. The result, expressed as a number, indicates the minimum education level required to comprehend the text; a higher number correlates with a higher required education level.[8] The Coleman–Liau ındex utilizes word length and the number of letters in a sentence, differing from other readability tests, which use the number of letters instead of syllables. Finally, the SMOG index calculates the proportion of words containing three or more syllables. This index is commonly used in health and education, as it was developed to gauge the comprehensibility of academic and medical texts.[10-12]

In addition to these parameters, the total number of words and sentences in each chatbot's response was also recorded.

The data were analyzed using IBM Statistical Packages for the Social Sciences (SPSS) Statistics Standard Concurrent User, version 25 (IBM Corp., Armonk, New York, USA). The Shapiro–Wilk test evaluated the data's suitability for normal distribution. The Kruskal–Wallis test was applied to non-normally distributed data, and differences between groups were analyzed using the Bonferroni correction. A P < 0.05 was accepted as the significance level.

Results

According to the ophthalmologists' evaluation of the responses generated by the LLMs chatbots, ChatGPT-3.5 generated 81%, ChatGPT-4o 100%, Gemini 98%, and Copilot 54% "correct and complete" responses. No incorrect responses were observed in Gemini and ChatGPT-4o, with a score of 2% in ChatGPT-3.5 and 5% in Copilot (Fig. 1). The Cohen's Kappa value was found to be 0.65 between the two ophthalmologists who made the evaluation, indicating moderate agreement between the evaluators.

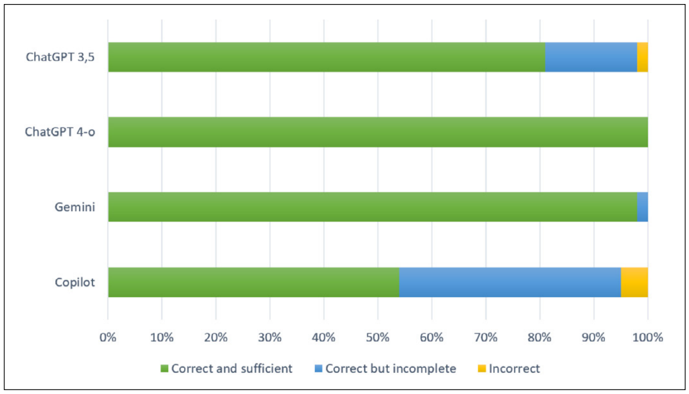

Fig. 1. Accuracy scoring of answers given in ChatGPT-3.5, ChatGPT-4o, Gemini, and Copilot.

The LLM chatbots showed a statistically significant difference (p<0.0001). Pairwise comparisons were conducted to evaluate statistical differences in accuracy scores among chatbots. After applying Bonferroni correction (adjusted significance threshold: p<0.0083), statistically significant differences were observed between ChatGPT-3.5 and ChatGPT-4o (p=0.0005) and ChatGPT-3.5 and Gemini (p=0.0079), indicating that both ChatGPT-4o and Gemini outperformed ChatGPT-3.5 in terms of accuracy. A significant difference was also found between ChatGPT-4o and Copilot (p=0.0065), supporting the higher performance of ChatGPT-4o. In contrast, comparisons between ChatGPT-3.5 and Copilot (p=0.3027) and ChatGPT-4o and Gemini (p=0.1594) did not yield statistically significant results. It shows that although ChatGPT-4o and Gemini provide superior accuracy compared to ChatGPT-3.5 and Copilot, their respective performance levels are not significantly different from each other.

The data for the readability ındex are presented in Table 2. Gemini yielded numerically higher Flesch Reading Ease scores and lower FKGL than the other models; however, these differences were not statistically significant (p=0.517 and p=0.354, respectively). Therefore, no definitive conclusions can be drawn regarding relative readability based on these measures. There was no significant difference between the groups in the Gunning Fog, Coleman-Liau, and SMOG index calculations (p=0.361, p=0.323, p=0.095, respectively). There was a statistically significant difference between the number of words and the number of sentences in the responses (p<0.0001 for both).

Discussion

The use of AI -based LMMs in health care is expanding rapidly. However, the accuracy of these technologies in medical information and their comprehensibility to patients remains a controversial issue.[13,14] In a comparative analysis of cataract and cataract surgery questions, Cohen et al.[5] observed analogous responses from both Google and ChatGPT. The analysis revealed that ChatGPT demonstrated superior accuracy in its responses. In light of the widespread use of AI tools in contemporary society, a comparative analysis was conducted involving four distinct AI tools. It was observed that ChatGPT-4o and Gemini yielded more accurate results. The literature has recently pointed out that the ChatGPT-4o version is generally good and performs better than the ChatGPT-3.5 version.[15,16] A recent study showed that ChatGPT responses were more accurate than Gemini in assessing vitreoretinal cases.[17] In our study, although ChatGPT-4o and Gemini produced responses with no incorrect content, there was no statistically significant difference between them, indicating comparable performance. In another study comparing the accuracy of ChatGPT, Gemini, and Copilot, chatbots were asked questions about keratoconus.[18] Similar to our research, Copilot received the lowest score.

As an element supporting the reliability of the evaluation process, a significant agreement was found at the Cohen's Kappa coefficient level between the classifications made by two independent ophthalmologists. This finding indicates that the consistency between the measurements is acceptable, thereby strengthening the internal validity of the applied evaluation from a methodological perspective.

Readability is a crucial factor influencing the ability to access and comprehend health information.[19] Various tests are used to assess the readability of health information materials. The flesh reading ease score was between 45 and 50 in all groups. The investigation revealed no statistically significant differences between the groups. This indicates that there are texts containing scientific and technical expressions, which are at a challenging level and require a university education. To improve readability, it is helpful

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

	ChatGPT-3.5	ChatGPT 4o	Gemini	Copilot	p*
Accuracy	1.21±0.45	1	1.02±0.14	1.51±0.59	<0.0001
Flesch reading ease score	45.90±7.95	45.78±7.87	47.90±7.49	45.64±10.89	0.517
Flesch-kincaid grade level	10.33±1.31	10.21±1.39	9.86±1.15	10.23±1.62	0.354
Gunning Fog	12.79±1.83	12.42±1.74	12.18±1.25	12.45±1.96	0.361
Coleman-Liau	11.89±1.28	11.93±1.2	11.8±1.29	12.27±1.62	0.323
SMOG index*	13.106±1.26	12.75±1.28	12.48±0.9	12.72±1.41	0.095
Word count	212.56±81.56	284.62±100.99	388.90±98.33	160.44±57.97	<0.0001
Sentence count	16.22±8	22.16±9.65	29.44±6.53	11.98±4.53	<0.0001

*SMOG index: Simple Measure of Gobbledygook.

to shorten sentences.[20] In this way, chatbots can also be educational for patients with a lower level of education. The FKGL scores suggested that Gemini produced slightly simpler responses than other models; however, the differences were not statistically significant, precluding any definitive interpretation regarding relative readability. The gunning fog index score gives information about the minimum education level. It was 12 for all chatbots. This means that the answers were at the level of a high school graduate. The calculation of the Coleman-Liau index was very similar to the Flesh Reading Ease score. When the SMOG index was evaluated, while the other chatbots gave results that were at a difficult level, the average score in ChatGPT-3.5 was 13.106 (very difficult; university and above).[21] In an evaluation where chatbots were asked questions about glaucoma, it was also noted that the readability of the answers was challenging for individuals with high educational levels.[22]

Another study comparing chatbots for retinopathy of prematurity highlighted that ChatGPT-4o was more accurate and Gemini was more readable.[23] While Gemini yielded numerically more favorable readability scores, the differences among chatbots were not statistically significant, and therefore, no clear superiority can be established. The practical implications of readability levels on patient education are a critical factor in determining the effectiveness of AI-powered chatbots. The findings of this study demonstrate that chatbot responses generally require a level of comprehensibility comparable to that of senior high school or university students, which may create accessibility issues for individuals with low health literacy. In particular, elderly patients or individuals with limited education may struggle to comprehend medical information that contains lengthy and complex sentences, which can lead to misunderstandings and potential health risks. The enhancement of chatbot efficacy in patient education can be achieved using simplified language, concise sentence structure, and employing patient-friendly expressions. The development of novel algorithms that facilitate the reduction of sentence lengths and enhance the comprehensibility of medical terminology holds promise in enhancing the readability levels of existing Chatbot models. In conclusion, optimizing readability is a pivotal aspect for AI-based health information systems to achieve widespread patient engagement.

If we examine the word count, we can see that Gemini generates answers with a very high word count. Similarly, in chatbot answers to questions about refractive surgery, it was observed that although Gemini had a higher word count, its readability was better.[10] Similarly, when examining the number of sentences, Gemini has the highest number. Although it uses more words to improve readability, it has compensated for this by increasing the number of sentences.

This study examines the relationship between the readability of AI-based chatbot responses and their effectiveness in patient education. The analysis, based on a comprehensive review of existing literature, finds that chatbot responses tend to exhibit an academic level of readability commensurate with the comprehension levels of university graduates. However, it is noteworthy that Flesch reading ease scores in the range of 45–50 suggest the presence of technical and scientific terminology, which may render the responses challenging for the average patient to comprehend. Nevertheless, to enhance the accessibility of chatbots in the context of patient education, algorithmic enhancements are necessary to reduce sentence lengths and employ more straightforward language. In this regard, optimizing chatbots to cater to individuals with lower educational levels to enhance patient-friendly communication may contribute to the augmentation of the efficacy of AI utilization in patient education processes.

With the increasing utilization of AI-powered chatbots in patient education, concerns about disseminating misinformation and potential legal liabilities have emerged as pivotal subjects in the discourse. Given that AI models derive responses through the process of learning from extensive datasets, there is a possibility that these responses may, in certain instances, be incomplete, misleading, or inaccurate, thereby jeopardizing patient safety. Although the incidence of misinformation was low in the present study, it is essential to recognize that AI chatbots should not replace medical counseling and advice. Patients must continue to rely on expert opinions to inform critical health decisions. In addition, the ambiguity surrounding the sources of chatbot responses can impede patients' access to reliable information, potentially leading to misunderstandings. Consequently, in developing AI-supported healthcare technologies, it is imperative that model outputs undergo regular review by healthcare professionals and that the responses explicitly state "for informational purposes and not as a substitute for medical advice." In the future, to minimize these risks, AI systems should be supported by mechanisms that enhance accuracy checks, and ethical and legal frameworks should be established to provide medical information.[7]

While earlier research has appraised the performance of AI chatbots in various ophthalmic diseases, these studies were confined mainly to generic accuracy analyses. They did not address the readability factor in depth. The present study makes a significant contribution to the existing literature by offering a thorough comparison that evaluates the performance of AI chatbots in terms of both accuracy and readability in the context of prevalent eye diseases, such as cataract and cataract surgery. The responses of ChatGPT-3.5, ChatGPT-4o, Gemini, and Microsoft Copilot were analyzed for accuracy by ophthalmologists and evaluated from a patient perspective using multiple readability metrics. The study's findings demonstrate that while AI-based chatbots are generally reliable regarding medical accuracy, the readability levels of the responses they produce require a high level of training and should be improved for patient education purposes. The analysis revealed that Gemini exhibited the highest readability score, indicating that AI models can employ diverse optimization strategies in language structures. All chatbots must be developed with innovative algorithms to generate concise and straightforward sentences, facilitating patient-friendly communication. In this context, our study contributes significantly to the existing literature by highlighting the need for innovative optimization methods to enhance the effectiveness of AI chatbots in ophthalmology and patient education processes.

A limitation of the study is that chatbots are constantly being updated. The same data may yield different results in subsequent periods due to the rapid advancement of technology. Our study aims to inspire future research to enhance the effectiveness of AI technologies in patient education processes by evaluating the informative potential of current LLMs in ophthalmology. Analyzing how chatbots improve accuracy and readability over time is a potential research topic. A notable constraint pertains to the limited number of expert evaluators involved in the assessment process. The chatbot responses were evaluated solely by two experienced cataract surgeons. While Cohen's Kappa coefficient demonstrated moderate inter-rater agreement, the limited panel size inherently constrains the external validity and generalizability of the findings. In studies that rely on expert opinion, the robustness of the conclusions is closely tied to the diversity and number of expert participants. A more extensive panel might have encompassed a broader array of clinical viewpoints, reducing the likelihood of evaluator bias. Future studies should incorporate multi-institutional and multidisciplinary expert panels to enhance methodological rigor and allow for more representative and reproducible assessments.

A further limitation that can be identified is that the mean values obtained from readability evaluations appear to be contingent upon a specific educational level. However, these values were not found to be statistically different. In the present study, no statistically significant difference was found. The responses were categorically at levels such as high school-university or 9th grade, 10th grade, and this was shared with the reader in detail. It is essential to emphasize that the current situation is characterized by categorical differences, given its potential to be perceived as a misleading comment, despite the absence of statistical significance.

In this study, the assessment of accuracy was based on the clinical practice-based judgments of two experienced cataract surgeons. However, to make such assessments more objective and reproducible, future studies are planned to use a systematic, scoring-based approach with predetermined information topics. This method will enable a more detailed measurement of both the comprehensiveness and accuracy of the responses, thereby increasing the comparability between different models.

A significant constraint of the search engine-based data collection method employed in the study pertains to the implementation of algorithmic personalization mechanisms, which render search results user specific. As stated in the referee's opinion, this situation is a potential source of bias that may affect the reproducibility of the study by different researchers. To mitigate the potential impact of personalization on the study's findings, comprehensive measures were implemented. These measures included conducting searches in incognito mode, ensuring users did not log in or utilize location information, and adhering to stringent research protocols. This methodological decision is predicated on the principle of algorithmic neutrality, which, in turn, enhances the internal and external validity of the findings obtained. If further studies of this nature are undertaken in this field, implementing a standardized definition of the digital environment is recommended to enhance the reliability of comparative analyses.

In the categorical distribution of the questions, a significant concentration was observed in topics related to the surgical process and postoperative care. This distribution was obtained naturally through the Google "People Also Ask" system, utilizing a patient-centered approach, and reflects that users' information needs are concentrated on these topics. However, the imbalance between categories was considered a potential limitation in the results, as the language models' performance was biased toward a certain area.

With all these implications in mind, AI developers can implement various strategies to improve readability without sacrificing medical accuracy. Algorithms that condense and simplify long, complex sentences can be developed. Medical terminologies can be supported with simple and patient-friendly explanations, and chatbots can adapt technical expressions to everyday language using a patient-centered language. In addition, they can incorporate adaptive mechanisms that dynamically adjust their responses according to the patient's educational level. Furthermore, integrating automated editing systems that analyze the readability of responses using metrics such as Flesch Reading Ease or FKGL can simplify complex phrases above a certain threshold. By making model updates based on user feedback, the patient-friendly language used by chatbots can be improved over time. These methods will help AI-based health chatbots become more accessible for patient education while maintaining medical accuracy and reliability.

Conclusion

The findings of this study indicated that ChatGPT-4o and Gemini exhibited a marked improvement in the precision of their responses to FAQs concerning cataracts and cataract surgery, surpassing the performance of ChatGPT-3.5 and Copilot. However, a lack of statistically significant differences was observed in readability scores across the models. Furthermore, the majority of responses were written at a level that may present a challenge for patients with limited health literacy. These findings underscore the necessity of optimizing the language outputs of LLMs to ensure both medical accuracy and comprehensibility. Subsequent endeavors should prioritize the implementation of algorithmic enhancements designed to optimize readability, the integration of patient-centered communication methodologies, and the establishment of ethical safeguards to mitigate the dissemination of misinformation through AI-driven health communication tools.

Ethics Committee Approval: Since this study involved no human participants, ethics committee approval wasn't necessary.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept: A.B.O., S.A.D.; Design: A.B.O., S.A.D.; Supervision: A.B.O., S.A.D.; Resource: A.B.O.; Materials: A.B.O.; Data Collection and/or Processing: A.B.O., S.A.D.; Analysis and/or Interpretation: A.B.O., S.A.D.; Literature Search: A.B.O.; Writing: A.B.O.; Critical Reviews: A.B.O., S.A.D.

Conflict of Interest: None declared.

Use of AI for Writing Assistance: Not declared.

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

References

1. Nagy ZZ. History of cataract surgery from ancient times to today: Honorary lecture at the 13th Conference of the Hungarian Medical Association of America–Hungary Chapter (HMAA-HC) at 30–31 August 2019, in Balatonfüred, Hungary. Dev Health Sci 2020;2:88–92. [CrossRef]
2. Lim EJ, Chowdhury M, Higham A, McKinnon R, Ventoura N, He YV, et al. Can large language models safely address patient questions following cataract surgery? Invest Ophthalmol Vis Sci 2023;64:1214. [CrossRef]
3. 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;36:384–91. [CrossRef]
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;388:1233–9. [CrossRef]
5. Cohen SA, Brant A, Fisher AC, Pershing S, Do D, Pan C. Dr. Google vs. Dr. ChatGPT: Exploring the use of artificial intelligence in ophthalmology by comparing the accuracy, safety, and readability of responses to frequently asked patient questions regarding cataracts and cataract surgery. Semin Ophthalmol 2024;39:472–9. [CrossRef]
6. Dihan Q, Chauhan MZ, Eleiwa TK, Brown AD, Hassan AK, Khodeiry MM, et al. Large language models: A new frontier in paediatric cataract patient education. Br J Ophthalmol 2024;108:1470–6. [CrossRef]
7. Monteith S, Glenn T, Geddes JR, Whybrow PC, Achtyes E, Bauer M. Artificial intelligence and increasing misinformation. Br J Psychiatry 2024;224:33–5. [CrossRef]
8. Aydın FO, Aksoy BK, Ceylan A, Akbaş YB, Ermiş S, Kepez Yıldız B, et al. 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;54:313–7. [CrossRef]
9. American Academy of Ophthalmology. Preferred Practice Pattern® Guidelines. https://www.aao.org/education/about-preferred-practice-patterns. Accessed June 26, 2025.
10. Huang G, Fang CH, Agarwal N, Bhagat N, Eloy JA, Langer PD. Assessment of online patient education materials from major ophthalmologic associations. JAMA Ophthalmol 2015;133:449–54. [CrossRef]
11. Hedman AS. Using the SMOG formula to revise a health-related document. Am J Health Educ 2008;39:61–4. [CrossRef]
12. Zhou S, Jeong H, Green PA. How consistent are the best-known readability equations in estimating the readability of design standards? IEEE Trans Prof Commun 2017;60:97–111. [CrossRef]
13. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open 2023;6:e2330320. [CrossRef]

14. Doğan L, Özçakmakcı GB, Yılmaz İE. The performance of chatbots and the AAPOS website as a tool for amblyopia education. J Pediatr Ophthalmol Strabismus 2024;61:325–31. [CrossRef]
15. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: A comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. Cureus 2023;15:e40822. [CrossRef]
16. Jiao C, Edupuganti NR, Patel PA, Bui T, Sheth V. Evaluating the artificial intelligence performance growth in ophthalmic knowledge. Cureus 2023;15:e45700. [CrossRef]
17. Carlà MM, Gambini G, Baldascino A, Giannuzzi F, Boselli F, Crincoli E, et al. Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. Br J Ophthalmol 2024;108:1457–69. [CrossRef]
18. Kayabaşı M, Köksaldı S, Durmaz Engin C. Evaluating the reliability of the responses of large language models to keratoconus-related questions. Clin Exp Optom 2024:1–8. [CrossRef]
19. McInnes N, Haglund BJ. Readability of online health information: Implications for health literacy. Inform Health Soc Care 2011;36:173–89. [CrossRef]
20. DuBay WH. The principles of readability. Costa Mesa: Impact Information; 2004.
21. McLaughlin GH. SMOG grading—A new readability formula. J Reading 1969;12:639–46.
22. Yalla GR, Hyman N, Hock LE, Zhang Q, Shukla AG, Kolomeyer NN. Performance of artificial intelligence chatbots on glaucoma questions adapted from patient brochures. Cureus 2024;16:e56766. [CrossRef]
23. Ermis S, Özal E, Karapapak M, Kumantaş E, Özal SA. Assessing the responses of large language models (ChatGPT-4, Claude 3, Gemini, and Microsoft Copilot) to frequently asked questions in retinopathy of prematurity: A study on readability and appropriateness. J Pediatr Ophthalmol Strabismus 2025;62:84–95. [CrossRef]