DOI:10.14744/bej.2025.76743
Beyoglu Eye J 2025; 10(3): 168-174

Original Article

Comparison of the Accuracy, Comprehensiveness, and Readability of ChatGPT, Google Gemini, and Microsoft Copilot on Dry Eye Disease

Dilan Colak,[1] Burcu Yakut,[2] Abdullah Agin[2]

[1]Department of Ophthalmology, University of Health Science, Beyoglu Eye Training and Research Hospital, Istanbul, Türkiye
[2]Department of Ophthalmology, University of Health Science, Haseki Training and Research Hospital, Istanbul, Türkiye

Abstract

Objectives: This study compared the performance of ChatGPT, Google Gemini, and Microsoft Copilot in answering 25 questions about dry eye disease and evaluated comprehensiveness, accuracy, and readability metrics.

Methods: The artificial intelligence (AI) platforms answered 25 questions derived from the American Academy of Ophthalmology's Eye Health webpage. Three reviewers assigned comprehensiveness (0–5) and accuracy (−2 to 2) scores. Readability metrics included Flesch-Kincaid Grade Level, Flesch Reading Ease Score, sentence/word statistics, and total content measures. Responses were rated by three independent reviewers. Readability metrics were also calculated, and platforms were compared using Kruskal–Wallis and Friedman tests with *post hoc* analysis. Reviewer consistency was assessed using the intraclass correlation coefficient (ICC).

Results: Google Gemini demonstrated the highest comprehensiveness and accuracy scores, significantly outperforming Microsoft Copilot ($p<0.001$). ChatGPT produced the most sentences and words ($p<0.001$), while readability metrics showed no significant differences among models ($p>0.05$). Inter-observer reliability was highest for Google Gemini (ICC=0.701), followed by ChatGPT (ICC=0.578), with Microsoft Copilot showing the lowest agreement (ICC=0.495). These results indicate Google Gemini's superior performance and consistency, whereas Microsoft Copilot had the weakest overall performance.

Conclusion: Google Gemini excelled in content volume while maintaining high comprehensiveness and accuracy, outperforming ChatGPT and Microsoft Copilot in content generation. The platforms displayed comparable readability and linguistic complexity. These findings inform AI tool selection in health-related contexts, emphasizing Google Gemini's strengths in detailed responses. Its superior performance suggests potential utility in clinical and patient-facing applications requiring accurate and comprehensive content.

Keywords: Artificial intelligence, ChatGPT, dry eye disease, Google Gemini, Microsoft Copilot

Introduction

Dry eye disease (DED) is a multifactorial disease of the ocular surface characterized by a loss of homeostasis of the tear film, and accompanied by ocular symptoms, in which tear film instability and hyperosmolarity, ocular surface inflammation and damage, and neurosensory abnormalities play etiological roles (1-4). The prevalence of DED varies widely, with estimates ranging from 5% to 50% of the adult

How to cite this article: Colak D, Yakut B, Agin A. Comparison of the Accuracy, Comprehensiveness, and Readability of ChatGPT, Google Gemini, and Microsoft Copilot on Dry Eye Disease. Beyoglu Eye J 2025; 10(3): 168-174.

Address for correspondence: Abdullah Agin, MD. Department of Ophthalmology, University of Health Science, Haseki Training and Research Hospital, Istanbul, Türkiye
Phone: +90 533 517 52 77 **E-mail:** abdullahagin@gmail.com

Submitted Date: May 21, 2025 **Revised Date:** July 16, 2025 **Accepted Date:** August 03, 2025 **Available Online Date:** September 25, 2025

Beyoglu Eye Training and Research Hospital - Available online at www.beyoglueye.com

population (2). The impact of DED is substantial, affecting millions worldwide and leading to discomfort, visual impairment, and a diminished quality of life (5). The variability in its presentation and the lack of a universally accepted diagnostic criterion further complicate its understanding and management (6).

The digital age has seen a surge in individuals turning to online resources for health information, and artificial intelligence (AI) chatbots have emerged as potential tools to provide readily accessible medical knowledge (7). However, the accuracy, comprehensiveness, and readability of the information provided by these AI platforms, especially for intricate medical conditions like DED, remain a critical concern. In addition to providing health information to patients, AI chatbots are increasingly integrated into medical education, supporting both undergraduate and postgraduate learners. These tools offer interactive learning experiences, assist in knowledge reinforcement, and serve as accessible resources for quick clinical reference, making them valuable for healthcare professionals and non-specialist users alike.

This study addresses this concern by evaluating the performance of three leading AI chatbots, ChatGPT, Google Gemini, and Microsoft Copilot, in answering a diverse set of questions about DED. The questions, formulated based on the American Academy of Ophthalmology's Eye Health webpage, encompass various aspects of the disease, including its definition, symptoms, causes, risk factors, diagnosis, treatment options, and impact on daily life. By systematically assessing the responses generated by these AI platforms for accuracy, comprehensiveness, and readability, we seek to provide a nuanced understanding of their capabilities and limitations in the context of DED. The study also examined inter-reviewer consistency in evaluating AI-generated responses, aiming to enhance the reliability and reproducibility of the evaluation process. Ultimately, this research endeavors to shed light on the potential and challenges of AI chatbots in disseminating accurate and comprehensible medical information about DED. By providing a comprehensive assessment of their performance, we hope to empower both healthcare providers and patients to make informed decisions about utilizing AI chatbots as adjuncts in the pursuit of improved patient education and healthcare outcomes. The insights gained from this study will contribute to the ongoing dialogue about the role of AI in healthcare communication, fostering a more informed and judicious use of these tools.

Methods

Data Source: Twenty-five questions were created based on the AAO Eye Health webpage and used to prompt each AI platform. Although these questions were not pilot-tested or formally validated, they were derived from a reputable patient information source to ensure clinical relevance and clarity. The questions cover a wide range of topics, including definitions, symptoms, causes, risk factors, diagnosis, treatment, and impact on daily life. The questions were designed to assess the AI platforms' ability to provide accurate, comprehensive, and readable information on this complex topic. Since the study does not involve any procedures related to patients, ethical committee approval is not required. The questions are highlighted in Table 1.

- AI platforms: Three prominent AI platforms were selected for evaluation:
- ChatGPT: A large language model developed by OpenAI, known for its conversational abilities and general knowledge.
- Google Gemini: A cutting-edge AI model developed by Google, designed to excel in natural language understanding and generation tasks.
- Microsoft Copilot: An AI-powered code completion and generation tool developed by Microsoft that is also capable of providing information on various topics.

Evaluation

- Independent reviewers: Three ophthalmologists with expertise in DED independently assessed the responses generated by each AI platform. This ensured an unbiased and expert evaluation of the information provided. All responses were anonymized before evaluation, and reviewers were blinded to the identity of the AI model that generated each response to minimize potential bias.
- Comprehensiveness: Each response was rated on a scale of 0 to 5, where 0 indicated no relevant information, and 5 indicated a fully comprehensive answer that addressed all aspects of the question.
- Accuracy: Each response was rated on a scale of −2 to 2, where −2 indicated utterly inaccurate information, 0 indicated partially accurate or incomplete information, and 2 indicated exact information.
- Readability metrics: To assess the readability of the AI-generated responses, several established metrics were calculated.
- Flesch–Kincaid grade level (FKGRL): This metric estimates the U.S. school grade level required to understand the text.
- Flesch reading ease score (FRES): This metric indicates how easy the text is to read, with higher scores representing easier readability.
- Average words per the sentence: This metric measures the average length of sentences in the text.
- Average syllables per word: This metric assesses the complexity of words used in the text.
- Total number of sentences: This metric provides the total count of sentences in the response.

Table 1. Patient-oriented questions on dry eye disease answered by ChatGPT, Google Gemini, and Microsoft Copilot

1.	What is the definition of dry eye disease?
2.	What are the common symptoms of dry eye disease?
3.	What are the primary causes of dry eye disease?
4.	What are the main tests used in the diagnosis of dry eye disease?
5.	How is the Schirmer test performed, and what does it measure?
6.	What are the effects of tear film layer disruption on vision?
7.	What is the prevalence of dry eye disease?
8.	What are the risk factors for dry eye disease?
9.	What is blepharitis, and how is it related to dry eye disease?
10.	What are the components of the tear film layer, and what functions do they serve?
11.	What are the primary treatment methods for dry eye disease?
12.	What is the long-term prognosis for dry eye disease?
13.	What are the effects of computer use on dry eye disease?
14.	What autoimmune diseases are associated with dry eye disease?
15.	How does dry eye disease affect the daily lives of patients?
16.	How is the tear break-up time (TBUT) test performed, and what does it measure?
17.	What corneal findings may be observed in patients with dry eye disease?
18.	How does contact lens use affect dry eye disease?
19.	How do hormonal changes impact dry eye disease?
20.	What are the effects of environmental factors on dry eye disease?
21.	How are Lissamine Green and Rose Bengal dyes used, and what do they measure?
22.	What is the prevalence of dry eye syndrome in children and adolescents?
23.	What pharmacological treatments are used in the management of dry eye disease?
24.	What are the non-pharmacological treatment methods for dry eye disease?
25.	What are the social and economic impacts of dry eye disease?

- Total number of words: This metric gives the total word count of the response.

Statistical Analysis

All statistical analyses were conducted using Statistical Package for the Social Sciences Statistics 23 (IBM Corp., Armonk, New York, USA). Initially, descriptive statistics (mean±standard deviation) were calculated for each variable across the three AI platforms. The Shapiro–Wilk test was conducted to evaluate the normality of the data for each variable. As the variables were not normally distributed, the Kruskal–Wallis test was used to compare differences among the three models. This test was used for Comprehensiveness, Accuracy, total number of sentences, and total number of words to evaluate whether at least one platform performed significantly differently from the others. If the Kruskal–Wallis test revealed a significant difference, Mann–Whitney U tests were conducted as *post-hoc* pairwise comparisons to identify which AI platforms differed from one another. This test was applied separately for each pair of platforms (ChatGPT vs. Google Gemini, ChatGPT vs. Microsoft Copilot, and Google Gemini vs. Microsoft Copilot). Bonferroni correction was applied to control for Type I error (adjusted P-value threshold: $p<0.0167$ for three comparisons. To analyze differences in FKGRL and FRES scores, a Friedman test was conducted, as these metrics were assessed across all three AI platforms on the same dataset. To assess inter-rater agreement in comprehensiveness and accuracy scores, the intraclass correlation coefficient (ICC) was calculated for each AI platform separately. A statistical significance threshold of $p<0.05$ was used.

Results

Google Gemini demonstrated the highest comprehensiveness ($p<0.001$) and accuracy ($p=0.003$) scores among the three AI platforms, followed by ChatGPT, while Microsoft Copilot consistently underperformed. Although Gemini

outperformed ChatGPT in comprehensiveness (p=0.014), their accuracy scores did not differ significantly (p=0.280) (Fig. 1). Readability scores such as FKGRL and FRES showed no significant differences among models (p=0.468 and p=0.289, respectively), but ChatGPT produced significantly longer responses in terms of sentence count and total word count (both p<0.001) (Table 2). Copilot had a higher average syllables-per-word score than ChatGPT (p=0.007) (Table 3 and Fig. 2). Inter-observer agreement was strongest for Gemini (ICC = 0.701, p<0.001) and weakest for Copilot (ICC = 0.495, p=0.022), suggesting greater consistency in expert evaluation for Gemini (Table 4 and Fig. 3).

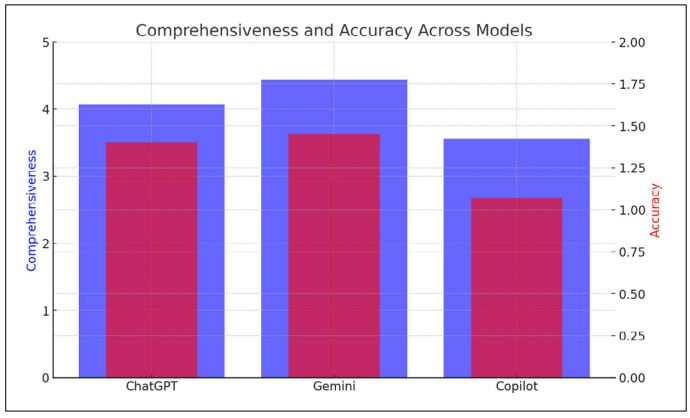

Figure 1. Comprehensiveness and accuracy across models. The blue bars represent the mean Comprehensiveness scores (left y-axis), while the red bars indicate the mean Accuracy scores (right y-axis) for each artificial intelligence model (ChatGPT, Gemini, and Copilot). Comprehensiveness scores (blue) range from 0 to 5, and accuracy scores (red) range from -2 to +2.

Discussion

Our study provides a comprehensive comparison of ChatGPT, Google Gemini, and Microsoft Copilot in answering DED-related questions, assessing their performance in terms of accuracy, comprehensiveness, readability, and inter-observer reliability. The findings highlight key differences in how these AI platforms generate medical information, offering valuable insights for both healthcare professionals and patients seeking reliable online health content.

Consistent with previous research on AI-generated medical information, ChatGPT and Google Gemini demonstrated comparable accuracy, confirming the ability of large language models to provide medically relevant responses. However, our study uniquely underscores the substantial differences in response comprehensiveness, with Google Gemini significantly outperforming both ChatGPT and Microsoft Copilot. Google Gemini's more detailed responses may enhance understanding of complex DED concepts, but they also introduce the potential risk of information overload, which could be challenging for some users. While extensive answers can be beneficial for professionals or highly engaged patients, they may also make it harder for general users to extract key takeaways efficiently.

Microsoft Copilot, initially developed for code generation, performed the weakest in both accuracy and comprehensiveness. This result reinforces the importance of task-specific AI tools. It highlights the need for users to carefully consider an AI model's intended purpose before relying on it for medical information. The significant performance gap between Microsoft Copilot and the other two platforms suggests that general-purpose AI models may not always be suitable for specialized domains such as medical education and patient counseling.

Table 2. Descriptive statistics and group comparisons of AI models

Metric	ChatGPT (mean±SD)	Gemini (mean±SD)	Copilot (mean±SD)	Range (min-max)	X^2	p
Comprehensiveness	4.07±0.55	4.44±0.56	3.56±0.61	2–5	21.72	<0.001*
Accuracy	1.40±0.32	1.45±0.50	1.07±0.52	–1–2	11.76	0.003*
Total sentence count	3.24±2.79	1.48±0.96	1.12±0.44	0–14	29.20	<0.001*
Total word count	47.04±13.26	25.04±7.32	20.56±10.50	7–90	43.08	<0.001*
Average sentence length	21.68±15.15	19.66±5.72	18.09±3.88	5–86	5.63	0.060
Average syllables per word	1.84±0.24	1.93±0.26	1.99±0.31	1.3–3.0	7.45	0.024*
FKGRL	15.11±4.89	15.91±4.50	15.39±4.81	7.12–27.17	1.52	0.468
FRES	28.14±16.78	21.38±21.30	21.49±23.06	–48.99–76.05	2.21	0.331

*Statistically significant. FRES: Flesch reading ease score; higher scores indicate easier readability, FKGRL: Flesch–Kincaid grade reading level; indicates U.S. school grade level required for comprehension, Avg. syllables/word: Average number of syllables per word, X^2: Kruskal–Wallis H test statistics, used for group comparison of non-normally distributed variables, Accuracy (–2 to +2), Comprehensiveness (0–5). SD: Standard deviation

Table 3. *Post-hoc* comparisons for statistically significant variables

Metric	Comparison	p	Direction of difference
Comprehensiveness	Gemini versus ChatGPT	0.014*	Gemini > ChatGPT
	ChatGPT versus Copilot	<0.001*	ChatGPT > Copilot
	Gemini versus Copilot	<0.001*	Gemini > Copilot
Accuracy	Gemini versus ChatGPT	0.280	No significant difference
	ChatGPT versus Copilot	<0.001*	ChatGPT > Copilot
	Gemini versus Copilot	<0.001*	Gemini > Copilot
Total sentence count	ChatGPT versus Gemini	<0.001*	ChatGPT > Gemini
	ChatGPT versus Copilot	<0.001*	ChatGPT > Copilot
	Gemini versus Copilot	0.014*	Gemini > Copilot
Total word count	ChatGPT versus Gemini	<0.001*	ChatGPT > Gemini
	ChatGPT versus Copilot	<0.001*	ChatGPT > Copilot
	Gemini versus Copilot	0.014*	Gemini > Copilot
Average syllables per word	ChatGPT versus Gemini	0.086	No significant difference
	ChatGPT versus Copilot	0.007*	Copilot > ChatGPT
	Gemini versus Copilot	0.210	No significant difference

*Statistically significant.

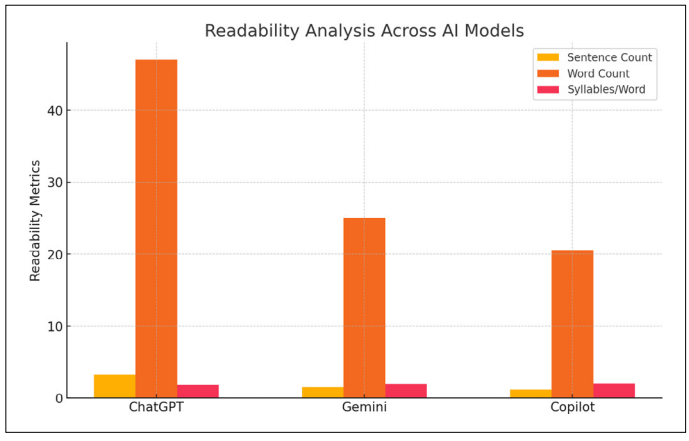

Figure 2. Readability analysis across artificial intelligence models. Readability metrics across the three models. Yellow bars represent the average sentence count, orange bars represent the average word count, and pink bars represent the average syllables per word for responses generated by each model.

Readability remains a critical factor in ensuring that AI-generated health information is accessible to a diverse audience. While ChatGPT produced the highest number of sentences and words, potentially making its responses more detailed, our analysis found no significant differences among the platforms in FKGRL and FRES readability scores. This suggests that, despite variations in response length, all three models generated content that is generally suitable for educated laypersons. However, given the well-documented correlation between readability and patient comprehension, further refinements in AI-generated medical content may be necessary to ensure accessibility for individuals with lower health literacy.

A particularly noteworthy finding of our study is the variation in inter-observer reliability. Google Gemini exhibited the highest agreement among evaluators (ICC = 0.701), suggesting that its responses were perceived as more consistently accurate and comprehensive across different reviewers. ChatGPT followed with moderate reliability (ICC = 0.578), while Microsoft Copilot had the lowest agreement (ICC = 0.495), indicating higher variability in how its responses were evaluated. This reinforces the notion that specific AI models may produce more stable and trustworthy outputs, which is a crucial consideration for both medical professionals and AI developers aiming to refine chatbot performance.

This finding aligns with studies demonstrating the efficacy of large language models in providing medically relevant information (1,2). However, our study uniquely highlights the significant difference in response length between these two platforms. While both ChatGPT and Google Gemini provided accurate and comprehensive answers, Google Gemini consistently generated more extensive responses. This may offer a deeper understanding of DED concepts, but it also raises concerns about information overload for some users. Microsoft Copilot, designed primarily for code generation, exhibited the lowest performance in this medical context, reinforcing the importance of task-specific AI mod-

Table 4. Inter-observer reliability

Model	Single measures ICC	Average measures ICC	p	Reliability level
ChatGPT	0.314	0.578	0.005*	Moderate
Gemini	0.439	0.701	<0.001*	High
Copilot	0.247	0.495	0.022*	Low to moderate

*Statistically significant. ICC: Intraclass correlation coefficient.

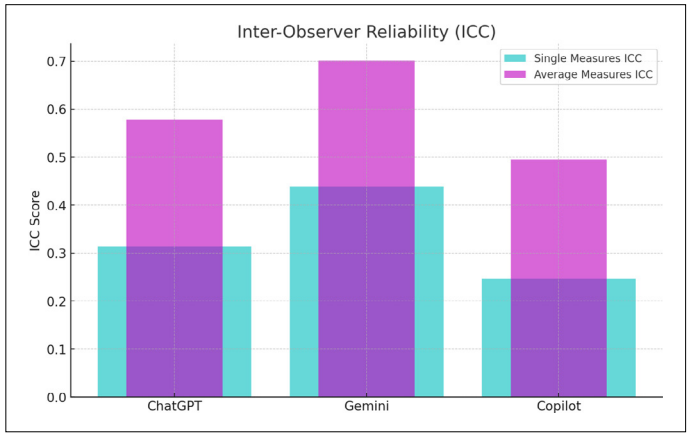

Figure 3. Inter-observer reliability (ICC). Turquoise bars show the single measures ICC, and purple bars show the average measures ICC for inter-observer agreement on scoring across the three artificial intelligence models.

els. Users should carefully consider an AI tool's intended purpose before relying on it for medical information. Our readability analysis revealed that ChatGPT generated significantly more sentences and words than both Google Gemini and Microsoft Copilot, suggesting that response length may impact readability perceptions. However, there were no significant differences among the platforms in terms of FKGRL and FRES readability scores, indicating that all three produced text suitable for educated laypersons. However, the complexity of the language used may pose a challenge for individuals with lower health literacy. This aligns with the broader concern raised in the literature regarding the need for AI-generated health information to be tailored to diverse audiences (5-7). Interestingly, Google Gemini exhibited the highest inter-reviewer consistency, indicating greater agreement among experts regarding the quality of its responses. This finding suggests that Google Gemini may be a more reliable in providing consistent and trustworthy information.

Haddad et al. (8) highlighted that ChatGPT's responses often required a higher level of education for comprehension compared to other platforms, which may hinder patient understanding. This complexity in readability is particularly concerning, given that many patients seek straightforward information about their conditions. The correlation between readability and patient comprehension is well-documented, with studies indicating that overly complex materials can lead to confusion and misinformation (9). Therefore, it is essential for AI tools to balance detail with accessibility to ensure that patients can quickly grasp the information provided. The reliance on AI chatbots for medical information is particularly pertinent in the context of chronic conditions such as glaucoma. Guler and Ertan Baydemir reported that approximately 43% of glaucoma patients utilize the Internet for medical information, highlighting the need for high-quality, reliable content (10). This shift in patient behavior underscores the importance of ensuring that AI-generated information is not only accurate but also presented in a manner that is easily digestible for patients with varying levels of health literacy. Moreover, the potential of AI chatbots in medical education has been explored in various studies. Haddad et al. (8) assessed the ability of ChatGPT to answer ophthalmology-related questions across different levels of training, indicating its utility as a supplementary educational tool for medical professionals. This aligns with findings from Davis et al., (9) who evaluated the application of AI in generating patient-centered information in other medical fields, suggesting a broader applicability of AI in enhancing patient education across specialties. The integration of AI into medical education and patient information dissemination could significantly improve the quality of care provided to patients. Desideri et al. (11) work further contributes to this discourse by examining the accuracy and applicability of AI chatbots in providing information about age-related macular degeneration. Her study categorized patient inquiries into general medical advice and pre- and post-intravitreal injection advice, revealing that while AI platforms provided accurate information, there were notable gaps in comprehensiveness and specificity (11). This highlights the necessity for continuous refinement of AI tools to ensure they meet the diverse informational needs of patients effectively. In addition, a study conducted by Guler and Ertan Baydemir evaluated the accuracy of responses provided by ChatGPT to 50 frequently asked questions by glaucoma patients, demonstrating a high level of concordance among ophthalmologists and generally accurate responses without observed significant inaccuracies

with potential harm (10). The research examined ChatGPT's overall accuracy in responding to typical patient inquiries related to glaucoma, highlighting its potential to address medical concerns and alleviate patient anxieties promptly. Despite the promising capabilities of AI chatbots, challenges remain in ensuring the accuracy and readability of the information they provide. The variability in performance among different platforms, as demonstrated by our study, suggests that continuous evaluation and refinement of AI tools are necessary to enhance their effectiveness in clinical settings. In addition, the complexity of responses generated by AI underscores the importance of tailoring information to meet the needs of diverse patient populations (8).

Conclusion

Our study has several limitations. The relatively small sample size of questions and reviewers may not fully capture the nuances of AI performance across a broader range of medical topics. Although the 0.4-point difference in mean accuracy was statistically significant, its clinical relevance remains context-dependent and may vary based on the complexity of the medical topic and the informational needs of the user. Since all prompts and responses were in English, the findings may not be generalizable to non-native speakers or multilingual populations. Language proficiency and cultural context could affect both comprehension and perceived quality of AI-generated content. Furthermore, the study did not assess qualitative attributes such as tone, empathy, or perceived trustworthiness of the responses, which may play a critical role in patient engagement and trust in AI-generated health content. Our comparative study reveals that while ChatGPT and Google Gemini can provide accurate and comprehensive information on DED, Google Gemini tends to offer more extensive responses. Microsoft Copilot, while proficient in other domains, may not be the optimal choice for complex medical queries. All platforms produced text suitable for educated laypersons, but further efforts are needed to improve readability for diverse audiences. This research emphasizes the importance of selecting the appropriate AI chatbot for specific tasks and highlights the potential of AI in revolutionizing healthcare communication. However, continued research and development are necessary to optimize AI's role in providing accessible, accurate, and user-friendly medical information. Future research should expand the scope of inquiry and explore objective DED measures of AI performance.

Disclosures

Ethics Committee Approval: Since the study does not involve any procedures related to patients, ethical committee approval is not required.

Conflict of Interest: None declared.

Funding: The authors declare that this study has received no financial support.

Use of AI for Writing Assistance: Not declared.

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

Peer-review: Externally peer-reviewed.

References

1. Craig JP, Nichols KK, Akpek EK, Caffery B, Dua HS, Joo CK, et al. TFOS DEWS II definition and classification report. Ocul Surf 2017;15:276–83. [CrossRef]
2. The epidemiology of dry eye disease: Report of the Epidemiology Subcommittee of the International Dry Eye WorkShop (2007). Ocul Surf 2007;5:93–107. [CrossRef]
3. Yucekul B, Mocan MC, Kocabeyoglu S, Tan C, Irkec M. Evaluation of long-term silicone hydrogel use on ocular surface inflammation and tear function in patients with and without meibomian gland dysfunction. Eye Contact Lens 2019;45:61–6. [CrossRef]
4. Colak D, Kocabeyoglu S, Karakaya J, Irkec M. Association of ocular surface and meibomian gland alterations with silicone hydrogel contact lens wear. Cont Lens Anterior Eye 2024;47:102093. [CrossRef]
5. Bron AJ, de Paiva CS, Chauhan SK, Bonini S, Gabison EE, Jain S, et al. TFOS DEWS II pathophysiology report. Ocul Surf 2017;15:438–510. Erratum in: Ocul Surf 2019;17:842. [CrossRef]
6. Wolffsohn JS, Arita R, Chalmers R, Djalilian A, Dogru M, Dumbleton K, et al. TFOS DEWS II Diagnostic Methodology report. Ocul Surf 2017;15:539–74. [CrossRef]
7. Li JO, Liu H, Ting DSJ, Jeon S, Chan RVP, Kim JE, et al. Digital technology, tele-medicine and artificial intelligence in ophthalmology: A global perspective. Prog Retin Eye Res 2021;82:100900. [CrossRef]
8. Haddad F, Saade JS. Performance of ChatGPT on ophthalmology-related questions across various examination levels: Observational study. JMIR Med Educ 2024;10:e50842. [CrossRef]
9. Davis RJ, Ayo-Ajibola O, Lin ME, Swanson MS, Chambers TN, Kwon DI, et al. Evaluation of oropharyngeal cancer information from revolutionary artificial intelligence chatbot. Laryngoscope 2024;134:2252–7. [CrossRef]
10. Güler MS, Baydemir EE. Evaluation of ChatGPT-4 responses to glaucoma patients' questions: Can artificial intelligence become a trusted advisor between doctor and patient? Clin Exp Ophthalmol 2024;52:1016–9. [CrossRef]
11. Ferro Desideri L, Roth J, Zinkernagel M, Anguita R. "Application and accuracy of artificial intelligence-derived large language models in patients with age related macular degeneration". Int J Retina Vitreous 2023;9:71. [CrossRef]