RESEARCH ARTICLE

Araştırma Makalesi

Yazışma adresi
Correspondence address
Bedia KESİMAL
Dışkapı Yıldırım Bayezıt Training and Research Hospital, Ophthalmology Department, Ankara, Türkiye

bediakesimal@gmail.com

Geliş Tarihi / *Received* : January 08, 2026
Kabul Tarihi / *Accepted* : May 12, 2026

Bu makalede yapılacak atıf
Cite this article as
Kesimal B, Kocamış Sİ. Appropriateness and Readability of Large Language Model Chatbot Responses to Frequently Asked Questions About Dry Eye Disease: Cross-Sectional Study

Akd Med J 2026;12: 1-8

Bedia KESİMAL
Dışkapı Yıldırım Bayezıt Training and Research Hospital, Ophthalmology Department, Ankara, Türkiye

Sücattin İlker KOCAMIŞ
Dışkapı Yıldırım Bayezıt Training and Research Hospital, Ophthalmology Department, Ankara, Türkiye

Appropriateness and Readability of Large Language Model Chatbot Responses to Frequently Asked Questions About Dry Eye Disease: Cross-Sectional Study

Kuru Göz Hastalığı Hakkında Sık Sorulan Sorulara Büyük Dil Modeli Sohbet Botlarının Yanıtlarının Uygunluğu ve Okunabilirliği: Kesitsel Araştırma

ABSTRACT

Objective

Large language model chatbots are increasingly consulted for medical information. This study evaluated the accuracy and readability of chatbot responses to common patient questions on dry eye disease.

Material and Methods

This cross-sectional study analysed responses from four chatbots (ChatGPT 3.5, ChatGPT 4.0, Google Gemini, and Microsoft Copilot) to fifty standardised questions about dry eye disease. Two ophthalmologists independently rated accuracy on a five-point Likert scale, with inter-rater agreement measured by Cohen's kappa. Readability was assessed using Flesch-Kincaid Grade Level, Gunning Fog Index, Coleman–Liau Index, Simple Measure of Gobbledygook, Flesch Reading Ease, and Reach score. Statistical tests included repeated-measures analysis of variance or Friedman tests with Bonferroni correction.

Results

Agreement between raters was excellent (kappa = 0.88). Google Gemini showed the highest accuracy (4.84 ± 0.37), followed by Microsoft Copilot (4.76 ± 0.48), ChatGPT 4.0 (4.42 ± 0.50), and ChatGPT 3.5 (4.32 ± 0.59; $p < 0.001$). Gemini and Copilot significantly outperformed both ChatGPT versions. Readability differed significantly ($p < 0.001$). ChatGPT 4.0 produced the simplest texts, with the lowest grade levels, highest Flesch Reading Ease and broadest Reach. Gemini and ChatGPT 3.5 generated more complex responses, while Copilot showed intermediate values.

Conclusions

Chatbots demonstrated complementary strengths. Gemini and Copilot were most accurate, whereas ChatGPT 4.0 was most readable and accessible. Chatbots may aid patient education in dry eye disease, but professional oversight remains necessary.

Keywords

Artificial intelligence, Chatbots, Dry eye disease, Large language models, Readability

DOI: 10.53394/akd.1859263

ÖZ

Amaç

Büyük dil modeli sohbet botları tıbbi bilgi için giderek daha fazla başvurulan kaynaklar haline gelmiştir. Bu çalışma, kuru göz hastalığına ilişkin sık sorulan hasta sorularına verilen yanıtların doğruluğunu ve okunabilirliğini değerlendirmeyi amaçladı.

Gereç ve Yöntemler

Bu kesitsel çalışmada dört sohbet botunun (ChatGPT 3.5, ChatGPT 4.0, Google Gemini ve Microsoft Copilot) kuru göz hastalığı ile ilgili elli standart soruya verdikleri yanıtlar analiz edildi. İki oftalmolog yanıtların doğruluğunu beş basamaklı Likert ölçeğiyle bağımsız olarak puanladı ve değerlendiriciler arası uyum Cohen'in kappa katsayısı ile ölçüldü. Okunabilirlik; Flesch-Kincaid Grade Level, Gunning Fog Index, Coleman–Liau Index, Simple Measure of Gobbledygook, Flesch Reading Ease ve Reach skoru ile değerlendirildi. İstatistiksel analizlerde tekrarlayan ölçümlerde varyans analizi veya Friedman testi ve Bonferroni düzeltmesi kullanıldı.

Bulgular

Değerlendiriciler arası uyum mükemmeldi (kappa = 0,88). En yüksek doğruluk Google Gemini'de (4,84 ± 0,37) izlendi; bunu Microsoft Copilot (4,76 ± 0,48), ChatGPT 4.0 (4,42 ± 0,50) ve ChatGPT 3.5 (4,32 ± 0,59; $p < 0,001$) takip etti. Gemini ve Copilot her iki ChatGPT versiyonuna anlamlı olarak üstün bulundu. Okunabilirlik anlamlı farklılık gösterdi ($p < 0,001$). ChatGPT 4.0 en basit metinleri üreterek en düşük sınıf düzeyleri, en yüksek Flesch Okuma Kolaylığı ve en geniş Reach skoruna ulaştı. Gemini ve ChatGPT 3.5 daha karmaşık yanıtlar üretirken Copilot orta düzey değerler sergiledi.

Sonuçlar

Sohbet botları tamamlayıcı güçlü yönler ortaya koydu. Gemini ve Copilot en yüksek doğruluğa sahipken, ChatGPT 4.0 en okunabilir ve erişilebilir yanıtları sundu. Sohbet botları kuru göz hastalığında hasta eğitimi için faydalı olabilir; ancak profesyonel gözetim gerekliliğini korumaktadır.

Anahtar Kelimeler

Yapay zeka, Sohbet botları, Kuru göz hastalığı, Büyük dil modelleri, Okunabilirlik

INTRODUCTION

Dry eye disease (DED) is a multifactorial disorder of the ocular surface that affects millions of individuals worldwide (1). It is characterised by tear film instability, increased osmolarity, ocular surface inflammation, and neurosensory abnormalities, leading to symptoms such as ocular discomfort, visual disturbance, and fluctuating vision (2). Prevalence estimates vary widely, ranging from 5% to 50%, depending on diagnostic criteria, geographic location, and population characteristics (3). The condition has a substantial impact on visual function, daily activities, and quality of life, making patient awareness and education essential for effective management (3).

In recent years, the internet has become a primary source of health information for patients (4). With advances in natural language processing, large language models (LLMs) based chatbots, including OpenAI's ChatGPT, Google's Gemini, and Microsoft's Copilot, have emerged as tools capable of generating detailed, context-specific responses to user queries (5). Their rapid adoption in healthcare related searches reflects their potential to improve access to medical information. However, the reliability, completeness, and readability of their outputs particularly in specialised fields like ophthalmology remain subjects of debate.

For patients with DED, the accuracy and clarity of information are crucial. Misleading, incomplete, or overly complex responses can lead to misunderstanding, poor treatment adherence, and delayed clinical intervention. Previous studies in ophthalmology have evaluated the performance of LLMs chatbots in subspecialties such as refractive surgery and corneal diseases (6, 7). Nevertheless, only a limited number of studies have systematically assessed chatbot performance in dry eye disease, despite its high prevalence and the need for clear, reliable patient education materials.

The purpose of this study was to evaluate the appropriateness and readability of responses generated by four widely used LLMs chatbots to a standardised set of 50 frequently asked questions about DED. By systematically analysing both the accuracy and accessibility of these responses, we aim to inform clinicians, patients, and developers about the potential benefits and limitations of AI-based tools in DED patient education.

MATERIALS and METHODS

Study Design and Aim

This cross-sectional study did not include human participants; therefore, approval from an ethics committee was not applicable. This study was designed to evaluate the accuracy and readability of responses generated by four LLMs based chatbots ChatGPT 3.5, ChatGPT 4.0 (OpenAI, San Francisco, CA, USA), Google Gemini (Google DeepMind, London, UK), and Microsoft Copilot (Microsoft, Redmond, WA, USA) to frequently asked questions related to DED. The primary aim was to compare the quality, reliability, and linguistic accessibility of chatbot generated content using validated assessment tools.

Question Selection

A total of 50 questions regarding DED were included. These were identified through a combined strategy: (1) reviewing the most commonly asked questions by patients during outpatient consultations, and (2) extracting the most frequently searched queries on the internet (Google Trends and related search results). This dual approach was adopted to ensure that the dataset captured both real-world patient concerns and online information-seeking behaviour.

Chatbot Response Collection

Each of the 50 questions was individually submitted to the four chatbots during March 2025. Responses were collected in their original, unedited form and stored for further analysis. All models were accessed in their publicly available versions without customisation or fine-tuning.

Accuracy Assessment

The accuracy of each response was evaluated by two experienced ophthalmologists (B.K. and S.İ.K). The ratings provided by the two ophthalmologists were compared using Cohen's kappa coefficient to assess inter-rater agreement. A five-point Likert scale was used:

1 = Strongly disagree (very poor or unacceptable, with high risk of harm);
2 = Disagree (inaccurate, with potentially harmful errors);
3 = Neutral (moderate inaccuracies, limited clinical significance);
4 = Agree (minor, non-harmful inaccuracies; generally good quality);
5 = Strongly agree (highly accurate, no errors, no harm risk).

Readability Analysis

Readability metrics were calculated for each response using the Readable software (Readable.com, London, UK). The following indices were extracted:

• Flesch-Kincaid Grade Level: estimates the U.S. school grade required to understand the text.
• Gunning Fog Index: measures textual complexity based on sentence length and the proportion of complex words.
• Flesch Reading Ease Score: evaluates ease of comprehension on a 0–100 scale, with higher scores indicating simpler readability.
• Coleman-Liau Index: determines readability based on characters per word and words per sentence.
• Simple Measure of Gobbledygook (SMOG): estimates the years of education needed to understand a piece of writing.
• Reach score (%): reflects the percentage of the general population expected to understand the text.
• Word and character counts: provide objective measures of text length.

All statistical analyses were conducted using Jamovi (version 2.3.28, Sydney, Australia). Normality was assessed using the Shapiro–Wilk test. For normally distributed variables, repeated-measures analysis of variance (ANOVA) was performed, while the Friedman test was used for non-normally distributed data. Post-hoc pairwise comparisons were conducted with Bonferroni correction where applicable. A p-value of <0.05 was considered statistically significant.

RESULTS

Accuracy of Chatbot Responses

The two ophthalmologists demonstrated a high level of agreement in their ratings (Cohen's kappa = 0.88), and the mean of their scores was included in the analysis. Google Gemini achieved the highest accuracy, followed by Microsoft Copilot, ChatGPT 4.0, and ChatGPT 3. The mean Likert accuracy scores are summarised in Table I and illustrated in Figure 1.

Table I. Likert accuracy scores of chatbot responses.

Chatbot	Mean ± SD	Median	Min–Max
ChatGPT-3.5	4.32 ± 0.59	4.0	3–5
ChatGPT-4	4.42 ± 0.50	4.0	4–5
Google Gemini	4.84 ± 0.37	5.0	4–5
Microsoft Copilot	4.76 ± 0.48	5.0	3–5

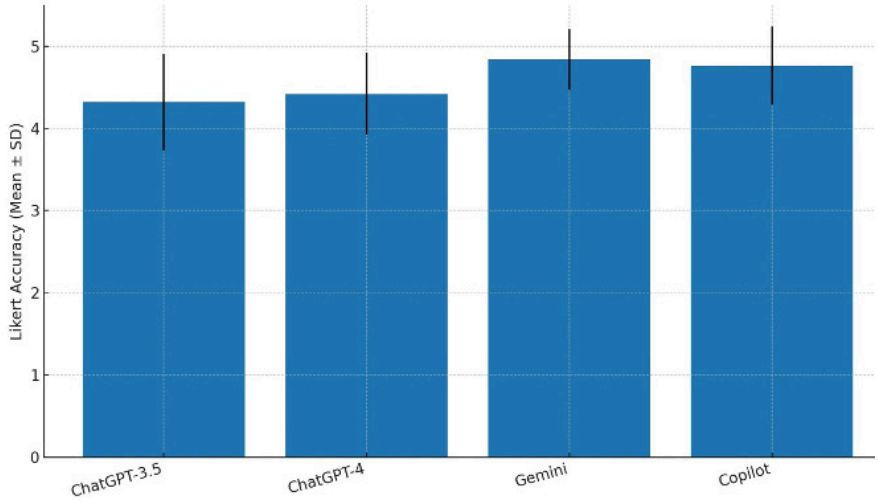

Figure 1. Accuracy of chatbot responses on dry eye disease questions.
Bar chart showing mean Likert accuracy scores (± SD) for ChatGPT 3.5, ChatGPT 4.0, Google Gemini, and Microsoft Copilot. Gemini and Copilot achieved significantly higher accuracy compared with both versions of ChatGPT ($p < 0.001$), while no significant difference was observed between Gemini and Copilot or between the two ChatGPT versions.

The Friedman test revealed significant differences among the four models ($\chi^2 = 33.4$, df = 3, $p < 0.001$). Post-hoc analysis demonstrated that Gemini and Copilot significantly outperformed both ChatGPT versions (all $p < 0.001$), while no difference was found between Gemini and Copilot or between ChatGPT 3.5 and ChatGPT 4.0. Detailed pairwise post-hoc comparisons are provided in Table II.

Table II. Post-hoc pairwise comparisons for Likert accuracy scores (Durbin–Conover test).

Comparison	Statistic	p-value
ChatGPT-3.5 vs ChatGPT-4	0.772	0.441
ChatGPT-3.5 vs Gemini	5.293	<0.001
ChatGPT-3.5 vs Copilot	4.521	<0.001
ChatGPT-4 vs Gemini	4.521	<0.001
ChatGPT-4 vs Copilot	3.749	<0.001
Gemini vs Copilot	0.772	0.441

Readability: Grade-Level Indices

Readability grade-level metrics are presented in Table III. Significant differences were found for all indices (all $p < 0.001$). ChatGPT 3.5 and Gemini consistently produced the most complex responses, whereas ChatGPT 4.0 was the most readable, with Copilot showing intermediate values. Post-hoc tests confirmed that ChatGPT 4.0 was significantly easier to read than all other models ($p < 0.01$ for all grade-level indices), whereas Gemini and ChatGPT 3.5 did not differ significantly from each other.

Readability: Ease, Reach, and Response Length

Flesch Reading Ease, Reach, and response length are presented in Table IV and illustrated in Figure 2. ChatGPT 4.0 achieved the highest readability and the broadest audience reach, while also generating the shortest responses. Google Gemini produced the longest responses with relatively low readability and reach. Microsoft Copilot showed intermediate performance for both readability and response length, while ChatGPT 3.5 resembled Gemini in producing more complex and less accessible content. Statistical comparisons confirmed significant differences across all models for these measures (all $p < 0.001$). Comprehensive post-hoc comparisons for readability indices are shown in Supplementary Table V.

Table III. Readability grade-level indices (mean ± SD).

Metric (Index)	GPT-3.5	GPT-4	Gemini	Copilot	Main finding
Flesch–Kincaid	13.0 ± 2.6	10.8 ± 2.4	13.3 ± 2.6	11.8 ± 2.0	GPT-4 easiest, Gemini hardest
Gunning–Fog	16.3 ± 3.2	13.5 ± 3.0	16.2 ± 3.0	14.8 ± 2.5	GPT-4 easiest, 3.5/Gemini hardest
Coleman–Liau	15.5 ± 2.5	13.9 ± 2.9	15.3 ± 2.9	14.6 ± 2.2	GPT-4 easiest, 3.5/Gemini hardest
SMOG	14.5 ± 2.0	12.2 ± 2.0	14.6 ± 2.1	13.5 ± 1.4	GPT-4 easiest, 3.5/Gemini hardest

Table IV. Readability ease, reach, and response length across chatbots (mean ± SD).

Metric	GPT-3.5	GPT-4	Gemini	Copilot
Flesch Reading Ease	28.8 ± 17.2	39.2 ± 17.4	29.2 ± 16.7	34.8 ± 15.1
Reach (%)	63.5 ± 18.0	79.0 ± 17.2	61.2 ± 18.4	71.2 ± 14.7
Word count	32.8 ± 5.9	29.8 ± 5.2	56.0 ± 15.3	42.3 ± 5.8
Character count	186 ± 28	163 ± 27	316 ± 96	234 ± 28

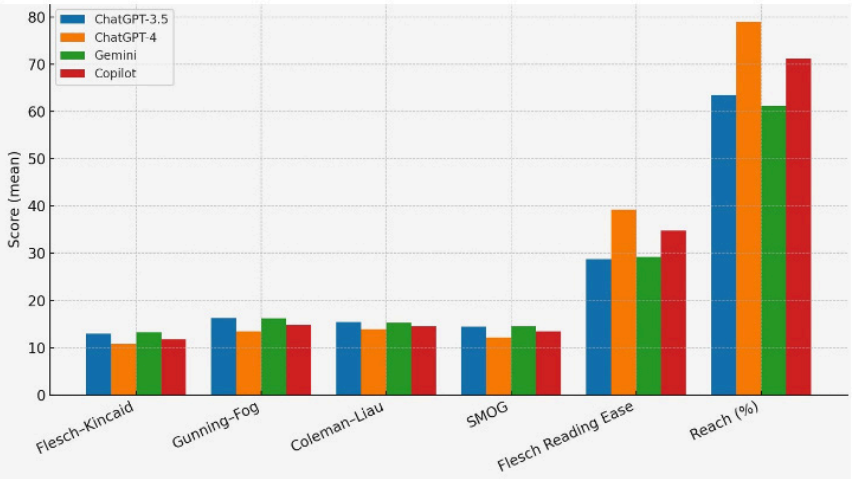

Figure 2. Readability metrics across chatbots. Grouped bar chart comparing six readability metrics (Flesch–Kincaid, Gunning–Fog, Coleman–Liau, SMOG, Flesch Reading Ease, and Reach). ChatGPT 4.0 consistently demonstrated the most readable outputs, with lower grade-level indices, higher Reading Ease, and broader Reach. ChatGPT 3.5 and Gemini generated more complex responses, whereas Copilot showed intermediate readability performance.

Table V. Post-hoc pairwise comparisons for readability indices (Bonferroni/Friedman).

Metric	Comparison	Mean Difference	Test statistic (t/χ^2)	p-value
Flesch–Kincaid	GPT-3.5 vs GPT-4	+2.25	7.79	<0.001
	GPT-3.5 vs Gemini	−0.26	−0.62	1.000
	GPT-3.5 vs Copilot	+1.18	3.59	0.005
	GPT-4 vs Gemini	−2.50	−7.48	<0.001
	GPT-4 vs Copilot	−1.06	−3.39	0.008
	Gemini vs Copilot	+1.44	4.19	<0.001
Gunning–Fog	GPT-3.5 vs GPT-4	+2.85	7.34	<0.001
	GPT-3.5 vs Gemini	+0.13	0.31	1.000
	GPT-3.5 vs Copilot	+1.52	3.31	0.010
	GPT-4 vs Gemini	−2.72	−7.30	<0.001
	GPT-4 vs Copilot	−1.34	−3.10	0.019
	Gemini vs Copilot	+1.38	3.46	0.007
Flesch Reading Ease	GPT-3.5 vs GPT-4	−10.40	−6.04	<0.001
	GPT-3.5 vs Gemini	−0.35	−0.15	1.000
	GPT-3.5 vs Copilot	−5.96	−3.11	0.019
	GPT-4 vs Gemini	+10.05	4.86	<0.001
	GPT-4 vs Copilot	+4.44	2.36	0.133
	Gemini vs Copilot	−5.61	−2.87	0.037
Coleman–Liau	GPT-3.5 vs GPT-4	+1.67	5.36	<0.001
	GPT-3.5 vs Gemini	+0.25	0.60	1.000
	GPT-3.5 vs Copilot	+0.91	2.70	0.057
	GPT-4 vs Gemini	−1.42	−4.17	<0.001
	GPT-4 vs Copilot	−0.75	−2.27	0.165
	Gemini vs Copilot	+0.66	2.04	0.282
SMOG	GPT-3.5 vs GPT-4	+2.29	9.44	<0.001
	GPT-3.5 vs Gemini	−0.06	−0.21	1.000
	GPT-3.5 vs Copilot	+1.01	3.44	0.007
	GPT-4 vs Gemini	−2.35	−9.91	<0.001
	GPT-4 vs Copilot	−1.28	−4.72	<0.001
	Gemini vs Copilot	+1.07	4.21	<0.001
Reach (%)	GPT-3.5 vs GPT-4	−15.5	6.49	<0.001
	GPT-3.5 vs Gemini	+2.3	1.03	0.302
	GPT-3.5 vs Copilot	−7.7	3.39	<0.001
	GPT-4 vs Gemini	+17.8	7.53	<0.001
	GPT-4 vs Copilot	+7.8	3.10	0.002
	Gemini vs Copilot	−10.0	4.42	<0.001

DISCUSSION

Artificial intelligence (AI), particularly LLMs, has rapidly emerged as a transformative tool in ophthalmology and other medical fields (8, 9). Patients increasingly turn to AI-based platforms for health information, often prior to consulting healthcare professionals (10). This trend highlights both the opportunities and risks associated with AI-generated medical content. On the one hand, these tools can improve access to knowledge, empower patients, and bridge information gaps in conditions such as DED (11). On the other hand, concerns remain regarding the accuracy, comprehensibility, and clinical appropriateness of their outputs. In this context, our study provides timely insights into the performance of widely used LLMs chatbots in addressing frequently asked questions about DED, emphasising key differences in accuracy and readability among the models evaluated.

In this study, Gemini and Copilot demonstrated superior accuracy, which is consistent with previous evidence indicating that newer large language models trained on broader and domain-specific datasets tend to achieve greater factual reliability (12). Gemini achieved the highest mean accuracy followed closely by Copilot, and both significantly outperformed ChatGPT 3.5 and 4.0. Similar trends have been reported in refractive surgery and corneal disease, where Gemini and Copilot provided more clinically appropriate outputs than ChatGPT models (6, 13). These results highlight the steady progress of large language model development, with successive versions offering incremental gains in precision.

Despite their comparatively lower accuracy, ChatGPT 4.0 responses were consistently the most readable, with the lowest grade-level requirements and the highest Flesch Reading Ease and Reach scores. This suggests that ChatGPT 4.0 is more accessible to a wider audience, an essential factor in patient education. Readability plays a critical role in ensuring that accurate information is actually understood and applied by patients (14). Prior studies have also noted that medical content, even when factually correct, may be ineffective if conveyed in overly complex language (15). Thus, ChatGPT 4.0's ability to produce simpler and shorter responses may represent a meaningful advantage in patient-centered communication.

Our findings revealed a trade-off between accuracy and readability: Gemini and Copilot achieved higher factual correctness, whereas ChatGPT 4.0 provided shorter and more accessible responses. This aligns with the results of Shi et al., who noted that advanced models often generate more accurate but linguistically complex outputs, while ChatGPT 4.0 produced content with greater readability (16). A recent study by Dihan et al. further highlighted this balance, showing that ChatGPT 4.0 could generate readable and high-quality patient education materials on dry eye disease, although these were sometimes limited in actionability (11). Taken together, these complementary findings suggest that Gemini and Copilot may be more suitable for clinician-oriented information, while ChatGPT 4.0 holds greater potential for patient education due to its accessible language.

The broader implications of these findings are twofold. First, AI-driven chatbots could play an increasingly valuable role in DED education, potentially improving patient understanding, enhancing self-management, and reducing the communication burden during clinical encounters (17). Second, the risks of misinformation, delayed treatment, and overreliance on imperfect outputs cannot be ignored. Importantly, none of the evaluated chatbots achieved perfect accuracy, underscoring the need for human oversight (18). Clinicians should advise patients to view chatbot responses as supplementary information rather than as substitutes for professional care.

This study has several limitations. First, the analysis was restricted to English-language outputs, and the findings may not generalise to other languages where training data are less robust. Second, only publicly available versions of the chatbots were evaluated, without customisation or fine-tuning for medical use. Third, the evaluation relied on expert-based assessments of accuracy and readability, rather than real-world patient comprehension or actionability, which should be addressed in future studies. In addition, the scope was limited to fifty questions, which may not fully capture the diversity of concerns expressed by patients with dry eye disease. Finally, because large language models evolve rapidly, our results represent a snapshot in time and may change with subsequent model updates. Despite these limitations, the integration of such tools into teleophthalmology platforms, combined with clinician oversight, may provide a promising strategy to balance accuracy and readability in patient education.

CONCLUSION

In conclusion, this study shows that current LLMs chatbots exhibit complementary strengths when responding to common questions about DED. Gemini and Copilot offer superior accuracy, while ChatGPT 4.0 provides greater readability and accessibility. These findings suggest that while LLMs hold considerable promise as adjunct tools for patient education in ophthalmology, they should be used cautiously and under clinical supervision. As LLMs continue to evolve, prioritizing both factual correctness and linguistic clarity will be critical to maximizing their utility in patient care.

Ethics Committee Approval
Ethics Committee approval was not required, as the study did not involve human participants, patient data, or any clinical intervention and was based exclusively on the evaluation of publicly available artificial intelligence–generated content. The study was conducted in accordance with relevant national regulations, institutional policies, and the principles of the Declaration of Helsinki.

Informed Consent
The study did not involve any direct patient contact, intervention, or identifiable personal data; therefore, written informed consent was not obtained.

Author Contributions
Concept – B.K., S.İ.K.; Design – B.K., S.İ.K.; Supervision – B.K., S.İ.K.; Resources – B.K., S.İ.K.; Materials – B.K., S.İ.K.; Data Collection and/or Processing – B.K., S.İ.K.; Analysis and/or Interpretation – B.K., S.İ.K.; Literature Search – B.K., S.İ.K.; Writing Manuscript – B.K., S.İ.K.; Critical Review – B.K., S.İ.K.

Conflict of Interest
The authors have no conflict of interest to declare.

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

Declaration of AI and AI-Assisted Technologies in the Writing Process: The authors declare that no artificial intelligence (AI) or AI-assisted technologies were used in the writing, preparation, editing, analysis, or revision of this manuscript. All content was created, reviewed, and approved solely by the authors.

REFERENCES

1. Pflugfelder SC, de Paiva CS. The Pathophysiology of Dry Eye Disease: What We Know and Future Directions for Research. Ophthalmol 2017;124(11S):S4-S13.

2. Pflugfelder SC, De Paiva CS, Villarreal AL, Stern ME. Effects of sequential artificial tear and cyclosporine emulsion therapy on conjunctival goblet cell density and transforming growth factor- beta2 production. Cornea 2008; 27(1):64–9.

3. Stapleton F, Alves M, Bunya VY, Jalbert I, Lekhanont K, Malet F, Na KS, Schaumberg D, Uchino M, Vehof J, Viso E, Vitale S, Jones L. TFOS DEWS II epidemiology report. Ocul surf 2017;15(3):334-65.

4. Swoboda CM, Van Hulle JM, McAlearney AS, Huerta TR. Odds of talking to healthcare providers as the initial source of healthcare information: updated cross-sectional results from the Health Information National Trends Survey (HINTS). BMC Fam Pract. 2018;19(1):146.

5. Rossettini G, Rodeghiero L, Corradi F, Cook C, Pillastrini P, Turolla A, Castellini G, Chiappinotto S, Gianola S, Palese A. Comparative accuracy of ChatGPT-4, Microsoft Copilot and Google Gemini in the Italian entrance test for healthcare sciences degrees: a cross-sectional study. BMC Med Educ. 2024;24(1):694.

6. Aydın FO, Aksoy BK, Ceylan A, Akbaş YB, Ermiş S, Kepez Yıldız B, Yıldırım Y. Readability and Appropriateness of Responses Generated by ChatGPT 3.5, ChatGPT 4.0, Gemini, and Microsoft Copilot for FAQs in Refractive Surgery. Turk J Ophthalmol. 2024;54(6):313-17.

7. Nichani PAH, Ong Tone S, AlShaker SM, Teichman JC, Chan CC. Use of Online Large Language Model Chatbots in Cornea Clinics. Cornea 2024;44(6):788-94.

8. Owens OL, Leonard M. A Comparison of Prostate Cancer Screening Information Quality on Standard and Advanced Versions of ChatGPT, Google Gemini, and Microsoft Copilot: A Cross-Sectional Study. Am J Health Promot. 2025;39(5):766-76.

9. Sevgi M, Antaki F, Keane PA. Medical education with large language models in ophthalmology: custom instructions and enhanced retrieval capabilities. Br J Ophthalmol. 2024;108(10):1354–61.

10. Raddatz N, Kettinger W, Coyne J. Giving to Get Well: Patients' Willingness to Manage and Share Health Information on AI-Driven Platforms. Communications of the Association for Information Systems. 2023;52:1017-49.

11. Dihan QA, Brown AD, Chauhan MZ, Alzein AF, Abdelnaem SE, Kelso SD, Rahal DA, Park R, Ashraf M, Azzam A, Morsi M, Warner DB, Sallam AB, Saeed HN, Elhusseiny AM. Leveraging large language models to improve patient education on dry eye disease. Eye (Lond). 2025;39(6):1115-22.

12. Chen X, Zhao Z, Zhang W, Xu P, Wu Y, Xu M, Gao L, Li Y, Shang X, Shi D, He M. EyeGPT for Patient Inquiries and Medical Education: Development and Validation of an Ophthalmology Large Language Model. J Med Internet Res. 2024;26:e60063.

13. Şensoy E, Çıtırık M. Performance of ChatGPT-3.5, Copilot and Gemini in answering English and Turkish questions related to ocular surface diseases and cornea: A comparison study. Turkish Journal of Clinical and Experimental Ophthalmology. 2025;20(6).

14. Wang J, Shi R, Le Q, Shan K, Chen Z, Zhou X, He Y, Hong J. Evaluating the effectiveness of large language models in patient education for conjunctivitis. Br J Ophthalmol. 2025;109(2):185-91.

15. Masoni M, Guelfi MR. Going beyond the concept of readability to improve comprehension of patient education materials. Intern Emerg Med. 2017;12(4):531-3.

16. Shi R, Liu S, Xu X, Ye Z, Yang J, Le Q, Qiu J, Tian L, Wei A, Shan K, Zhao C, Sun X, Zhou X, Hong J. Benchmarking four large language models' performance of addressing Chinese patients' inquiries about dry eye disease: A two-phase study. Heliyon. 2024;10(14):e34391.

17. Yang Z, Wang D, Zhou F, Song D, Zhang Y, Jiang J, Kong K, Liu X, Qiao Y, Chang RT, Han Y, Li F, Tham CC, Zhang X. Understanding natural language: Potential application of large language models to ophthalmology. Asia Pac J Ophthalmol (Phila). 2024;13(4):100085.

18. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a Large Language Model's Responses to Questions and Cases About Glaucoma and Retina Management. JAMA Ophthalmol. 2024;142(4):371-5.