GÜSBD 2026; 15(1): 148 - 156 | **Gümüşhane Üniversitesi Sağlık Bilimleri Dergisi** | **Araştırma Makalesi**
GUJHS 2026; 15(1): 148 - 156 | Gümüşhane University Journal of Health Sciences | Original Article

A Study on Accuracy and Readability of Frequently Asked Questions in Corneal Transplantation: Cross-Linguistic Evaluation of Large Language Models in Ophthalmic Patient Communication

Kornea Transplantasyonunda Sık Sorulan Soruların Doğruluğu ve Okunabilirliği Üzerine Bir Çalışma: Oftalmik Hasta İletişiminde Büyük Dil Modellerinin Çapraz Dilbilimsel Değerlendirmesi

Ayse BOZKURT OFLAZ[1], Muhammed Fatih DEMİRTAŞ[2], Abdullah ERDEM[3], Şule ACAR DUYAN[4]

ABSTRACT

This study assessed the accuracy, relevance, and readability of large language models (LLMs) in answering frequently asked questions (FAQs) on corneal transplantation. ChatGPT-3.5, ChatGPT 4o, Gemini, Copilot, and DeepSeek were evaluated using responses to 10 FAQs in Turkish and English. Accuracy was independently rated by two ophthalmologists, while readability was measured using the Ateşman and Çetinkaya indices for Turkish and six established English indices. Response length was analyzed by character, word, and sentence counts. The proportions of "correct and complete" responses were 72% for ChatGPT-3.5, 80% for ChatGPT 4o, 76% for Gemini, 70% for Copilot, and 71% for DeepSeek (p=0.47). In Turkish, Copilot produced the most complex texts, whereas DeepSeek achieved the highest readability based on the Ateşman index (58.2); according to the Çetinkaya index, ChatGPT 4o had the highest Turkish readability score (19.98). In English, DeepSeek demonstrated the highest readability across indices, while Gemini consistently generated the most complex outputs. Response length differed significantly between Turkish and English (p<0.001). Overall, LLMs provided largely accurate responses to patient-oriented corneal transplantation queries; however, variability in readability across languages indicates limitations in cross-linguistic communication. Enhancing linguistic adaptability and content accuracy remains essential for multilingual medical applications.

Keywords: Corneal transplantation, ChatGPT, Gemini, Copilot, DeepSeek

ÖZ

Bu çalışma, kornea transplantasyonuna ilişkin sık sorulan sorulara (SSS) büyük dil modellerinin (BDM) verdiği yanıtların doğruluk, uygunluk ve okunabilirliğini değerlendirdi. ChatGPT-3.5, ChatGPT 4o, Gemini, Copilot ve DeepSeek, Türkçe ve İngilizce 10 SSS'ye verdikleri yanıtlar üzerinden incelendi. Doğruluk iki oftalmolog tarafından bağımsız olarak değerlendirildi; okunabilirlik Türkçe için Ateşman ve Çetinkaya, İngilizce için altı yerleşik indeksle ölçüldü. Yanıt uzunlukları harf, kelime ve cümle sayıları temelinde analiz edildi. "Doğru ve eksiksiz" yanıt oranları ChatGPT-3.5 için %72, ChatGPT 4o için %80, Gemini için %76, Copilot için %70 ve DeepSeek için %71 olarak saptandı (p=0,47). Türkçede Copilot en karmaşık metinleri üretirken, Ateşman indeksine göre DeepSeek (58,2) en yüksek okunabilirliği sağladı; Çetinkaya indeksine göre ise ChatGPT 4o (19,98) en yüksek Türkçe okunabilirlik skoruna ulaştı. İngilizcede DeepSeek tüm indekslerde en yüksek okunabilirliği gösterirken, Gemini en karmaşık çıktıları üretti. Yanıt uzunlukları iki dil arasında anlamlı farklılık gösterdi (p<0,001). Genel olarak BDM'ler, hasta odaklı kornea transplantasyonu sorularına büyük ölçüde doğru yanıtlar sunmuştur; ancak diller arası okunabilirlik farklılıkları, çapraz dilsel iletişimde sınırlılıklara işaret etmektedir. Çok dilli tıbbi uygulamalarda hem dilsel uyarlanabilirliğin hem de içerik doğruluğunun artırılması gerekmektedir.

Anahtar Kelimeler: Kornea nakli, ChatGPT, Gemini, Copilot, DeepSeek

Highlights

* Artificial intelligence–based chatbots provide accurate answers to corneal transplant FAQs.
* Readability varies significantly across models and languages.
* Language adaptation is needed to improve clinical usefulness of AI tools.

Etik kurul izni gerekmemektedir

[1]Doktor Öğretim Üyesi, Ayse BOZKURT OFLAZ, Göz Hastalıkları, Selçuk Üniversitesi Tıp Fakültesi Hastanesi Göz Hastalıkları Bölümü, draysebozkurtoflaz@yahoo.com, ORCID: 0000-0001-5894-0220
[2]Doktor, Muhammed Fatih DEMİRTAŞ, Göz Hastalıkları, Selçuk Üniversitesi Tıp Fakültesi Hastanesi Göz Hastalıkları Bölümü, fth.drgn2@gmail.com, ORCID: 0009-0004-0182-9272
[3]Doktor, Abdullah ERDEM, Göz Hastalıkları, Selçuk Üniversitesi Tıp Fakültesi Hastanesi Göz Hastalıkları Bölümü, erdemabd@gmail.com, ORCID: 0009-0003-6253-7049
[4]Doktor Öğretim Üyesi, Şule ACAR DUYAN, Göz Hastalıkları, Selçuk Üniversitesi Tıp Fakültesi Hastanesi Göz Hastalıkları Bölümü, dr.sulenuracar@gmail.com, ORCID: 0000-0002-9319-0477

İletişim / Corresponding Author: Ayse BOZKURT OFLAZ
e-posta/e-mail: draysebozkurtoflaz@yahoo.com

Geliş Tarihi / Received: 30.06.2025
Kabul Tarihi/Accepted: 20.01.2026

INTRODUCTION

In recent years, there has been a notable surge in interest in using artificial intelligence-powered chatbots within the healthcare domain. These bots have demonstrated significant potential in disease diagnosis, treatment algorithms, and beyond (1, 2). However, incorporating chatbots into patient-oriented health education and patient information delivery remains a subject of debate, as some studies highlight their potential to improve access to medical information and patient engagement, whereas others raise concerns regarding information accuracy, readability, health literacy alignment, and the risk of misinformation (3, 4).

Notably, patients also frequently use chatbots as a source of information. While this trend appears to reduce reliance on health professionals and expedite the information process, it is a domain that necessitates validation concerning its accuracy, reliability, and comprehensibility, thus warranting further discussion.

As patients increasingly consult large language models (LLMs) for health-related information, concerns have emerged regarding the readability and comprehensibility of the responses they generate. Patients require access to accurate and reliable information that is presented in a clear and understandable manner. Previous evaluations across various medical conditions have demonstrated that chatbot-generated responses are often written at an academic level, underscoring the need for more user-friendly algorithms tailored to individuals with limited health literacy (3-5). Similar investigations have also addressed patient questions related to ophthalmic diseases (6-15). However, emerging LLMs such as DeepSeek have not yet been systematically evaluated in the context of ophthalmic patient education, representing a notable gap in the current literature.

Moreover, the readability indices applied in most previous studies are primarily standardized for English texts (9-12). Given the fundamental linguistic differences between Turkish and English—particularly in terms of suffixal structure, word formation, and sentence construction—direct application of English-based readability formulas to Turkish texts may yield incomplete or misleading assessments (16).

This study aimed to evaluate the accuracy, readability, and linguistic characteristics of responses generated by multiple LLMs to frequently asked patient questions regarding corneal transplantation in both Turkish and English. The primary research question was whether these models differ in terms of response accuracy, readability indices, and response length across languages, with the broader objective of informing the development of patient-oriented, user-friendly language models and improving patient–physician communication.

MATERIAL AND METHODS

As this study did not involve human participants, patient data, or identifiable personal information, ethics committee approval was not required, and the study was conducted in accordance with national ethical principles and regulations.

In this study, we evaluated the performance of artificial intelligence–based LLMs in responding to frequently asked patient questions related to corneal transplantation. The questions were derived from commonly encountered patient inquiries in our ophthalmology clinic and were designed to cover key domains, including the surgical procedure of corneal transplantation, postoperative recovery and follow-up, visual outcomes, graft rejection and failure, postoperative medication use, potential complications, and long-term prognosis.

A total of 10 questions were posed to ChatGPT-3.5, ChatGPT 4o, Google Gemini, Microsoft Copilot, and DeepSeek. The

questions were posed to each model in the same format, and the answers given by the models were recorded as they were, without alteration. There was no intervention in the content of the responses during the data collection process. Each chatbot was queried with the same set of ten identical questions on five separate occasions on the same day, with the chat history cleared between each session. The responses were evaluated according to two main criteria: appropriateness and readability. The appropriateness assessment was performed by two ophthalmologists, who categorized the responses into three groups: 'correct and adequate', 'correct but incomplete', and 'incorrect'. Inter-rater reliability between the two ophthalmologists was assessed using Cohen's kappa coefficient.

Turkish responses were scored using the Ateşman and Çetinkaya readability indices, which are validated tools specifically developed for Turkish texts. An analysis of English responses was conducted using multiple established readability indices, including the Flesch Reading Ease, Flesch–Kincaid Grade Level, Gunning Fog Index, Simple Measure of Gobbledygook (SMOG) Index, Coleman–Liau Index, and Automated Readability Index, which collectively assess text difficulty based on sentence length, word complexity, syllable structure, and character-based metrics.

In 1997, Ateşman adapted the Flesch Reading Ease calculation for Turkish texts. The Ateşman readability score is calculated as follows: 198.825- 40.175 x word length (total syllables divided by total words)- 2.610 x sentence length (total words divided by total sentences). The resulting readability score ranges from 0 to 100 (11-12). The Çetinkaya formula, developed in 2010, is expressed as follows: 118.823– 25.987x average word length – 0.971x average sentence length. Increases in the score indicate an increase in the readability of the text, irrespective of the formula employed (16).

The online tool Readable (https://app.readable.com/text/) was utilized to assess the readability of the text. The Flesch Reading Ease Score is a mathematical formula that considers both word and sentence length, and the resulting score ranges from 1 to 100. A higher score indicates greater readability; for example, a score of 70-80 corresponds to a child's reading level, while 30-49 means university-level readability (9).

The Flesch-Kincaid Grade Level is a measure of text comprehensibility according to the US educational system. The resulting number indicates the minimum grade level required to understand the text, with higher scores indicating more difficulty and lower scores indicating easier readability (16, 17).

The Gunning Fog Index is calculated based on the average sentence length of the number of complex words in the text. The result, expressed as a number, indicates the minimum level of education required to comprehend the text; a higher number is associated with a higher level of required education (6, 18).

The Coleman-Liau Index is distinct from other readability tests in that it utilizes the number of letters instead of syllables. Finally, the SMOG Index calculates the proportion of words containing three or more syllables. The index was developed to measure the comprehensibility of academic and medical texts, and it is widely used in health and education (9).

In addition to these parameters, the total number of letters, words, and sentences in each chatbot's response was recorded in both languages.

The data were analyzed using IBM SPSS Statistics 21 software. A one-way ANOVA test was employed for each metric to ascertain the disparities between the modules, and Tukey HSD post-hoc tests were conducted on the metrics that exhibited significant differences. $p<0.05$ was accepted as the significance level.

RESULTS

The inter-rater agreement between the two ophthalmologists was substantial, with a Cohen's kappa value of 0.78.

Based on the ophthalmologists' evaluations of the responses generated by the LLM chatbots, ChatGPT-3.5 produced 72%, ChatGPT 4o 80%, Gemini 76%, Copilot 70%, and DeepSeek 71% "correct and complete" responses (Figure 1). There was no statistically significant difference between the accuracy rates of the LLMs' responses (p=0.47).

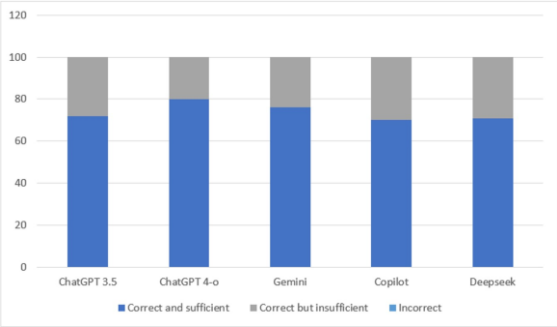

Figure 1. Accuracy assessment of responses generated by ChatGPT-3.5, ChatGPT 4o, Gemini, Copilot and DeepSeek

The Turkish and English readability index data are summarised in Table 1. A statistically significant difference was identified among the five chatbots in the Ateşman readability scores for the Turkish responses (p<0.001).

According to this index, DeepSeek achieved the highest mean readability score (58.2), indicating the most comprehensible Turkish texts, followed by ChatGPT 4o (56.98), whereas Copilot obtained the lowest score (43.57).

Similarly, based on the Çetinkaya readability index, ChatGPT 4o achieved the highest mean score (19.98), followed by DeepSeek (17.86), while Copilot again demonstrated the lowest readability level (7.62), with statistically significant differences observed among the groups (p<0.001). Notably, Gemini exhibited lower standard deviation values, suggesting more consistent readability performance.

Readability of the English responses was assessed using multiple established indices that evaluate text difficulty based on sentence length, word complexity, syllable structure, and character-based metrics.

According to the Flesch Reading Ease Score, DeepSeek achieved the highest readability, whereas Gemini produced the most complex texts. ChatGPT 4o exhibited the highest variability in English readability scores. Statistically significant differences were observed among the chatbots for the Flesch–Kincaid Grade Level, Gunning Fog Index, SMOG Index, Coleman–Liau Index, and Automated Readability Index (all p<0.001), with DeepSeek demonstrating the most favorable overall readability performance and Gemini the least across these measures.

Turkish and English response texts were further evaluated in terms of letter, word, and sentence counts. Significant differences were observed between languages and among chatbots (Table 2). The responses generated by ChatGPT-3.5 demonstrated substantial variability across multiple evaluated parameters, including readability indices and response length metrics, suggesting less consistency in output structure. In contrast, ChatGPT 4o and Gemini produced more homogeneous and balanced responses, characterized by narrower standard deviations and greater uniformity in sentence construction and word usage. Copilot responses were generally shorter and exhibited lower variance, indicating a tendency toward concise and standardized output generation. While this brevity may enhance accessibility, it may also limit the depth of information provided. DeepSeek, in comparison, consistently yielded the highest values across nearly all quantitative metrics, including letter, word, and sentence counts, and displayed pronounced extremes in both response length and readability measures, reflecting a more expansive but less controlled response generation pattern (Table 2).

Table 1. Mean values and standard deviation of readability index parameters in Turkish and English responses

	ChatGPT-3.5	ChatGPT 4o	Gemini	Copilot	DeepSeek	*p* value
Answers in Turkish						
Ateşman Readability Index	47.76±8.48	56.98±11.16	47.75±5.5	43.57±17.34	58.2±7.91	<0.001
Çetinkaya Readability Index	10.22±9.56	19.98±13.81	10.32±5.6	7.62±11.21	17.86±8.63	<0.001
Answers in English						
Flesch Reading Ease	44.96±8.09	43.05±12.66	38.28±5.84	41.26±7.46	45.17±6.69	<0.001
Flesch-Kincaid Grade Level	10.92±1.41	10.86±1.91	12.10±1	11.43±1.67	10.05±1.28	<0.001
Gunning Fog Index	14.01±1.63	14.39±2.05	15.59±1.17	14.77±1.86	13.28±1.43	<0.001
SMOG* Index	12.98±1.12	13.08±1.31	14.12±0.79	13.45±1.22	12.23±1.13	<0.001
Coleman-Liau Index	12.56±1.53	13.49±2.07	13.8±1.27	13.72±1.65	12.78±1.36	<0.001
Automated Readability Index	10.74±1.65	10.99±2.10	12.13±1.35	11.68±2.27	9.7±1.64	<0.001

**SMOG= Simple Measure of Gobbledygook*

Table 2. The mean and standard deviation of the letter count, word count, and sentence count in the Turkish and English responses across all chatbots.

	ChatGPT-3.5	ChatGPT 4o	Gemini	Copilot	DeepSeek	*p* value
Answers in Turkish						
Letter Count	1325.68±816.76	1689.22±360.18	1859.94±304.48	812.7±216.84	1979.1±594.33	<0.001
Word Count	195.98±123.26	256.86±51.63	274.72±43.97	119.14±32.63	296.86±90	<0.001
Sentence Count	15.96±11.38	23.2±6.48	21.24±3.64	8.74±2.63	32.56±13.57	<0.001
Answers in English						
Letter Count	1311.28 ±798.23	1788.36±380.42	1932.26±302.92	812.5±212.13	2064.62±628.1	<0.001
Word Count	255.72±157.11	336.08±72.94	362.28±55.99	152.68±42.39	392.24±124.45	<0.001
Sentence Count	16.88±11.56	23.78±6.61	21.64±3.61	9.82±2.57	33.52±13.56	<0.001

DISCUSSION

In recent years, an increase has been observed in patients utilizing artificial intelligence-supported applications to address medical queries. Ensuring the accuracy of information accessed through these applications is paramount (1, 10). LLMs undergo a daily process of self-updating to provide precise information promptly. This is a natural progression in technological advancement, whereby the accuracy of responses from the same model version improves over time. It has been asserted that all chatbots provide an acceptable level of correct information in their responses; however, difficulties in understanding may arise because this information exceeds the health literacy level of patients (3, 9, 19-21). In our study, we analyzed chatbot responses in two main metrics: accuracy and readability.

In the present study, questions concerning corneal transplantation were directed to ChatGPT-3.5, ChatGPT 4o, Gemini, Copilot, and DeepSeek. It was observed that none of the responses received could be categorized as 'incorrect'; instead, the answers were either 'correct and complete' or 'correct but incomplete'. While no statistically significant difference was identified between the accuracy rates of the five chatbots, the highest percentage of correct answers was observed in ChatGPT 4o, followed by DeepSeek. In contrast, the lowest percentage of 'correct and complete' answers was seen in Copilot. It is noteworthy that, given the recent introduction of DeepSeek as an AI tool, there is currently a paucity of literature about this chatbot. The relatively lower proportion of “correct and complete” responses observed for Copilot may be related to its response generation strategy, which appears to prioritize brevity

and generalization over detailed medical explanation. While concise outputs may enhance accessibility for lay users, this approach can result in the omission of clinically relevant details, particularly in complex topics such as corneal transplantation. As a consequence, responses that are broadly accurate but lack sufficient depth may be more frequently classified as "correct but incomplete" rather than fully correct.

In the Ateşman calculation of Turkish texts, DeepSeek received the highest score, while Copilot received the lowest score. Unlike the other modules, which scored in the range of 40-49, DeepSeek and ChatGPT 4o were suitable for the 11th-12th grade level with an average score of 58.2 and 56.98, respectively. According to Çetinkaya's calculation, the average performance of all chatbots received scores ranging from 0 to 29, indicating a university-level difficulty in terms of readability. In particular, ChatGPT 4o and DeepSeek produced more comprehensible texts, while Copilot produced the most challenging texts according to this measure. The minimum standard deviation of the scores obtained from Gemini's responses indicates a more balanced distribution in terms of readability. The study of Ateşman's readability index in medical data is limited to Duymaz et al. study of laryngeal cancer websites, Dağdelen's study of amblyopia information on large hospital websites, and Akkuş's study of prospectus readability (16, 17, 22). We have not yet found any information in the literature on using chatbots in responses.

Despite the statistically significant differences in the responses of the chatbots to the Flesch Reading Ease Score, all of them can be categorized as being between 30-49 points, i.e. the difficult level (university level). In the Flesch-Kincaid Grade Level calculation, a score above 12 points corresponds to the high school graduate level, a definition only met by Gemini. The readability of the others was at an easier level on this scale. The Gunning Fog Index indicated levels of 13-15 points, categorized as 'difficult' (university level), while the SMOG Index indicated levels of 13-15 points, categorized as 'university level'. The Coleman-Liau Index interpreted DeepSeek and ChatGPT-3.5 as high school level, while the remaining bots were measured as university level. The Automated Readability Index, except for Gemini, exhibited a high school level range, while Gemini demonstrated a university level range. In the study conducted by Duymaz et al. (22), Gemini produced more readable answers than ChatGPT in retinopathy of prematurity questions. In our study, the most difficult responses regarding readability for questions related to corneal transplantation were Copilot for Turkish texts and Gemini for English texts. This finding may be related to differences in how individual models adapt their response generation strategies to distinct linguistic structures. In Turkish responses, the lower readability observed for Copilot may be associated with the use of dense informational content combined with relatively complex sentence constructions and longer word forms. Given the agglutinative nature of the Turkish language, such linguistic characteristics can disproportionately reduce readability scores, even when the overall response length is relatively short.

The observed differences in accuracy and readability among the evaluated LLMs may be attributed to several underlying factors. First, variations in training data composition and optimization objectives may influence how models balance informational completeness with linguistic simplicity. Models trained with a stronger emphasis on conversational or user-oriented outputs may prioritize clearer sentence structures and shorter word lengths, resulting in higher readability scores.

Second, differences in response generation strategies, such as tendencies toward more concise versus more detailed explanations, may have contributed to variability in both readability indices and response length. While longer responses may allow for more comprehensive explanations, they can also increase linguistic complexity, particularly in medical contexts. It should be emphasized

that these interpretations are exploratory and based on observable output characteristics, as the proprietary nature of model architectures and training processes limits definitive conclusions regarding causal mechanisms.

While the number of sentences was quite low in ChatGPT-3.5 and Copilot, DeepSeek responses gave different results each time. Regarding word count, ChatGPT 4o and Gemini responses were more similar. Again, the DeepSeek answers were widely distributed. ChatGPT-3.5 and DeepSeek had a variable distribution for the number of letters.

A subsequent analysis of the English and Turkish versions revealed that the English version contained more sentences and words in all chatbots. This phenomenon can be attributed to the linguistic structure of English and Turkish. A study by Aydın et al. found that ChatGPT exhibited fewer words and sentences than other chatbots, while Gemini demonstrated a higher number of words and sentences in questions related to refractive surgery (9). In the present study, the lowest number of letters, words, and sentences in both versions of the answers was in Copilot and the highest number was in DeepSeek. In the Turkish readability analysis, in contrast, Copilot was more difficult to read with a shorter text, while DeepSeek was more easily read with a longer text.

A notable strength of this study is the inclusion of DeepSeek, a recently introduced LLM for which published data in medical and ophthalmic contexts remain limited. By directly comparing DeepSeek with more established models such as ChatGPT, Gemini, and Copilot, this study contributes novel evidence regarding the performance of emerging language models in patient-oriented ophthalmic communication.

CONCLUSION AND RECOMMENDATIONS

The study revealed that the level of accuracy exhibited by the chatbots in their responses to patients' inquiries regarding corneal transplantation is deemed to be within an acceptable range. However, discrepancies were observed between the Turkish and English metrics of the chatbots, suggesting that they might not fully align with international standards when dealing with disparate language structures. These findings underscore the necessity to enhance not only the precision of chatbot responses but also their comprehensibility in terms of linguistic structure.

From a clinical perspective, LLM-based chatbots may support patient education in ophthalmology; however, clinician supervision remains essential, particularly in languages with complex morphological structures such as Turkish.

This study has several limitations that should be considered when interpreting the findings. First, the number of frequently asked questions included in the analysis was limited. Although the selected questions reflect common patient concerns regarding corneal transplantation, a larger and more diverse question set might yield different readability and accuracy profiles.

Second, the performance of LLMs is inherently dependent on their specific versions and update cycles. As these models undergo continuous training and optimization, the accuracy and readability of their responses may change over time. Therefore, the findings of this study represent a time-specific evaluation and may not fully reflect future model performance.

Third, although multiple LLMs were evaluated, including an emerging model such as DeepSeek, the rapidly evolving nature of artificial intelligence systems limits the generalizability of direct comparisons across models and versions. Differences in training data, update frequency, and response generation strategies may have influenced the observed outcomes.

From a practical perspective, these findings have important implications for healthcare professionals involved in patient education. Although LLMs can generally

provide accurate information, the readability of their responses may exceed the health literacy level of many patients; therefore, chatbot-generated content should be used as a supplementary resource rather than a substitute for direct patient–physician communication. Healthcare providers may guide patients toward reliable digital tools while actively verifying and contextualizing the information obtained, and collaboration between clinicians and developers may further support the optimization of language models aligned with patient literacy levels and clinical communication needs.

A further limitation of this study is the relatively small number of questions included in the analysis. Although the selected ten questions reflect common patient concerns regarding corneal transplantation, the limited sample size may have reduced the statistical power of the analyses. Consequently, some differences between models may not have reached statistical significance despite observable trends. Future studies incorporating a larger and more diverse set of patient questions may provide more robust statistical power and allow for more definitive comparisons across models and readability metrics.

Disclosure Statement

The authors declare that they have no competing interests.

Funding

No funding was received to assist with the preparation of this manuscript.

Financial Interest

The authors have no relevant financial or non-financial interests to disclose.

Authors' Contributions

M.F.D. and A.E.; Data collection, Formal analysis, Methodology, Writing: A.B.O. and Ş.A.D.; Research, Conceptualization, Data collection, Formal analysis, Methodology, Writing-Review, Editing, Supervision, Project Management. All authors have read and accepted the published version of the article.

REFERENCES

1. Thirunavukarasu AJ. How Can the Clinical Aptitude of AI Assistants Be Assayed? Journal of Medical Internet Research. 2023; 25: e51603. doi:10.2196/51603.

2. Shakhovska N, Shebeko A, Prykarpatskyy Y. A Novel Explainable AI Model for Medical Data Analysis. Journal of Artificial Intelligence and Soft Computing Research. 2024; 14(2): 121-137.https://doi.org/10.2478/jaiscr-2024-0007.

3. Abreu AA, Murimwa GZ, Farah E, et al. Enhancing Readability of Online Patient-Facing Content: The Role of AI Chatbots in Improving Cancer Information Accessibility. Journal of the National Comprehensive Cancer Network. 2024; 22(2D): e237334. https://doi.org/10.6004/jnccn.2023.7334.

4. Lamb LR, Baird GL, Roy IT, et al. Are English-language online patient education materials related to breast cancer risk assessment understandable, readable, and actionable? The Breast. 2022; 61: 29-34. https://doi.org/10.1016/j.breast.2021.11.012.

5. Eppler MB, Ganjavi C, Knudsen JE, et al. Bridging the gap between urological research and patient understanding: the role of large language models in automated generation of layperson's summaries. Urology practice. 2023; 10(5): 436-443. doi:10.1097/UPJ.0000000000000428.

6. Eid K, Eid A, Wang D, et al. Optimizing ophthalmology patient education via ChatBot-generated materials: readability analysis of AI-generated patient education materials and the American Society of Ophthalmic Plastic and Reconstructive Surgery Patient Brochures. Ophthalmic Plastic & Reconstructive Surgery. 2024; 40(2): 212-216. doi:10.1097/IOP.0000000000002549.

7. Yalla GR, Hyman N, Hock LE, et al. Performance of Artificial Intelligence Chatbots on Glaucoma Questions Adapted From Patient Brochures. Cureus. 2024; 16(3). doi:10.7759/cureus.56766.

8. Sonmezoglu BG, Sonmezoglu HI. Comparative Analysis of AI Chatbots ChatGPT, Gemini, and Copilot's Answers to Common Cataract Questions. Pakistan Journal of Ophthalmology. 2024; 40(4). https://doi.org/10.36351/pjo.v40i4.1887.

9. Aydın FO, Aksoy BK, Ceylan A, et al. Readability and Appropriateness of Responses Generated by ChatGPT-3.5, ChatGPT 4.0, Gemini, and Microsoft Copilot for FAQs in Refractive Surgery. Turkish Journal of Ophthalmology. 2024; 54(6): 313-7. doi: 10.4274/tjo.galenos.2024.28234.

10. Doğan L, Özçakmakcı GB, Yılmaz İE. The Performance of Chatbots and the AAPOS Website as a Tool for Amblyopia Education. Journal of Pediatric Ophthalmology & Strabismus. 2024; 61(5): 325-331. https://doi.org/10.3928/01913913-20240409-01.

11. Pushpanathan K, Lim Z, Yew S, et al. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. iScience. 2023; 26. https://doi.org/10.1016/j.isci.2023.108163.

12. Tan D, Tham Y, Koh V, et al. Evaluating Chatbot responses to patient questions in the field of glaucoma. Frontiers in Medicine. 2024; 11. https://doi.org/10.3389/fmed.2024.1359073.

13. Oflaz AB, Duyan SA. Evaluating the accuracy, readability, and relevance of answers generated by large language models for frequently asked questions about cataract and cataract

surgery. 2025; 5(2): 103-110. https://doi.org/10.14744/eer.2025.44154.

14. Sabaner M, Anguita R, Antaki F, et al. Opportunities and Challenges of Chatbots in Ophthalmology: A Narrative Review. Journal of Personalized Medicine. 2024; 14. https://doi.org/10.3390/jpm14121165.

15. Güler M, Baydemir E. Evaluation of ChatGPT-4 responses to glaucoma patients' questions: Can artificial intelligence become a trusted advisor between doctor and patient?. Clinical & Experimental Ophthalmology. 2024; 52. https://doi.org/10.1111/ceo.14451.

16. Akkuş M. Evaluation of the Readability of Antidepressant Drug Package Inserts Commonly Used in Psychiatry. Osmangazi Tıp Dergisi. 2023; 45(5): 689-695. https://doi.org/10.20515/otd.1260211.

17. Dağdelen K. Readability Assessment of Patient Information Texts on Amblyopia on the Websites of Major Hospitals in Turkey. British Journal of Multidisciplinary and Advanced Studies. 2023; 4(4): 45-51. https://doi.org/10.37745/bjmas.2022.0269.

18. Gunning R. The technique of clear writing. 1952; McGraw-Hill.

19. Ahmed HS, Thrishulamurthy CJ. Evaluating ChatGPT's efficacy and readability to common pediatric ophthalmology and strabismus-related questions. European Journal of Ophthalmology. 2024; 11206721241272251. https://doi.org/10.1177/11206721241272251.

20. Engin CD, E Karatas, Ozturk T. Exploring the role of ChatGPT-4, BingAI, and Gemini as virtual consultants to educate families about Retinopathy of Prematurity. Children. 2024; 11(6): 750. doi: 10.3390/children11060750.

21. Huang G, Fang CH, Agarwal N, et al. Assessment of online patient education materials from major ophthalmologic associations. JAMA Ophthalmology. 2015; 133(4): 449-454. doi:10.1001/jamaophthalmol.2014.6104

22. Duymaz YK, Erkmen B, Sahin S, Tekin AM. Evaluation of the readability of Turkish online resources related to laryngeal cancer. European Journal of Therapeutics. 2024; 29(2): 168-172. https://doi.org/10.58600/eurjther.20232902-449.y.