



DOI: 10.14744/eer.2025.44366
Eur Eye Res 2025;5(3):242–247

EUROPEAN
EYE
RESEARCH

ORIGINAL ARTICLE

Accuracy and legibility of the answers to the questions of the patients for cataract surgery: Comparison of 3 different large language models

 Ali Ceylan,  Yusuf Berk Akbas

Department of Ophthalmology, University of Health Sciences, Basaksehir Cam and Sakura City Hospital, Istanbul, Türkiye

Abstract

Purpose: The aim of the study is to evaluate the accuracy and legibility of the answers given by 3 different large language models (LLMs) to common patient questions about cataract surgery.

Methods: Three distinct LLMs (ChatGPT, Microsoft Copilot, and Google Gemini) were queried on 30 common inquiries about cataract surgery. The accuracy of the responses was evaluated using a Likert scale, based on the consensus opinion of two specialists. The readability of the responses was evaluated using three distinct readability indices: Flesch-Kincaid Grade Level, Coleman-Liau, and Flesch Reading Ease.

Results: None of the responses from LLMs received a score of 1 for any question. All responses generated by ChatGPT were rated four or higher. For comparison, 90% of Gemini's responses and 27% of Copilot's responses achieved scores of four or above. In consideration of legibility, it was observed that all three LLMs were challenging to read. However, Copilot exhibited slightly superior readability, followed by Gemini and ChatGPT, respectively.

Conclusion: While the responses provided by ChatGPT exhibited a slightly lower level of readability, they nonetheless proved to be the most proficient in answering cataract surgery-related questions. LLMs may support patient education, but their readability must be improved to ensure effective communication. Future work should focus on making AI-generated responses clearer and more accessible.

Keywords: Artificial intelligence; cataract surgery; large language models.

In recent years, there has been a notable increase in the number of individuals seeking information about their health concerns online.^[1-3] The pros and cons of utilizing online health data are still under discussion. Nevertheless, a greater proportion of patients now have internet access than ever before, and it is anticipated that the number

of patients utilizing the internet for health information will continue to grow over the next decade.^[4,5] The use of large language models (LLMs) generated by artificial intelligence (AI) is becoming an increasingly valuable source of information for patients seeking to enhance their understanding of their own health.^[6] LLMs receive training



Cite this article as: Ceylan A, Akbas YB. Accuracy and legibility of the answers to the questions of the patients for cataract surgery: Comparison of 3 different large language models. *Eur Eye Res* 2025;5(3):242–247.

Correspondence: Yusuf Berk Akbas, M.D. Department of Ophthalmology, University of Health Sciences, Basaksehir Cam and Sakura City Hospital, Istanbul, Türkiye

E-mail: yusufberkakbas@gmail.com

Submitted Date: 07.04.2025 **Revised Date:** 26.06.2025 **Accepted Date:** 08.07.2025 **Available Online Date:** 17.12.2025

OPEN ACCESS This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



based on existing digital knowledge and demonstrate proficiency in a variety of language processing tasks, including text summarization and user query response.^[7] The most frequently utilized LLMs in current use include ChatGPT-4 (OpenAI, San Francisco, California, USA), Gemini (Google, Mountain View, California, USA), and Copilot (Microsoft, Redmond, Washington, USA), among others. The objective of these models is to facilitate human-like responses to a multitude of text-based queries by capitalizing on an extensive training dataset and a sophisticated algorithmic approach.

Cataract surgery is one of the most frequently performed surgical procedures in the present era.^[8] Patients may have numerous inquiries pertaining to surgical procedures and frequently utilize the Internet as a source of information. The quality and readability of the information provided are of great consequence, as they directly impact patient comprehension and decision-making. Although LLMs have the potential to provide medical advice, it is imperative to conduct a comprehensive evaluation of their efficacy in offering accurate and intelligible information. The objective of this study is to evaluate the accuracy and readability of the responses provided by three different LLMs to patients' inquiries regarding cataract surgery.

Materials and Methods

This study was designed to examine the adequacy and readability of the information provided by three common LLMs. Two specialists (A.C. and Y.B.A.) compiled a collection of 30 real-life patient questions from previous patients who had undergone cataract surgery. The questions were divided into three main sections: general information, pre-surgery, and post-surgery, with each section consisting of 10 questions. On October 07, 2024, the three LLMs, ChatGPT, Copilot, and Gemini, were instructed with the following prompt: "Suppose you're an experienced ophthalmologist specializing in cataract surgery, and I'm a patient who has cataracts. Can you please answer my questions?" The questions were then presented to the LLMs one at a time after being carefully checked for grammatical precision. Given the LLMs' intrinsic capacity to generate disparate responses to a singular inquiry, only the initial reply to each query was documented for analytical purposes. To ensure that previous queries do not impact responses to subsequent ones, a separate conversation page was created for each new search request in LLMs. As the study did not involve any patients, approval from the ethics committee was not required. No personal or identifiable patient data were used in this study, and all procedures were conducted

in accordance with institutional ethical standards.

To assess the accuracy of the responses provided by LLMs, a Likert-type scale was constructed, ranging from 1 to 5: 1 signifying very poor or unacceptable inaccuracies, corresponding to "strongly disagreed" responses; 2 indicating poor accuracy with potentially harmful errors, corresponding to "disagreed" responses; 3 reflecting moderate inaccuracies that may lead to misinterpretation, corresponding to "neither agreed nor disagreed" responses; 4 denoting good quality with only minor, non-harmful inaccuracies, corresponding to "agreed" responses; and 5 representing very high accuracy with no inaccuracies, corresponding to "strongly agreed" responses. The evaluations were conducted independently by three specialists with over a decade of experience in cataract surgery. To ensure unbiased responses, the source of the answers was masked. Discrepancies in the Likert scale ratings were resolved through discussion, and the consensus score was adopted as the final evaluation.

To determine the readability of each response, the answers were submitted for analysis via the online readability application, Readable (<https://app.readable.com/text/>). The readability and comprehension criteria and standardization utilized in the study were based on those typical of the English language. All queries and responses were provided in English. To assess the readability of each response, three different indices were applied: The Coleman-Liau Index, the Flesch Reading Ease Score, and the Flesch-Kincaid Grade Level.^[9,10] The Flesch Reading Ease Score, which is used to assess the readability of texts at various educational levels – from fifth grade to postgraduate studies – is calculated based on two key parameters: the average sentence length, expressed in words, and the average word length, expressed in syllables.^[11] These variables are used to determine a text's score on a scale ranging from 0 to 100. It functions as an indicator of text comprehensibility, with higher scores indicative of greater readability.^[12] The Flesch-Kincaid Grade Level is a widely utilized readability score that was initially developed by the United States military and subsequently validated in prior studies.^[12,13] The Flesch-Kincaid Grade Level is significantly influenced by the number of words and syllables. The Coleman-Liau Index is a quantitative measure of text comprehension.^[14,15] The test postulates a relationship between word length and text readability, suggesting that the average number of letters per 100 words and average sentence length are better predictors of text readability than syllables.^[16,17]

The statistical analysis was conducted using the Statistical Package for the Social Sciences version 22.0 for Windows (IBM Corp., Armonk, NY, USA). Continuous variables were expressed as mean±standard deviation or median (minimum–maximum). The normality of the data was evaluated using the Shapiro-Wilk test. To compare LLMs, Friedman and Wilcoxon tests with Bonferroni correction were conducted. A power analysis was conducted using G*Power 3.1 for a Friedman test involving three related groups. Assuming a medium effect size ($W = 0.3$), an alpha level of 0.05, and a power of 0.80, the analysis indicated that a minimum of 28 observations was required. A p-value of less than 0.05 was considered statistically significant.

Results

The LLM responses to each question are presented in Table 1, organized according to the identified subgroups. Not a single response from LLMs was deemed worthy of a score of 1 for any given question. All responses generated by ChatGPT were assigned a score of four or above. In comparison, 90% of responses generated by Gemini and 27% of responses generated by Copilot were also awarded a score of four or above. Table 2 illustrates the median values of Likert scoring for all questions and for each subgroup of questions. A statistically significant difference was identified ($p < 0.001$ for all) in each group. The median scores of responses to all questions from ChatGPT were found to

Table 1. Evaluation of the accuracy of the responses provided by the LLMs according to the Likert scale

Questions	ChatGPT	Gemini	Copilot
General Information			
1- What is cataract surgery?	5	5	3
2- How do I know if I need cataract surgery?	5	5	2
3- What are cataracts, and why do they occur?	5	4	2
4- Can cataracts be treated without surgery?	5	4	3
5- How long does cataract surgery take?	5	4	4
6- What happens during cataract surgery?	5	4	3
7- Will I feel pain during cataract surgery?	5	4	3
8- How safe is cataract surgery?	5	4	3
9- What is the recovery time after cataract surgery?	5	4	3
10- Is cataract surgery performed on both eyes at the same time?	5	3	3
Pre-surgery			
11- How do I prepare for cataract surgery?	5	4	3
12- Do I need to stop taking medications before surgery?	5	4	2
13- What type of anesthesia is used during cataract surgery?	4	5	3
14- What risks are involved in cataract surgery?	5	4	4
15- What tests are done before cataract surgery?	5	3	2
16- Are there any special considerations for diabetics before surgery?	5	4	3
17- What type of lens will be implanted during cataract surgery?	5	5	3
18- What is the difference between monofocal and multifocal lenses?	5	4	3
19- How is the right lens for me selected?	5	4	2
20- Can cataract surgery correct other vision problems like astigmatism or presbyopia?	5	5	3
Post-surgery			
21- How long will I need someone to assist me after surgery?	5	4	2
22- Will I still need glasses after cataract surgery?	5	4	3
23- Can I drive myself home after cataract surgery?	5	5	4
24- Can cataracts come back after surgery?	5	5	4
25- What activities should I avoid after cataract surgery?	5	4	3
26- How should I care for my eye after surgery?	5	5	4
27- What are the signs of complications after cataract surgery?	5	4	4
28- Is it normal to feel discomfort or see halos after surgery?	5	4	3
29- How long will the artificial lens last?	5	5	4
30- Are there any long term side effects of cataract surgery?	5	3	4

Table 2. The median values of the Likert scale evaluation of question subgroups

	ChatGPT Median (Min-Max)	Gemini Median (Min-Max)	Copilot Median (Min-Max)	p*
General Information	5.0 (5-5)	4.0 (3-5)	3.0 (2-4)	<0.001
Pre-surgery	5.0 (4-5)	4.0 (3-5)	3.0 (2-4)	<0.001
Post-surgery	5.0 (5-5)	4.0 (3-5)	4.0 (2-4)	<0.001
All Questions	5.0 (4-5)	4.0 (3-5)	3.0 (2-4)	<0.001

*Friedman test.

be significantly higher than those from Copilot and Gemini ($p<0.001$ for both). Furthermore, the median score of the Gemini was found to be significantly higher than that of Copilot ($p<0.001$). A comparison of the question subgroups revealed that ChatGPT achieved a statistically significantly higher score than Copilot in each subgroup ($p=0.012$ for all). A comparison of the performance of ChatGPT and Gemini on general information questions revealed that ChatGPT scored significantly higher than Gemini ($p=0.021$) on these questions. However, when the same comparisons were made on pre-surgery and post-surgery questions, the differences in performance were not statistically significant ($p=0.105$ and $p=0.060$, respectively). In addition, Gemini exhibited significantly higher scores than Copilot in general information ($p=0.030$) and pre-surgery ($p=0.018$) questions, whereas no significant difference was observed in the post-surgery ($p=0.063$) questions.

A summary of the readability indices is provided in Table 3. A comparison of the Coleman-Liau Index between the LLMs reveals that Copilot exhibits a significantly lower index than both ChatGPT ($p<0.001$) and Gemini ($p=0.006$). In a similar analysis, a comparison of the Flesch Reading Ease Score between the LLMs revealed that Copilot exhibited

a significantly higher score than both ChatGPT ($p<0.001$) and Gemini ($p=0.006$). However, when comparing the Flesch–Kincaid Grade Level of LLMs, it becomes evident that ChatGPT displays a significantly higher level than both Copilot ($p<0.001$) and Gemini ($p=0.012$).

Discussion

The implementation of AI is becoming increasingly prevalent on a global scale. Furthermore, a multitude of novel AI models are currently being developed. These include language models that have been trained to utilize pre-acquired data to navigate the internet and generate immediate responses in chatbot conversations.^[18] This paper offers an examination of the influence of this variability on LLM performance and output quality. It highlights that the discrepancies between the responses of different LLMs are predominantly attributable to variations in the algorithms employed. Nevertheless, in the present era, an increasing number of patients are seeking medical information from LLMs. Accordingly, the responses to these inquiries must be precise and readily comprehensible. In the present study, the responses of various LLMs to queries pertaining to cataract surgery were subjected to comparison.

Although the responses exhibited slightly lower readability, ChatGPT-4 demonstrated significantly higher accuracy, achieving a 100% agreement rate on all questions, with responses of either “strongly agree” or “agree.” This rate was 90% and 27% for Gemini and Copilot, respectively. A comparative analysis of the results reveals that Copilot exhibits a notable deficiency in accuracy when compared to the other two LLMs. Tepe et al.^[19] posed questions regarding breast imaging to several chatbots and, as in our own investigation, discovered that ChatGPT demonstrated greater accuracy and lower readability compared to Gemini

Table 3. Evaluation of the readability of the responses provided by the LLMs.

	ChatGPT Mean±SD	Gemini Mean±SD	Copilot Mean±SD	p*	p**	
Coleman-Liau Index	12.94±1.51	12.46±1.62	11.14±2.52	<0.001	ChatGPT versus Gemini	0.792
					ChatGPT versus Copilot	<0.001
					Gemini versus Copilot	0.006
Flesch Reading Ease Score	42.30±8.92	45.53±9.22	52.48±13.96	<0.001	ChatGPT versus Gemini	0.354
					ChatGPT versus Copilot	<0.001
					Gemini versus Copilot	0.006
Flesch-Kincaid Grade Level	11.65±1.59	10.36±1.58	9.54±2.02	0.001	ChatGPT versus Gemini	0.012
					ChatGPT versus Copilot	<0.001
					Gemini versus Copilot	0.096

*Friedman test; **Wilcoxon signed-rank test with Bonferroni adjustment; LLMs: Large language models.

and Copilot. In another study that compares ChatGPT and Gemini on hypertension education, the results were comparable.^[20]

A review of the Flesch-Reading Ease scores reveals that Copilot exhibits a notably higher level of readability than the other two LLMs. While ChatGPT and Gemini necessitated college-level reading proficiency within the US educational framework, Copilot required a reading skill commensurate with that of a 10th–12th grade student. An evaluation of the Flesch-Kincaid Grade Level results reveals that ChatGPT exhibits a considerably higher score than Gemini and Copilot, indicating a lower level of readability. Upon evaluation of the results, it becomes evident that ChatGPT, Gemini, and Copilot are at a level of complexity that can be understood by students in the 11th, 10th, and 9th US grades, respectively. An analysis of the Coleman-Liau Index revealed that ChatGPT and Gemini were deemed comprehensible by American 12th graders, while Copilot was considered so by 11th graders. It is notable that despite the considerable discrepancy in their respective scores, all three LLMs are classified as “quite hard to read.” These results demonstrated that ChatGPT’s output necessitated the most advanced US educational qualifications for comprehension, which might restrict its accessibility to a broader demographic of patients. Conversely, Copilot exhibited the lowest required reading level, indicating its potential to become a significantly more accessible resource for patient education.

In a study conducted by Hillmann et al.^[21] the relationship between atrial fibrillation and cardiac implantable electronic devices was analyzed with the support of LLM. Their findings aligned with those of our study, revealing that ChatGPT exhibited lower readability scores. Xie et al.^[22] presented a series of progressively intricate clinical scenarios to ChatGPT-4, Gemini, and Copilot. The results demonstrated that ChatGPT-4 exhibited superior readability and reliability compared to the other two LLM models. Another study utilizing ChatGPT 4.0 reported readability results comparable to our findings, indicating a reading level suited for undergraduate and graduate audiences and a text that is relatively difficult to comprehend.^[23]

The present study is subject to certain limitations. Initially, the research was restricted to 30 questions. The manner in which queries are phrased when interacting with LLMs can have a significant impact on the quality and nature of the responses produced. Furthermore, there is a lack of consensus regarding the consistency of LLMs’ responses to identical or similar questions posed at different times. In

this study, each question was presented to the LLMs on a single occasion. Another limitation of the study is the lack of a patient cohort for real-world evaluation.

While LLMs offer valuable support in delivering accessible health information, their unregulated use may lead to misinformation and potentially weaken the doctor–patient relationship. Therefore, these tools must be used with appropriate clinical oversight, serving as adjuncts rather than substitutes for professional medical advice. Looking ahead, the role of LLMs in medicine is expected to expand significantly. Future models may demonstrate improved accuracy, personalization, and real-time adaptability, making them useful tools in clinical decision-making and patient education. However, ongoing research, ethical oversight, and regulatory frameworks will be essential to ensure their safe and effective integration into healthcare systems.

Conclusion

ChatGPT demonstrated the highest accuracy in answering cataract surgery questions, followed by Gemini. The accuracy of Copilot responses was found to be inferior to that of the other two LLM models. In terms of readability, all LLM outputs were challenging to process, although Copilot exhibited a comparatively greater ease of comprehension. Despite the complexity of these models, they are notable for their readability, accuracy, and reliability. However, future research should include a wider variety of patient queries covering all aspects of cataract surgery and involve blinded participants to confirm these findings, ensuring the effectiveness and reliability of LLMs in clinical applications.

Ethics Committee Approval: As the study did not involve any patients, approval from the ethics committee was not required.

Peer-review: Externally peer-reviewed.

Authorship Contributions: Concept: A.C.; Design: A.C.; Supervision: Y.B.A.; Data Collection and/or Processing: A.C., Y.B.A.; Analysis and/or Interpretation: A.C., Y.B.A.; Literature Search: Y.B.A.; Writing: A.C., Y.B.A.; Critical Reviews: A.C., Y.B.A.

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. Bujnowska-Fedak MM, Węgierek P. The impact of online health information on patient health behaviours and making decisions concerning health. *Int J Environ Res Public Health* 2020;17:880. [\[CrossRef\]](#)
2. Van Riel N, Auwerx K, Debbaut P, Van Hees S, Schoenmakers

- B. The effect of Dr Google on doctor–patient encounters in primary care: a quantitative, observational, cross-sectional study. *BJGP Open* 2017;1. [\[CrossRef\]](#)
3. Burzyńska J, Bartosiewicz A, Januszewicz P. Dr. Google: Physicians—the web—patients triangle: Digital skills and attitudes towards e-Health Solutions among Physicians in South Eastern Poland—A cross-sectional study in a Pre-COVID-19 era. *Int J Environ Res Public Health* 2023;20:978. [\[CrossRef\]](#)
 4. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online health information seeking among US adults: measuring progress toward a healthy people 2020 objective. *Public Health Rep* 2019;134:617–25. [\[CrossRef\]](#)
 5. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *J Med Internet Res* 2013;15:e1933. [\[CrossRef\]](#)
 6. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell* 2023;6:1169595. [\[CrossRef\]](#)
 7. Devlin J. Bert. Pre-training of deep bidirectional transformers for language understanding. *arXiv Prepr arXiv:181004805*. 2018.
 8. Alió JL. *Cataract surgery: from today's standards to future progress*. Lippincott Williams & Wilkins; 2017. p. 309.
 9. Michel C, Dijanic C, Abdelmalek G, Sudah S, Kerrigan D, Gorgy G, et al. Readability assessment of patient educational materials for pediatric spinal conditions from top academic orthopedic institutions. *J Child Orthop* 2023;17:284–90. [\[CrossRef\]](#)
 10. Patel AJ, Kloosterboer A, Yannuzzi NA, Venkateswaran N, Sridhar J, editors. Evaluation of the content, quality, and readability of patient accessible online resources regarding cataracts. *Semin Ophthalmol* 2021;1–7.
 11. Doğan L, Özçakmakçı GB, Yılmaz İE. The Performance of Chatbots and the AAPOS Website as a Tool for Amblyopia Education. *J Pediatr Ophthalmol Strabismus* 2024;1–7. [\[CrossRef\]](#)
 12. Friedman DB, Hoffman-Goetz L. A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Educ Behav* 2006;33:352–73. [\[CrossRef\]](#)
 13. Ley P, Florio T. The use of readability formulas in health care. *Psychol Health Med* 1996;1:7–28. [\[CrossRef\]](#)
 14. 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:e27850. [\[CrossRef\]](#)
 15. 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;14:635–7. [\[CrossRef\]](#)
 16. Valentine MJ, Cottone G, Kramer HD, Kayastha A, Kim J, Pettinelli NJ, et al. Lower back pain imaging: a readability analysis. *Cureus* 2023;15. [\[CrossRef\]](#)
 17. Coleman M, Liau TL. A computer readability formula designed for machine scoring. *J Appl Psychol* 1975;60:283. [\[CrossRef\]](#)
 18. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health* 2023;2:e0000198. [\[CrossRef\]](#)
 19. 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. [\[CrossRef\]](#)
 20. Lee TJ, Campbell DJ, Patel S, Hossain A, Radfar N, Siddiqui E, et al. Unlocking health literacy: the ultimate guide to hypertension education from ChatGPT versus Google Gemini. *Cureus* 2024;16. [\[CrossRef\]](#)
 21. Hillmann HA, Angelini E, Karfoul N, Feickert S, Mueller-Leisse J, Duncker D. Accuracy and comprehensibility of chat-based artificial intelligence for patient information on atrial fibrillation and cardiac implantable electronic devices. *Europace* 2024;26:euad369. [\[CrossRef\]](#)
 22. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Seifman MA. Investigating the impact of innovative AI chatbot on post-pandemic medical education and clinical assistance: a comprehensive analysis. *ANZ J Surg* 2024;94:68–77. [\[CrossRef\]](#)
 23. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. *Ophthalmol Retina* 2023;7:862–8. [\[CrossRef\]](#)