Title: Performance evaluation of ChatGPT and Google Bard for frequently asked glaucoma-related questions

Sahil Thakur*, MS[1] ,Aditi Rathore, MD[2], Devanshi Rathore,DNB[3], Sidharth Puri, MD[4]

[1]Singapore Eye Research Institute, Singapore; [2]Rohilkhand Medical College and Hospital, Rohilkhand, India; [3]Government Rukmani Devi Beni Parsad Jaipuria Hospital, Jaipur, India; [4]Paras Hospital, Panchkula, India

Corresponding Author:

Dr Sahil Thakur, MBBS, MS

Department of Ocular Epidemiology

Singapore Eye Research Institute

Singapore

drsahilthakur@gmail.com

+917888667891, +6583070039

Word Count :

Keywords :

LLM, ChatGPT, Google Bard, glaucoma, patient information, QUEST

Disclosure:

No financial or commercial disclosure by any of the authors.

Precis

Chatbots can be used to answer frequently asked questions about glaucoma, but generated responses lack proper attribution and have low readability scores. These findings are important in light of increasing use of artificial technology for healthcare purposes.

Abstract:

Purpose: Chatbots like OpenAI's ChatGPT and Google's Bard are computer programs that simulate human conversation and can provide patients with a personalized and interactive way of learning about glaucoma, allowing them to ask questions, receive feedback, and access resources in real time. However, the effectiveness of these chatbots in providing accurate and reliable information has not been evaluated. We designed this study to evaluate the side-by-side responses of ChatGPT and Bard to a curated list of 15 frequently asked questions (FAQ) related to glaucoma.

Methods: We used ChatGPT 3.5 (March 23 Version) and Bard (Update: 2023.04.21). The 15 questions were curated using the FAQ sections from the BrightFocus Foundation and Glaucoma Research Foundation websites. The responses were evaluated using the QUEST tool by 2 independent reviewers on the basis of 5C's of credibility, currency, content, construction, and clarity. Readability of the responses was also analysed using standard scores like the Flesch Kincaid Reading Ease and Grade Level.

Results: On the QUEST tool, ChatGPT scored 10.4 while Bard scored 10.2 but the difference was not statistically significant (P= 0.55). The chatbots had a kappa coefficient of 0.6296 (95% CI: -0.4505 to 1.7098) and Lin's coefficient of concordance of 0.6352 (95% CI: 0.2531 to 0.8458) indicating moderate agreement between the chatbots. They scored poorly on authorship, attribution, and currency of information. However, the responses generated were unbiased and supported the initiation or building up of the patient-physician relationship. Bard performed better on the readability scores indicating better reader comprehension.

Conclusions: This study demonstrates the limitations of chatbots as the information provided cannot be referenced or attributed to any source and the responses are difficult for general public to understand. Talking to an eye doctor was frequently highlighted in the chatbot responses indicating complementarity between chatbots and healthcare professionals. Though chatbots offer an exciting medium to improve patient knowledge about glaucoma, our results highlight the need for developing specialised chatbots for use in healthcare scenarios.

Introduction

Large Language Models (LLMs) represent the current frontier of research in artificial intelligence. When deployed as chatbots, such as OpenAI's ChatGPT and Google's Bard, LLMs have the ability to simulate human conversation, facilitating user interaction through messaging platforms, websites, and mobile applications. Although these models have showcased robust performance while testing medical knowledge as in structured medical examinations, there remain significant concerns regarding their impact on patients and the broader healthcare sector in the foreseeable future.[1-6]

Evidence suggests that such chatbots hold promise in deploying personalized management plans for chronic conditions like diabetes and obesity, thereby presenting innovative alternatives to traditional professional support in these domains.[7, 8] Improved knowledge, awareness and self-care related to chronic diseases have been demonstrated to improve disease outcomes.[9-11] This is particularly crucial for conditions like glaucoma, where public knowledge and awareness are notably deficient.[12, 13] In the realm of glaucoma management, these chatbots may offer patients an individualized and interactive learning platform about the disease, enabling them to ask questions, receive feedback, and access resources in real-time.

However, the assurance of these chatbots in delivering accurate and reliable information remains unevaluated and unverified.[14] Preliminary research has indicated that these chatbots are capable of disseminating misleading, biased, and outdated information.[15-17] Therefore, it is of paramount importance to rigorously assess the output generated by these chatbots, especially given their free accessibility and integration into browsers and search engines.

This study was conceived with the objective of evaluating the performance of these chatbots utilizing the Quality Evaluation Scoring Tool (QUEST).[18] QUEST is a comprehensive tool designed to assess online health-related articles, focusing on various dimensions such as credibility, currency, content, construction, and clarity of information. Additionally, the readability of the responses was analysed using standard scores like the Flesch Kincaid Reading Ease and Grade levels.[19] The findings of our study aim to assist readers in making informed decisions regarding the efficacy of chatbots in providing information about frequently asked questions pertaining to glaucoma.

Methodology

Selection of Chatbots

Two widely-used and accessible LLMs, ChatGPT 3.5 (March 23 Version) and Bard (Update: 2023.04.21), were selected for this study due to their prevalence and their application in simulating human conversation across various digital platforms. (**Figure 1**)

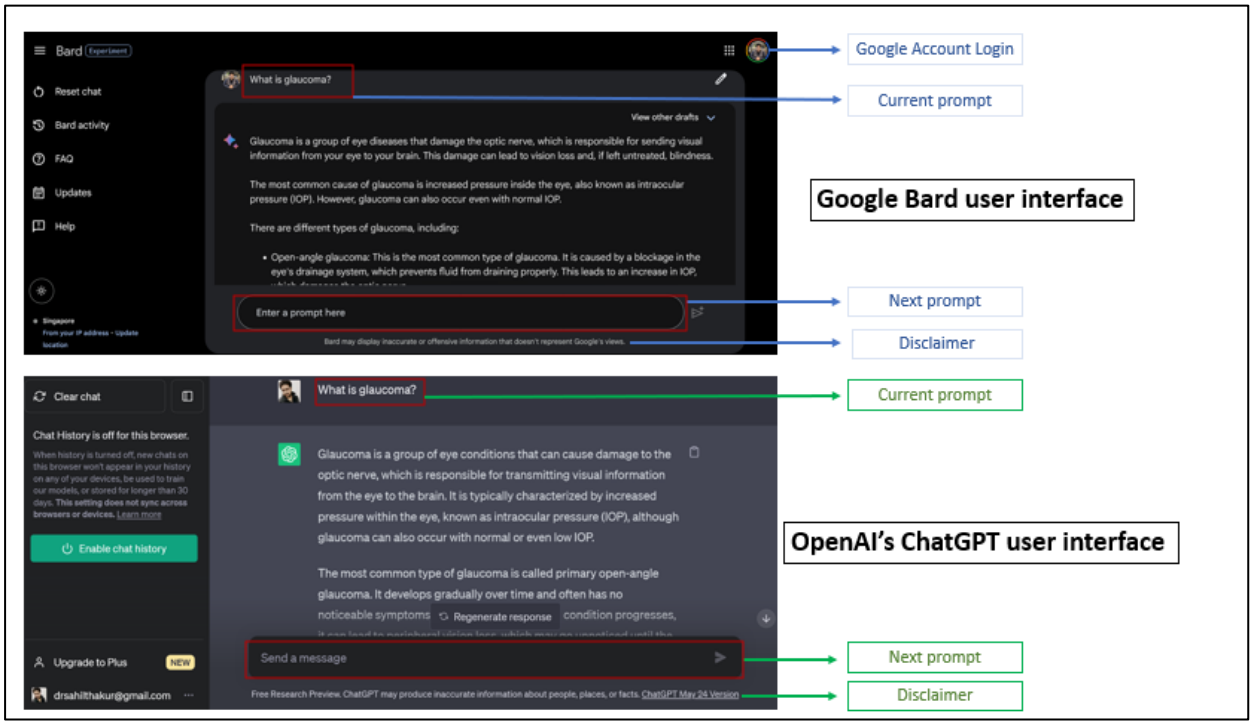

Figure 1: Screenshots of the chatbots with their user interface. Notice the need to log in with an account for using both chatbots.

Development of Query Set

A comprehensive set of queries was developed, encompassing frequently asked questions and scenarios related to glaucoma using the BrightFocus Foundation and Glaucoma Research Foundation websites and are shown in **Table 1.** These queries were designed to evaluate the chatbots' ability to provide accurate, relevant, and personalized information and advice.

Table 1: Fifteen frequently asked question (FAQ) curated from BrightFocus Foundation and Glaucoma Research Foundation websites

S.No	Frequently Asked Question
1	What is glaucoma?
2	What causes glaucoma?
3	Who is at risk of getting glaucoma?
4	How is glaucoma harmful to vision?
5	Will I go blind from glaucoma?
6	How can I tell if I have glaucoma?
7	How is glaucoma detected or diagnosed?
8	How is glaucoma treated?

9	Will my vision be restored after treatment?
10	Can glaucoma be prevented?
11	What are different types of glaucoma?
12	What is normal eye pressure?
13	How does pressure inside the eye and other factors affect vision and glaucoma?
14	What resources are available to help glaucoma patients and caregivers?
15	Where can I find more information?

Interaction with Chatbots

The interactions with each chatbot were conducted utilizing the query set. These interactions took place within a controlled environment to guarantee both consistency and reliability across sessions. To initiate these sessions, the chatbots were accessed through a browser in incognito mode, subsequent to clearing the cache, ensuring a clean slate for each interaction. While both chatbots necessitated login via account credentials, the search history feature was deactivated during the interactions to prevent any influence on the responses. All replies from the chatbots were recorded in an Excel spreadsheet for detailed subsequent analysis. To better simulate the variability and randomness of real-world interactions, each chatbot was tested twice using the same set of queries. This dual-phase testing approach aimed to identify any inconsistencies or variations in the responses generated by the chatbots across different sessions.

Evaluation Using QUEST

The Quality Evaluation Scoring Tool (QUEST) was employed to assess the responses from the chatbots. (**Table 2**) The assessment focused on five key dimensions: credibility, currency, content, construction, and clarity of the information provided. Each response was scored according to predefined criteria in each dimension, and an overall QUEST score was calculated for each chatbot. A qualitative analysis was also conducted on the chatbots' responses to identify any patterns of misinformation, bias, or outdated information. The analysis involved a detailed examination of the content and the context in which the information was provided.

Table 2: QUEST tool adapted from Robillard et al. Maximum score can be summed up to 28.

Authorship	0: No indication of authorship or username	**(Score x 1)**
	1: All other indications of authorship	
	2: Author's name and qualification clearly stated	
Attribution	0: No sources	**(Score x 3)**
	1: Mention of expert source, research findings (though with insufficient information to identify the specific studies), links to various sites, advocacy body, or other	
	2: Reference to at least one identifiable scientific study, regardless of format (e.g., information in text, reference list)	

	3: Reference to mainly identifiable scientific studies, regardless of format (in >50% of claims)	
	For all articles scoring 2 or 3 on attribution: *Type of study*	**(Score x 1)**
	0: In vitro, animal models, or editorials	
	1: All observational work	
	2: Meta-analysis, randomized controlled trials, clinical studies	
Conflict of interest	0: Endorsement or promotion of intervention designed to prevent or treat condition (e.g.: supplements, brain training games, foods) within the article	**(Score x 3)**
	1: Endorsement or promotion of educational products & services (books, care home services)	
	2: Unbiased information	
Currency	0: No date present	**(Score x 1)**
	1: Article is dated but 5 years or older	
	2: Article is dated within the last 5 years	
Complementarity	0: No support of the patient-physician relationship	**(Score x 1)**
	1: Support of the patient-physician relationship	
Tone	0: Fully supported (authors fully and unequivocally support the claims, strong vocabulary such as “cure”, “guaranteed” and “easy”, mostly use of non-conditional verb tenses (“can”, “will”), no discussion of limitations)	**(Score x 3)**
	1: Mainly supported (authors mainly support their claims but with more cautious vocabulary such as “can reduce your risk” or “may help prevent”, no discussion of limitations)	
	2: Balanced/cautious support (author’s claim are balanced by caution, includes statements of limitations and/or contrasting findings)	

Readability Assessment

The readability of the chatbots’ responses was assessed employing a variety of standard readability scores, utilizing the readability checker available at https://originality.ai/readability-checker.[20] The comprehensive suite of scores used for this analysis comprised the Flesch Kincaid Reading Ease, Flesch Kincaid Grade Level, Gunning Fog Index, SMOG Index, Powers Sumner-Kearl, FORCAST Grade Level, Coleman Liau Index, Automated Readability Index, Dale-Chall Readability Grade, Spache Readability Grade, and the Linsear Write Grade. (**Table 3**) We used different scales and grades, as, it allowed for a more holistic and nuanced evaluation of the text, as each readability formula has its own unique focus and calculation method, emphasizing different aspects such as sentence length, word complexity, and syllable count. These scores collectively offered valuable insights into the complexity and understandability of the information provided, thereby aiding in determining whether the responses from the

chatbots are accessible and comprehensible to a diverse audience, including individuals with varying levels of health literacy and education.

Table 3: Readability Indices used in the study and how they are computed.

Flesch Kincaid Reading Ease	The Flesch-Kincaid Reading Ease formula is designed to assess the readability of a text by examining the average sentence length and syllables per word. Higher scores indicate easier to read text. **FKRE = 206.835 – (1.015 x ASL) – (84.6 x ASW)** Where FKRE is Flesch Kincaid Reading Ease, ASL is the average sentence length, and ASW is the average number of syllables per word.
Flesch Kincaid Grade Level	The grade level translates the reading ease score to the equivalent U.S. school grade level. **FKGL = (0.39 x ASL) + (11.8 x ASW) – 15.59** Where FKGL is Flesch Kincaid Grade Level, ASL is the average sentence length, and ASW is the average number of syllables per word.
Gunning Fog Index	It accounts for sentence length and the number of complex words, which are defined as words with three or more syllables. **GFI = 0.4 x (ASL + PHW)** Where GFI is Gunning Fog Index, ASL is the average sentence length, and PHW is the percentage of hard words (words with three or more syllables).
SMOG Index	It estimates the years of education needed to comprehend a text by analysing the number of polysyllabic words in a sample. **SMOG = 1.043 x √(30 x PDW) + 3.1291** Where SMOG is the Simplified Measure of Gobbledygook (SMOG) Index and PDW is the number of polysyllabic words per 30 sentences.
Powers Sumner-Kearl Grade Level	It assesses text readability by analysing syllable patterns and word frequency data. **PSK-GL = 0.0778(ASL) + 0.0455(NS) – 2.2029** Where PSK-GL is Powers Sumner-Kearl grade level, ASL is average sentence length and NS is number of syllables
FORCAST Grade Level	It measures text readability based on the frequency of single syllable words. **FORCAST = 20 – (NOSW / NTW)** Where NOSW is the number of one-syllable words, and NTW is the total number of words in the sample.
Coleman Liau Index	It considers the number of characters per word and sentences per 100 words to estimate the U.S. grade level required to understand a text. It is unique in that it doesn't rely on syllable counts making it an efficient formula. **CLI = (0.0588 x L) – (0.296 x S) – 15.8** Where CLI is the Coleman-Liau Index, L is the average number of characters per 100 words, and S is the average number of sentences per 100 words.
Automated Readability Index	It uses characters per word and words per sentence to determine the readability of a text. The results are converted to the U.S. Grade level.

	ARI = (4.71 x CHW) + (0.5 x WPS) – 21.43 Where ARI is the Automated Readability Index, CHW is the average number of characters per word, and WPS is the average number of words per sentence.
Dale-Chall Readability Grade	It incorporates sentence length and percentage of difficult words, which are those not found in a pre-defined list of 3,000 familiar words. **DC = (0.1579 x PDW) + (0.0496 x ASL)** Where DC is the Dale-Chall Readability Grade, PDW is the percentage of difficult words, and ASL is the average sentence length.
Spache Readability Grade	It is specifically designed to analyse texts aimed at young readers. By looking at sentence length and unfamiliar words it can determine the rough age at which a reader would need to be for the text. **SRG = (ASL + PDW) / 2** Where SRG is the Spache Readability Grade, ASL is the average sentence length, and PDW is the percentage of difficult words.
Linsear Write Grade	It evaluates text readability by focusing on the number of simple and complex words in a sample of 100 words. **LWG = (SIMW + (COMW x 3))** Where LWG is the Linsear Write Grade, SIMW is the number of simple words, COMW is the number of complex words, and NTW is the total number of words in the sample.
Adopted from originality.ai, Available at: https://originality.ai/readability-checker	

Statistical Analysis

In the statistical analysis section of the manuscript, we applied a range of analytical techniques to gain insights from the data. First, descriptive statistics were employed to provide a concise summary of both the QUEST and readability scores, allowing for a better understanding of the dataset's central tendencies and variability. To assess the performance of the two chatbots, a comparative analysis was conducted using appropriate statistical tests after assessing the data distribution, specifically the Mann Whitney U test. This analysis aimed to detect any significant differences in how the two chatbots performed across the measured parameters. Additionally, we evaluated the agreement and correlation between the two chatbots using a comprehensive set of statistical metrics. These metrics included kappa coefficients, Lin's concordance coefficient, intraclass-correlation coefficients (ICC), and the Spearman correlation coefficients. These analyses helped us assess the consistency and relationship between the chatbots' outputs. In our analysis, a significance threshold of $P < 0.05$ was applied to identify statistically meaningful results. To conduct this data analysis, we utilized STATA software, Version 16, developed by Stata Corp LP, based in College Station, TX, USA.

Results

There were 15 questions in the query set that were assessed in the study. Two independent graders used QUEST to grade the responses generated by the chatbots. The grades were pooled together and subsequently analysed. On the QUEST, ChatGPT scored 10.4±1.05 while Bard scored 10.2±0.77, but the difference was not statistically significant (*P*= 0.55). The chatbots had a kappa coefficient of 0.629 (95% CI: -0.451 to 1.709) and Lin's coefficient of concordance of 0.635 (95% CI: 0.253 to 0.846) indicating moderate agreement between the chatbots. They scored poorly on authorship, attribution, and currency of information. However, the responses generated were unbiased and supported the initiation or building up of the patient-physician relationship.(**Supplementary Table 1**)

The readability assessment was done by using the Flesch Kincaid Reading Ease, Flesch Kincaid Grade Level, Gunning Fog Index, SMOG Index, Powers Sumner-Kearl, FORCAST Grade Level, Coleman Liau Index, Automated Readability Index, Dale-Chall Readability Grade, Spache Readability Grade, and the Linsear Write Grade. (**Table 4**) Bard (57.47±11.92) scored significantly higher than ChatGPT (39.87±13.46) on the Flesch Kincaid Reading Ease. This was consistent with other readability criteria as Bard consistently scored lower than ChatGPT on the other metrics. (all *P*<0.05) These results indicate that Bard's responses are generally more readable than ChatGPT's, as indicated by statistically significantly higher scores on the Flesch Kincaid Reading Ease and lower scores across other metrics.

Table 4: Readability assessment indicators for Open AI ChatGPT and Google Bard

Metric	Chat GPT Mean ±SD	Bard Mean ±SD	*P value
Flesch Kincaid Reading Ease	39.87±13.46	57.47±11.92	P<0.001
Flesch Kincaid Grade Level	11.20±1.27	8.53±1.43	P<0.001
Gunning Fog Index	15.66±2.03	11.47±1.52	P<0.001
SMOG Index	11.99±0.03	11.19±0.76	P<0.001
Powers Sumner-Kearl Grade	7.25±1.04	6.11±0.93	P=0.001
FORCAST Grade Level	12.13±1.27	10.96±1.09	P=0.009
Coleman Liau Index	11.99±0.03	11.35±0.81	P=0.010
Automated Readability Index	10.89±1.84	7.92±1.82	P<0.001
Dale-Chall Readability Grade	7.11±1.04	5.79±0.81	P<0.001
Spache Readability Grade	5.00±0.00	4.93±0.17	P=0.079
Linsear Write Grade	11.67±3.68	7.28±2.15	P=0.003

*P-value from Mann Whitney U test

The agreement and correlation between the readability scores for the two chatbots were also evaluated using ICC and Spearman correlation coefficients. (**Table 5**) Several metrics showed moderate to strong positive correlation (e.g., Gunning Fog Index and Flesch Kincaid Grade Level), indicating a consistent relationship between the scores assigned to the two chatbots. However, Coleman Liau Index showed a slight negative correlation, suggesting a discrepancy in this metric while assessing the two chatbots. This may also be due to its unique nature as it does not rely on syllable counts. The ICC between two chatbots had wide confidence intervals and indicated low agreement between the readability scores of the two chatbots.

Table 5: The intraclass correlation coefficients (ICC) and Spearman correlation coefficient (ρ) indicators for the readability scores for Open AI ChatGPT and Google Bard

Metric	Intraclass correlation coefficient (95% CI)	Spearman correlation coefficient (*P*-value)
Flesch Kincaid Reading Ease	0.381 (-0.089 to 0.804)	0.364 (0.182)
Flesch Kincaid Grade Level	0.140 (-0.088 to 0.486)	0.571 (0.026)
Gunning Fog Index	0.166 (-0.057 to 0.539)	0.656 (0.008)
SMOG Index	0.008 (-0.184 to 0.331)	0.284 (0.305)
Powers Sumner-Kearl Grade	0.421 (-0.113 to 0.784)	0.344 (0.209)
FORCAST Grade Level	0.262 (-0.127 to 0.637)	0.391 (0.149)
Coleman Liau Index	-0.009 (-0.278 to 0.379)	-0.235 (0.399)
Automated Readability Index	0.150 (-0.109 to 0.505)	0.239 (0.392)
Dale-Chall Readability Grade	0.269 (-0.114 to 0.657)	0.447 (0.095)
Spache Readability Grade	0 (-0.423 to 0.469)	NA
Linsear Write Grade	0.215 (-0.115 to 0.593)	0.453 (0.089)

Discussion

In this study, we sought to evaluate and compare the readability and quality of responses provided by two chatbots, ChatGPT and Bard, to frequently asked glaucoma related questions. The results present an intricate picture of the performance of these chatbots in delivering comprehensible and high-quality information.

The detailed analysis of readability metrics revealed varying degrees of readability between ChatGPT and Bard. While it was observed that Bard achieved a higher score on the Flesch Kincaid Reading Ease (FKR) metric, indicating more comprehensible text. The Mann-Whitney

U test indicated significant differences (all $P<0.05$) in several readability metrics, including FKR, Flesch Kincaid Grade Level , Gunning Fog Index, SMOG Index, Powers Sumner-Kearl, FORCAST Grade Level, Coleman Liau Index, Automated Readability Index, Dale-Chall Readability Grade, Spache Readability Grade, and Linsear Write Grade. However, the Spearman correlation coefficients varied, with some metrics showing moderate positive correlation between the chatbots, suggesting a level of consistency in relative readability across different questions.

Readability of healthcare material has also been studied previously. Symons and Davis showed that for a sample of 248 patient information sheets, the mean Flesch Reading Ease score was 49.3±5.7 and for the Flesch-Kincaid Grade Level 11.4±1.1. The mean SMOG score was 13.2±0.9. They also reported that commercial information sheets were more than twice as long as non-commercial, but statistically more readable ($P = 0.03$) when analysed using the SMOG formula.[19] Williamson and Martin also reported an average Flesch readability of all hospital patient information sheets as 60 (Range: 43.8-76.9), with a Flesch Kincaid Grade Level of 7.8 (Range: 5.4-10.2).[21] They highlighted a critical aspect that though the patient information sheets were well laid out and easy to read, majority would have exceeded patient comprehension. This potentially means that a large number of people who do not have the requisite linguistic ability or comprehension skills are being excluded from the benefit of a patient information sheet.[22] A systematic review has also highlighted that ophthalmic patient education materials are also consistently written at a level that is too high for many patients to understand.[23] The LLM's thus suffer from the same flaws (high quality input, high quality output or GIGO fallacy i.e. garbage in, garbage out) as other AI models, as the training data that they have been developed on is not suitable for all readers and thus similar issues persist in their output as well.[24-26]

The quality of the responses was evaluated using the QUEST tool.[18] It yielded insights into the reliability and robustness of the chatbots in providing accurate and consistent information. A Cohen's Kappa statistic of 0.634 indicates substantial agreement between ChatGPT and Bard in terms of the quality of responses, highlighting their potential as reliable sources of information. The chatbots' emphasis on early diagnosis, regular eye check-ups, and consulting an eye doctor underscores their role in complementing healthcare professionals and enhancing patient knowledge about glaucoma. These can be further seen in the **Supplementary Table 1** where we have included sample responses from the two chatbots.

Despite the promising results, the study also sheds light on the several limitations of chatbots. The inability to reference or attribute information can pose challenges for users seeking verified and credible information. This study demonstrates the limitations of chatbots as the information provided cannot be referenced or attributed to any source. This can be addressed by adding, 'with reference' or 'give reference' to the prompt in some cases, however several times the models addressed queries like this with a response like, 'I'm a text-based AI and can't assist with that.' or simply crash. (**Figure 2**) Moreover, general public would just type in their query without specifically asking for a reference. Thus, while strategies such as modifying prompts to request references can be explored, but the effectiveness of such approaches remains to be seen, given the varied responses from the chatbots. It has been demonstrated that these chatbots are likely to produce fake or incorrect references, thereby affecting the validity and accuracy of responses.[15, 17] Future studies can explore this aspect of LLMs while designing prompt query sets.

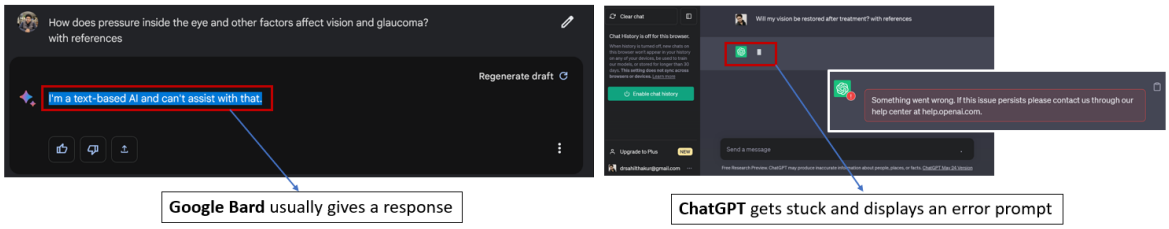

Figure 2: Examples of prompt failure when attribution/ references are requested from the chatbots

Our results also highlight the crucial consideration for healthcare professionals to assess the comprehension levels of their patients, before suggesting use of chatbots for general health-related queries. This additional layer of discernment ensures that the recommended digital platforms align with the patients' literacy levels, thereby facilitating effective communication and comprehension of health information. Consequently, there is also a clear opportunity for developers to improve readability of LLM models by implementing strategies such as simplifying language, breaking text into shorter sentences and paragraphs, using headings and bullet points, providing contextual clarity, and actively seeking and incorporating user feedback. These enhancements can offer responses that not only maintain accuracy and informativeness but also enhance overall user satisfaction and accessibility through improved readability. Some other alternative approaches also include fine tuning foundation models like Meta-AI (LLaMA) for medical chats.[27]

This study thus offers a multifaceted perspective on the complexity and understandability of information delivered by chatbots in the context of glaucoma. The application of a diverse set of readability measures and the standardized QUEST tool have contributed to a robust and reliable assessment, mitigating the limitations inherent in any single measure. The insights derived from this study pave the way for future research focused on improving the readability and reliability of chatbot responses, exploring user interaction dynamics, and developing advanced evaluation tools and techniques to assess the performance of chatbots in diverse healthcare scenarios.

Conclusion:

Chatbots have emerged as a promising and innovative medium for augmenting patient knowledge about glaucoma. For individuals in the general public exploring information on glaucoma, our findings underscore a distinct variation in readability between ChatGPT and Bard. This variation, along with a lack of consistent concordance between the chatbots coupled by issues like lack of appropriate referencing, signals the necessity for users to contemplate their individual preferences and comprehension requisites when utilizing these digital aids for health-related inquiries. Ultimately, the nuanced differences in readability and the personalized nature of chatbots underscore the collective responsibility of all stakeholders like users, developers, and healthcare providers to be thoughtful and considerate of individual informational needs and literacy levels. This collaborative approach can maximize the benefits of using chatbots as supplementary tools for health education and awareness, fostering an environment of informed, aware, and empowered healthcare consumers.

References

1. Guerra GA, Hofmann H, Sobhani S, et al. GPT-4 Artificial Intelligence Model Outperforms ChatGPT, Medical Students, and Neurosurgery Residents on Neurosurgery Written Board-Like Questions. World Neurosurg 2023.
2. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations. J Am Acad Orthop Surg 2023.
3. Gilson A, Safranek CW, Huang T, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ 2023;9:e45312.
4. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. JMIR Med Educ 2023;9:e48002.
5. Moshirfar M, Altaf AW, Stoakes IM, et al. Artificial Intelligence in Ophthalmology: A Comparative Analysis of GPT-3.5, GPT-4, and Human Expertise in Answering StatPearls Questions. Cureus 2023;15(6):e40822.

6. Wang H, Wu W, Dou Z, et al. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: Pave the way for medical AI. Int J Med Inform 2023;177:105173.
7. Arslan S. Exploring the Potential of Chat GPT in Personalized Obesity Treatment. Ann Biomed Eng 2023;51(9):1887-8.
8. Ismail AMA. Chat GPT in Tailoring Individualized Lifestyle-Modification Programs in Metabolic Syndrome: Potentials and Difficulties? Ann Biomed Eng 2023.
9. Schrauben SJ, Cavanaugh KL, Fagerlin A, et al. The Relationship of Disease-Specific Knowledge and Health Literacy With the Uptake of Self-Care Behaviors in CKD. Kidney Int Rep 2020;5(1):48-57.
10. Yeh J-Z, Wei C-j, Weng S-f, et al. Disease-specific health literacy, disease knowledge, and adherence behavior among patients with type 2 diabetes in Taiwan. BMC Public Health 2018;18(1):1062.
11. Berkman ND, Sheridan SL, Donahue KE, et al. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011;155(2):97-107.
12. Al Rashed WA, Bin Shihah AS, Alhomoud AS, et al. Knowledge, attitude, and practice toward glaucoma and its management among adult Saudi patients. Saudi J Ophthalmol 2020;34(4):261-5.
13. Rewri P, Kakkar M. Awareness, knowledge, and practice: a survey of glaucoma in north Indian rural residents. Indian J Ophthalmol 2014;62(4):482-6.
14. Banerjee A, Ahmad A, Bhalla P, Goyal K. Assessing the Efficacy of ChatGPT in Solving Questions Based on the Core Concepts in Physiology. Cureus 2023;15(8):e43314.
15. Májovský M, Černý M, Kasal M, et al. Artificial Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora's Box Has Been Opened. J Med Internet Res 2023;25:e46924.
16. Frosolini A, Franz L, Benedetti S, et al. Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. Eur Arch Otorhinolaryngol 2023.
17. Kumar M, Mani UA, Tripathi P, et al. Artificial Hallucinations by Google Bard: Think Before You Leap. Cureus 2023;15(8):e43313.
18. Robillard JM, Jun JH, Lai J-A, Feng TL. The QUEST for quality online health information: validation of a short quantitative tool. BMC Medical Informatics and Decision Making 2018;18(1):87.
19. Symons T, Davis JS. Creating concise and readable patient information sheets for interventional studies in Australia: are we there yet? Trials 2022;23(1):794.
20. originality.ai. Text Readability Checker. 2023. https://originality.ai/readability-checker.
21. Williamson JML, Martin AG. Analysis of patient information leaflets provided by a district general hospital by the Flesch and Flesch–Kincaid method. International Journal of Clinical Practice 2010;64(13):1824-31.
22. Posch N, Horvath K, Wratschko K, et al. Written patient information materials used in general practices fail to meet acceptable quality standards. BMC Fam Pract 2020;21(1):23.
23. Williams AM, Muir KW, Rosdahl JA. Readability of patient education materials in ophthalmology: a single-institution study and systematic review. BMC Ophthalmol 2016;16:133.
24. Harkness R, Hall G, Frangi AF, et al. The Pitfalls of Using Open Data to Develop Deep Learning Solutions for COVID-19 Detection in Chest X-Rays. Stud Health Technol Inform 2022;290:679-83.
25. Duffy G, Clarke SL, Christensen M, et al. Confounders mediate AI prediction of demographics in medical imaging. NPJ Digit Med 2022;5(1):188.

26. Ichhpujani P, Thakur S. Ethics and Artificial Intelligence: The Pandora's Box. In: Ichhpujani P, Thakur S, eds. Artificial Intelligence and Ophthalmology: Perks, Perils and Pitfalls. Singapore: Springer Singapore, 2021.
27. Li Y, Li Z, Zhang K, et al. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus 2023;15(6):e40895.