OriginalArticle

Comparative Analysis of AI Chatbots Chat GPT, Gemini, and Copilot's Answers to Common Cataract Questions

PJO – Official Journal of Ophthalmological Society of Pakistan

Busra Guner Sonmezoglu[1], Halil Ibrahim Sonmezoglu[2]
[1]Serdivan State Hospital Sakarya, Turkey, [2] Hendek StateHospital, Sakarya, Turkey

ABSTRACT

Purpose: To compare the readability and quality of the answers given by artificial intelligence (AI) chatbots to the twenty-five most frequent searches about cataracts on Google.

Study Design: Cross sectional comparative study.

Place and Duration of study: Serdivan State Hospital's Ophthalmology Department from March 2024 to April 2024.

Methods: The word 'Cataract' was entered into Google Trends, and 25 trending searches made worldwide since 2004 were identified. These 25 trending searches were entered separately into AI programs. The answers were examined for quality using the Ensuring Quality Information for Patients (EQIP) test, while the readability was evaluated using the Flesch-Kincaid Reading Ease (FKRE) and Flesch-Kincaid Grade Level (FKGL).

Results: The keywords searched most frequently were 'Cataract surgery,''Eye cataract and 'After Cataract'. The EQIP category of all the three AI chatbots was found to be with 'serious quality issues.' When EQIP scores were compared, Chat GPT had a lower median value than Gemini and Copilot (p:0.001, p:0.007, respectively), while there was no significant difference between Gemini and Copilot (p:0.098). When FKRE values were compared, Chat GPT had a lower median value than Gemini and Copilot (p:0.001, p:0.001, respectively), with no significant difference between Gemini and Copilot (p:0.557). When FKGL values were compared, Chat GPT had a higher median value than Gemini and Copilot (p:0.003, p:0.001, respectively), with no significant difference between Gemini and Copilot (p:0.245).

Conclusion: All three AI chatbots had an EQIP category of 'serious issues with quality. 'The readability of all three was not at the recommended level.

Key Words: Artificial intelligence, Cataract, Chatbot, Chat GPT, Gemini, Copilot.

How to Cite this Article: Sonmezoglu BG, Sonmezoglu HI. Comparatıve Analysıs of AI Chatbots Chat Gpt, Gemini, and Copilot's Answers to Common Cataract Questıons. 2024;40(4):370-375.
Doi: 10.36351/pjo.v40i4.1887

Correspondence: Busra Guner Sonmezoglu
Serdivan State Hospital Sakarya, Turkey
Email: busra-gnr1@hotmail.com

Received: July 07, 2024
Accepted: September 01, 2024

INTRODUCTION

Cataract is a leading cause of vision impairment and blindness worldwide, particularly among the elderly. The epidemiology of cataracts is critical to understanding the burden of this condition on a global scale. Cataracts are responsible for approximately 51% of global blindness, as stated by the World Health Organization. This highlights the substantial influence of this eye condition.[1]

AI chatbots are application programs that function as virtual assistants, offering services to users through conversations in natural language on social media networks or websites.[2] AI chatbots can be employed in various fields, including customer assistance,

healthcare interactions, and symptom recognition, to aid individuals in determining if they should consult a healthcare expert.[3,4] AI chatbots can be an essential tool for cataract patients. They can increase patient compliance in diagnosis, treatment, and post-surgical care, reduce patient concerns, and access the correct information quickly.[5] It is important to acknowledge that AI chatbot replies have limitations and hazards, such as the possibility of generating inaccurate outputs and concerns around data security. Some chatbots may provide accurate and detailed information, while others may have limitations regarding accuracy, clarity, and appropriateness of language.[6]

This study compares the readability and quality of Gemini, Copilot, and Chat GPT responses to the 25 most common search words about Cataract.

METHODS

This cross sectional study was conducted between March 1st 2024, and April 1st 2024, at Ophthalmology Department of Serdivan State Hospital. Since there was no process involving data from individuals, obtaining ethical committee authorization for this inquiry was unnecessary.[7] As a preventative move to mitigate any bias, all personal browser data was wiped before doing the searches.

The 25 most frequently searched words related to cataract were obtained from Google Trends (https://trends.google.com/) by entering the word cataract in a worldwide search from January 2004 to March 1st 2024. Three conditions, including "Eye," "Eye surgery," and "Glaucoma," were excluded from the analysis due to their lack of connection to the subject.

Following the original search order, The specified keywords were entered in sequence into Chat GPT 3.5 (24th version, https://chat.openai.com/), Gemini February Version (https://gemini.google.com/), and Copilot (https://copilot.microsoft.com/). Prior to initiating search queries, all browser records were thoroughly erased, and a separate account was created to engage with each AI chatbot, guaranteeing clear segregation. In order to maintain segregation and improve analytical procedures, each inquiry was treated on an individual chat panel. The obtained responses were stored to conduct assessments on legibility and excellence.

The obtained resources were evaluated for quality using the EQIP tool. This device evaluates various facets of the material, including its coherence and the caliber of its composition. The survey consisted of a total of 20 questions, where participants were given response options such as "yes," "partly," "no," or "does not apply." The rating system entailed multiplying the count of affirmative responses by 1, the count of partially affirmative responses by 0.5, and the count of negative responses by 0. The acquired value was divided by 20, which represented the total number of elements. If a response was deemed 'not legitimate', the number of responses were deducted from 20 and then split. Ultimately, the percentage was derived by multiplying the value by 100. Texts that received ratings between 76% and 100% were categorized as 'well written', suggesting exceptional quality. ratings between 51% and 75% indicated good quality with minor issues', while scores between 26% and 50% indicated 'serious quality issues'. Texts that scored between 0% and 25% were considered to have 'severe quality issues.[8] Two Ophthalmologists scored the EQIP.

The AI Chatbots' information readability was evaluated using the FKGL and FKRE metrics. To find the FKGL, one must follow a series of procedures, such as dividing the word count by the sentence count, multiplying the result by 0.39, dividing the word count by the syllable count, and finally multiplying the result by 11.8. Considering variables like sentence length and syllable count, the estimated comprehension was obtained by adding together the results and subtracting 15.59 from the final figure. A higher number showed that the language was difficult to understand, whereas a lower value showed that you can understand it. To find the FKRE, on the other hand, the text's readability was multiplied by 1.015 for the average sentence length (the average number of words per sentence) and 84.6 for the average syllable count per word. Subtracting the resultant difference from 206.835 yielded the final figure. Better readability was indicated by a lower Ease of Reading score, whereas greater complexity was indicated by a higher score.[9]

Statistical analysis was performed using SPSS version 27 (IBM, New York, USA). The normality of data was evaluated using Shapiro-Wilk test. Mean ± standard deviation was used to describe continuous data, while frequency was used to represent categorical data. The Kruskal-Wallis test was used to calculate the differences among groups. The significance level was set at 0.05.

RESULTS

The keywords most frequently searched were 'Cataract surgery,''Eye cataract,' and 'After Cataract' (Table 1). According to the EQIP Tool, 45.45% were conditions or illnesses, 36.36% were tests, operations, investigations, or procedures, 13.63% were discharged, or aftercare and 4.54% were medications or products.

Table 1: *Terms related to cataract from 2004–2024 are included in the Google Trends (Irrelevant Terms Are crossed out).*

Rank	Keyword
1	Cataract surgery
2	Eye cataract
3	Eye
4	After cataract
5	After cataract surgery
6	Eye surgery
7	Eye surgery cataract
8	Cataract lens
9	What is cataract
10	Cataract surgery cost
11	Cataract vision
12	Cataracts
13	Cataract eyes
14	Cataract operation
15	Cataract meaning
16	Cataract symptoms
17	Glaucoma
18	What is cataract surgery
19	Cataract laser surgery
20	Cataract treatment
21	Cataracts surgery
22	Cataract eye drops
23	ICD-10 cataract
24	Cataract surgery recovery
25	What is a cataract

Trinidad and Tobago, Singapore and Australia ranked highest in terms of search interest scores (Figure 1).

Figure 2 displays the graph illustrating the temporal evolution of cataract popularity as observed using Google Trend Analysis.

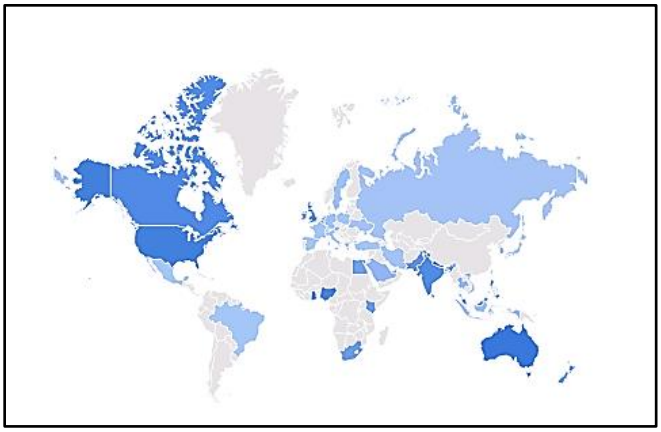

Figure 1: An analysis of the global demand for searches related to cataract in different regions, based on data from Google Trends (excluding regions with relatively little searches).

A statistically significant disparity in FKRE scores was seen among the chatbots (p=0.001). Upon applying the Bonferroni correction, a pairwise analysis was conducted, which indicated a statistically significant disparity in FKRE scores between Chat GPT, Gemini, and Copilot (p:0.001, p:0.001, respectively). Gemini achieved the highest score, while Chat GPT obtained the lowest score. There was no significant difference between Gemini and Copılot (p:0,557).

The FKGL scores between the chatbots also demonstrated significant differences (p<0.001). Upon using the Bonferroni adjustment for pairwise comparisons of FKGL scores, it was observed that Gemini and Copilot exhibited significantly lower FKGL scores compared to Chat GPT (p:0,003, p:0,001 respectively). There was no difference between Gemini and Copılot (p:0,245) (Table 2).

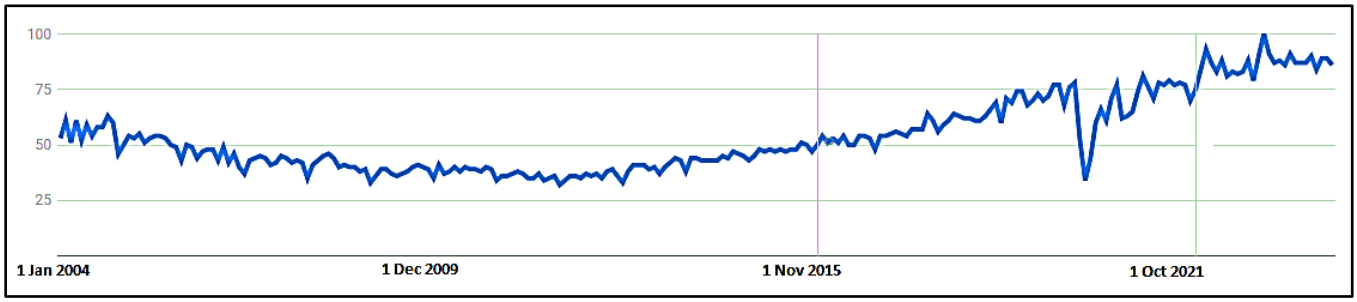

Figure 2: An analysis was conducted on the worldwide search interest between 2004 and 2024 using data obtainedfrom Google Trends.

Table2: When *analyzing chatbot-generated content, statistical measurements such as the median, minimum, maximum, mean, and standard deviation are employed.*

Chatbot		Median	Minimum	Maximum	Mean	Std. Deviation
Gemini	EQIP Score	44.31	25	53.33	43.98	7.91
	FKRE	54.95	36.70	64.20	51.92	7.85
	FKGL	9.65	7.50	13.40	10.06	1.69
Copilot	EQIP Score	39.64	20.58	53.84	40.39	7.49
	FKRE	51.65	35.10	67.50	50.97	7.90
	FKGL	9.40	6	11.70	9.30	1.55
Chat GPT	EQIP Score	34.16	23.07	50	34.80	6.53
	FKRE	40.20	18.70	70.50	39.90	10.58
	FKGL	11.70	8	15.80	12	2.21

The median values of EQIP scores were 44.31 in Gemini, 39.64 in Copilot, and 34.16 in Chat GPT. EQIP scores between chatbots differed significantly ($p<0.001$). Pairwise comparisons of EQIP scores using Bonferroni correction revealed that Gemini and Copilot were statistically significantly higher than Chat GPT (p:0.001, p:0.007, respectively). There was no significant difference between Gemini and Copilot (p:0,098).

DISCUSSION

Our results indicated that the AI chatbots' responses to inquiries connected to cataracts did not adhere to the prescribed standards of readability. It was determined that all chatbots exhibited significant quality deficiencies. Furthermore, the comparisons revealed that Gemini and Copilot generated content that was more legible and of superior quality in contrast to Chat GPT. While there was no statistically significant distinction observed between Gemini and Copilot, Gemini exhibited a superior level of readability and content quality.

According to the data, the three keywords that were most commonly searched were "Cataract surgery,""Eye cataract," and "After cataract." The frequency of these search terms suggests a significant inclination towards acquiring knowledge regarding the symptoms, indications, and therapeutic approaches for cataracts. This highlights the imperative for easily accessible, expeditious, and accurate information about Cataract. In order to meet this need, it is crucial to evaluate AI chatbot material, deliver training in AI chatbots using reliable sources, and re-educate AI chatbots under the supervision of a panel of healthcare experts to guarantee high-quality and comprehensible information.

The quality and readability of AI chatbot responses about cataracts can vary. Some AI chatbots may utilize sophisticated algorithms and natural language processing to generate detailed and accurate responses about cataracts. Others may have limitations in their knowledge base or need help to effectively communicate complex medical information in a user-friendly manner.[6]

Yilmaz et al, compared chatbots for cataract patient education and found that chat GPT provided the most accurate and comprehensive answers to cataract-related questions. In contrast, Bard (artificial intelligence chatbot developed by Google) provided the most understandable answers.[10]

In our study, the quality and readability levels of the texts generated by Gemini and Copilot were higher than those produced by Chat GPT, which may be attributed to their more recent development compared to Chat GPT. Although there was no statistical difference between Gemini and Copilot, Gemini had a higher EQIP score, likely due to the inclusion of visual content in its answers.

According to Ittarat et al's, findings, chatbots can guide patients through pre-cataract surgery, explaining the procedure, recovery time, and post-operative care. They can also address common concerns, such as potential risks, reduce anxiety and uncertainty about the surgery, and ensure that patients have realistic expectations.[6]

According to the National Institute of Health, it is recommended that health-related content be written at a reading level that is equivalent to or below that of an eighth-grade student.[11] The results of our investigation demonstrated that the text readability produced by chatbots regarding cataract was consistent with the expected comprehension levels of those who have completed around 16 to 17 years of formal education. In order to regulate the degrees of legibility, AI might

be educated using revised guidelines. At the same time, content generated under a panel of experts' supervision could be beneficial in adhering to the requisite readability criteria.

According to Pan et al's, findings, AI chatbots provide accurate information in response to significant cancer-related search queries. However, the responses were often lacking in practical application and were written at a comprehension level more appropriate for higher education learners.[12]

While AI chatbots offer instant responses about cataract that are convenient and accessible for users' inquiries, research suggests potential limitations regarding the quality of the content.[13] For example, Coşkun et al, reported that Chat GPT has difficulties in delivering precise and high-quality. patient information on prostate cancer.[14] Cocci et al, discovered that Chat GPT generated subpar information pertaining to patients in the field of urology.[15] Şahin et al, compared 5 different AI chatbots about erectile dysfunction and found that none of the chatbots had the required level of readability and quality.[16]

Although chatbot responses are generally understandable, the suitability of the content might differ, with specific responses omitting crucial elements. This highlights the need for caution when utilizing chatbots for medical information.[17] According to Temel et al, employing different approaches, such as streamlining sentence structures, utilizing straightforward language, offering explicit explanations, arranging the content efficiently, and incorporating visual features, might enhance the readability of the content generated by Chat GPT.[18]

Inconsistencies in chatbot responses may arise due to factors such as training data and conversation history, despite the chatbot's capacity to generate high-quality phrases and effectively handle diverse conversation subjects. Consequently, it becomes imperative to employ approaches to identify and rectify these inconsistencies.[19]

There were some limitations of our study. Initially, the search was limited to the initial 25 terms, which may have negatively affected the accuracy of the results. The inclusion of additional keywords in a comprehensive technique has the potential to yield more accurate and refined conclusions. In addition, expanding the utilization of non-English keywords could enhance the scope of the evaluation, leading to more generally applicable findings. Second, the chatbots used in our study are the publicly available Chat GPT 3.5 and Gemini. There are also upgraded, paid versions of these chatbots. This is one of the limitations of our study, given that the paid versions may provide better responses.[20] The study could be redesigned using the latest version of these chatbots.

CONCLUSION

While AI chatbots can offer exciting potential in providing information about cataract, users should exercise caution and not become overly reliant on them. Users must evaluate the sources critically and not rely solely on AI-generated data. AI chatbots should continue to be developed. As AI chatbots continue to evolve and improve, the quality and readability of their responses about cataract can be expected to improve.

Funding: This study was not funded by any organization.

Conflict of Interest: Authors declared no conflict of interest.

Ethical Statement: No ethical approval was needed because this is not a human study, but only online information was used.

REFERENCES

1. Organization WH. Cataract 2024. Available from: https://www.emro.who.int/health-topics/cataract/. Accessed August 30 2024
2. **Pérez-Soler S, Juarez-Puerta S, Guerra E, de Lara J.** IEEE Software. Choosing a chatbot development tool. 2021;**38(4):**94-103. Doi:10.1109/MS.2020.3030198
3. **Athota L, Shukla VK, Pandey N, Rana A.** Chatbot for healthcare system using artificial intelligence. In: 2020 8th International conference on reliability, infocom technologies and optimization (trends and future directions)(ICRITO) 2020 Jun 4 (pp. 619-622). IEEE.
4. **Sallam M.** Chat GPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023;**11(6):**887. Doi: 10.3390/healthcare11060887.
5. **Ghorashi N, Ismail A, Ghosh P, Sidawy A, Javan R.** AI-powered chatbots in medical education: potential applications and implications. Cureus. 2023;**15(8)**. Doi:10.7759/cureus.43271

6. **Ittarat M, Cheungpasitporn W, Chansangpetch S.** Personalized Care in Eye Health: Exploring Opportunities, Challenges, and the Road Ahead for Chatbots. J Pers Med. 2023;**13(12):**1679. Doi: 10.3390/jpm13121679.
7. **Bagcier F, Yurdakul OV, Temel MH.** Quality and readability of online information on myofascial pain syndrome. J Bodyw Mov Ther. 2021;**25:**61-66. Doi: 10.1016/j.jbmt.2020.11.001.
8. **Moult B, Franck LS, Brady H.** Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information. Health Expect. 2004;**7(2):**165-175. Doi:10.1111/j.1369-7625.2004.00273.x.
9. **Boles CD, Liu Y, November-Rider D.** Readability Levels of Dental Patient Education Brochures. J Dent Hyg. 2016;**90(1):**28-34.
10. **Yılmaz IBE, Doğan L.** Talking technology: exploring chatbots as a tool for cataract patient education. Clin Exp Optom. 2024:1-9. Doi: 10.1080/08164622.2023.2298812.
11. **Oliffe M, Thompson E, Johnston J, Freeman D, Bagga H, Wong PKK.** Assessing the readability and patient comprehension of rheumatology medicine information sheets: a cross-sectional Health Literacy Study. BMJ Open. 2019;**9(2):**e024582. Doi: 10.1136/bmjopen-2018-024582.
12. **Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti A.** Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol. 2023;**9(10):**1437-1440. Doi:10.1001/jamaoncol.2023.2947
13. **Ting DSJ, Tan TF, Ting DSW.** Chat GPT in ophthalmology: the dawn of a new era? Eye (Lond). 2023:1-4. Doi: 10.1038/s41433-023-02619-4
14. **Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O.** Can Chat GPT, an artificial intelligence language model, provide accurate and high-quality patient information on prostate cancer? Urology. 2023;**180:**35-58. Doi:10.1016/j.urology.2023.05.040
15. **Cocci A, Pezzoli M, Lo Re M, Russo GI, Asmundo MG, Fode M, et al.** Quality of information and appropriateness of Chat GPT outputs for urology patients. Prostate Cancer Prostatic Dis. 2024;**27(1):**103-108. Doi: 10.1038/s41391-023-00705-y.
16. **Şahin MF, Ateş H, Keleş A, Özcan R, Doğan Ç, Akgül M, et al.** Responses of Five Different Artificial Intelligence Chatbots to the Top Searched Queries About Erectile Dysfunction: A Comparative Analysis. J Med Syst. 2024;**48(1):**38. Doi: 10.1007/s10916-024-02056-0.
17. **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(1):**euad369.
18. **Temel MH, Erden Y, Bağcıer F.** Information Quality and Readability: Chat GPT's Responses to the Most Common Questions About Spinal Cord Injury. World Neurosurg. 2024;**181:**e1138-e44. Doi:10.1016/j.wneu.2023.11.062
19. **Prats JM, Estecha-Garitagoitia M, Rodrıguez-Cantelar M, Fernando L.** Automatic Detection of Inconsistencies in Open-Domain Chatbots. Proceedings of the Proceeding IberSPEECH. 2022:116-20.
20. **Massey PA, Montgomery C, Zhang AS.** Comparison of Chat GPT-3.5, Chat GPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations. J Am Acad Orthop Surg. 2023;**31(23):**1173-1179. Doi: 10.5435/JAAOS-D-23-00396.

Authors Designation and Contribution

Busra Guner Sonmezoglu; Medical Doctor: *Manuscript review.*

Halil Ibrahim Sonmezoglu; Medical Doctor: *Concepts, Design, Manuscript editing.*