DOI: 10.14744/eer.2025.93723
Eur Eye Res 2026;6(1):60–69

EUROPEAN EYE RESEARCH

ORIGINAL ARTICLE

Use of large language models in Turkish information materials for glaucoma patient education: evaluation of readability, accuracy and comprehensiveness

Ali Dal,[1] Murat Erdag,[2] Betul Dikme,[1] Bunyamin Kutluksaman[1]

[1]Department of Ophthalmology, Tayfur Ata Sokmen Faculty of Medicine, Mustafa Kemal University, Hatay, Turkiye
[2]Department of Ophthalmology, Fırat Faculty of Medicine, Fırat University, Elazığ, Turkiye

Abstract

Purpose: This study aims to evaluate the readability of the Turkish Ophthalmology Association's (TOA) glaucoma patient education brochure and to assess the capabilities of GPT-4.0, Gemini, and DeepSeek in generating Turkish patient education materials with respect to readability, accuracy, and comprehensiveness.

Methods: The TOA's patient education brochure on glaucoma was evaluated for readability using the Ateşman and Bezirci-Yilmaz formulae. The questions from the TOA booklets were presented independently to the GPT-4.0, Gemini, and DeepSeek models. The replies generated by these models were readability tested using the same formulas. In addition, qualified ophthalmologists evaluated the accuracy and comprehensiveness of the artificial intelligence (AI)-generated responses. AI-generated responses were converted to Q1 and Q2 formats to test text simplification. These versions were reevaluated for readability, accuracy, and comprehensiveness to see if simplification increased intelligibility without affecting medical accuracy.

Results: The TOA brochure had a higher readability level than the recommended patient education standard. Bezirci-Yilmaz scores showed that Gemini and DeepSeek had significantly lower readability than the TOA brochure (p=0.007 and p=0.033, respectively), whereas GPT-4.0 showed no significant difference (p=0.077). Ateşman scores indicated no significant difference between TOA and AI-generated texts. Gemini showed significantly higher comprehensiveness than GPT-4.0 (p=0.042), whereas accuracy scores did not differ significantly among the models. Readability improved for Gemini following simplification (p=0.013 and p=0.005, respectively), whereas GPT 4.0 and DeepSeek remained unchanged. After simplification, the comprehensiveness score decreased for Gemini, whereas GPT-4.0 and DeepSeek maintained their comprehensiveness.

Conclusion: While large language models hold promise for use as glaucoma patient information materials, it is essential to rigorously evaluate the accuracy and comprehensiveness of the content they produce.

Keywords: Glaucoma; large language models; readability.

Cite this article as: Dal A, Erdağ M, Dikme B, Kutluksaman B. Use of large language models in Turkish information materials for glaucoma patient education: Evaluation of readability, accuracy, and comprehensiveness. Eur Eye Res 2026;6(1):60–69.

Correspondence: Ali Dal, M.D. Department of Ophthalmology, Mustafa Kemal University, Hatay, Turkiye
E-mail: alidal19@hotmail.com

Submitted Date: 24.07.2025 **Revised Date:** 28.09.2025 **Accepted Date:** 09.10.2025 **Available Online Date:** 29.04.2026

OPEN ACCESS This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

Glaucoma represents a leading cause of irreversible blindness on a global scale. Current estimates indicate that approximately 3.6 million cases of blindness among individuals aged 50 and older can be attributed to glaucoma. Projections suggest that by the year 2040, this condition is anticipated to impact over 110 million individuals worldwide.[1,2] Due to the asymptomatic onset and gradual progression of the disease, diagnosing and treating glaucoma present significant challenges, even in developed countries with established screening programs.[3]

In managing glaucoma, effectively controlling intraocular pressure through the appropriate use of medications can significantly reduce the risk of vision loss for most patients.[4] Non-adherence to treatment is a critical factor that significantly impacts the long-term visual prognosis of patients.[5] Research indicates that adherence to glaucoma treatment is notably low, particularly within the initial 6 months. During this timeframe, there was a significant decline in the number of patients who continued their treatment.[6] Reasons for non-adherence to treatment include patients' insufficient understanding of glaucoma, apprehensions regarding treatment, potential side effects of medications, and the belief that the medications may not be effective. Research indicates that these issues can be addressed through enhanced patient education and the implementation of effective communication strategies.[7,8]

To ensure patient compliance with treatment, it is essential that they receive accurate and comprehensive information about the treatment process. At the outset of treatment, key aspects such as medication dosages, administration timing, intervals between medications, and the technique for instilling drops should be thoroughly explained. Written informational materials provided to patients serve as valuable resources to facilitate the correct implementation of the treatment.[9] International guidelines suggest that patient education materials should be written at a readability level of at least sixth grade to ensure they are easily understood by patients and their families.[10] In addition to the educational brochures offered to patients, there is a growing trend of individuals seeking health information through digital platforms, including the internet, social media, and artificial intelligence (AI)-driven chatbots.[11] Research indicates that patients who seek information from sources such as the internet tend to exhibit greater adherence to their treatment plans.[12]

Large language models (LLMs) are advanced AI systems that have been trained on extensive datasets, enabling them to generate coherent and contextually relevant natural language text.[13] These systems possess the capability to generate medical information and educate patients by analyzing content sourced from the Internet. Models such as ChatGPT from OpenAI, Gemini from Google, and DeepSeek from DeepSeek AI are being increasingly utilized in the medical sector for patient education and the creation of informational content.[8,14] There are concerns about the accuracy, comprehensiveness, and readability of the content generated by LLMs.[15] In addition, Bard and Bing, which are also widely used AI tools in the healthcare field, are developed by Google and Microsoft, respectively.

In Turkey, the Turkish Ophthalmology Association (TOA) provides educational resources for patients regarding glaucoma and various other eye conditions on its official website (https://oftalmoloji.org.tr). In the context of chronic diseases such as glaucoma, it is essential that the information provided is clear, precise, and thorough, enabling patients to make well-informed decisions.[16] This study evaluated glaucoma patient information brochures created by TOA, focusing on their readability through the Ateşman and Bezirci-Yilmaz formulas. Questions presented in a question-answer format were posed to language models, and the responses generated by GPT-4.0, Gemini, and DeepSeek were analyzed for accuracy, comprehensiveness, and readability. In addition, the study assessed whether simplifying the AI-generated responses could enhance readability.

The primary objective of this study is to assess the readability level of TOA patient information brochures and to evaluate the effectiveness, readability, accuracy, and comprehensiveness of LLMs in generating Turkish patient education materials specifically for glaucoma patients.

Materials and Methods

In our research, we utilized the information provided on the glaucoma unit's webpage of the TOA website, which serves public informational purposes, as our primary data source. This guide is structured around common questions related to glaucoma, such as "What is glaucoma?" and "What causes glaucoma?" Each response in the guide was assessed individually using the Ateşman and Bezirci-Yilmaz readability formulas. Our study exclusively relied on publicly accessible data and literature. Given that no animal or human subjects were involved, approval from an ethics committee and patient consent were not necessary.

Use of Language Models

The questions from the glaucoma unit's public web page were submitted, without alteration, to ChatGPT-4.0, DeepSeek,

and Gemini, which are the most widely utilized LLMs.[17] Each question was posed on a separate chat page, and the responses were documented to develop new patient education brochures. We evaluated three LLMs: ChatGPT-4 (OpenAI, April 2025 version), Gemini Advanced 1.5 (Google, 2025 version), and DeepSeek-Chat (2025 version). All responses were generated between May 10 and 20, 2025, to ensure consistency. Standardized prompts were used for all models (see Supplementary File 1 for the complete prompt texts). Each model was run with default parameters (temperature= 0.7, maximum token limit =1024, randomness/seed = default).

Furthermore, these responses were requested to be reorganized into "Question 1" and "Question 2" formats to evaluate the capability of LLMs to adapt the texts for lower education levels.

- **Question 1: "***Can you rearrange the text I shared below so that a 6*th *grader can understand it?***"**
- **Question 2: "***Can you edit the text I shared below to make it simpler to understand?***"**

A total of 72 patient-style questions were extracted from the official TOA glaucoma brochure and used as the study dataset. For each question, responses were generated from three LLMs in three different formats: Initial response (IR), Question 1 format (Q1), and Question 2 format (Q2). This process yielded a total of 648 individual responses ($72 \times 3 \times 3$). Each response was analyzed separately using both the Ateşman and Bezirci-Yilmaz readability indices, and was independently evaluated for accuracy and comprehensiveness by two glaucoma specialists. Inter-rater agreement between the two glaucoma specialists was assessed using intraclass correlation coefficients (ICC, two-way random effects model, and absolute agreement).

Readability Criteria

Ateşman readability criterion

The Ateşman readability measure assigns a score ranging from 0 to 100 based on the average length of sentences and words. In this system, scores of 90–100 are deemed suitable for individuals at the 4th-grade level and below. Scores between 80 and 89 correspond to the 5th or 6th grade, whereas scores of 70–79 are appropriate for the 7th or 8th grade. Scores ranging from 60 to 69 indicate a 9th or 10th grade level of education, and scores of 50–59 are suitable for the 11th or 12th grade. Scores between 40 and 49 reflect an education level equivalent to an associate degree (13–15th grade), whereas scores of 30–39 are indicative of undergraduate graduation. Finally, scores of 29 and below are associated with graduate-level education.[17]

Bezirci-Yilmaz readability criterion

The Bezirci-Yilmaz readability criterion assesses text complexity by calculating a score based on the average sentence length and the number of syllables in the words used. Scores ranging from 1 to 8 are deemed suitable for primary education, whereas scores between 9 and 12 are appropriate for high school. For undergraduate education, scores from 12 to 16 are recommended, and scores exceeding 16 are considered suitable for academic levels.[18]

Comprehensiveness and accuracy of LLMs' production of patient-targeted information.

Responses from language models were assessed for both accuracy and comprehensiveness by referencing the information available on the public website of the glaucoma unit. The evaluation of the accuracy and breadth of the materials was conducted by specialist physicians A.D. and B.K., who possess extensive knowledge of glaucoma and actively manage glaucoma patients in their clinical practice. The evaluations were conducted independently by both reviewers and then discussed together to reach a consensus on each response.

For comprehensiveness, responses were rated as "not comprehensive" (1 point) for responses lacking important detail, "somewhat comprehensive" (2 points) for responses that included minimal but essential information, "moderately comprehensive" (3 points) for responses that provided a reasonable level of detail, "comprehensive" (4 points) for responses that addressed the most critical issues, and "very comprehensive" (5 points) for responses that provided comprehensive and detailed information.

Accuracy was scored as "poor" (1 point) for responses that contained significant inaccuracies that could mislead patients and potentially cause harm, "fair" (2 points) for responses that may contain factual errors but are unlikely to mislead or harm patients, and "good" (3 points) for responses that were free of errors.[20]

Statistical Analysis

In the analysis of the data, a one-way analysis of variance (ANOVA) was used to determine whether there were statistically significant differences between the groups. The normality of the data distribution was assessed using the Shapiro–Wilk test, and the homogeneity of variances was evaluated using Levene's test. When the ANOVA test indicated significant differences between group means, the Tukey honestly significant difference (HSD) *post hoc* test was applied to identify which specific groups differed. Effect sizes (η^2 for ANOVA and Cohen's d for

pairwise comparisons) and 95% confidence intervals (CIs) were reported alongside *P*-values to ensure transparent interpretation of the results. Statistical analyses were conducted using IBM Statistical Package for the Social Sciences Statistics for Windows, Version 26.0 (IBM Corp., Armonk, NY, USA). $p<0.05$ was considered statistically significant.

Results

Ateşman Readability Scores

In the analyses of the Ateşman readability score (as the Ateşman readability score increases, the level of education required to understand the text decreases), it was observed that the first answers produced by GPT-4.0, Gemini, and DeepSeek models had higher readability scores compared to the TOA brochure, but these differences were not found to be statistically significant (TOA and GPT-4.0: p=0.758, TOA and Gemini: p=0.101, TOA and DeepSeek: p=0.082). In addition, no statistically significant difference was found in the pairwise comparisons between GPT-4.0, Gemini, and DeepSeek models ($p>0.05$) (Table 1).

The evaluation of readability scores for Ateşman revealed that the IRs generated by GPT-4.0 and Gemini were assessed at a 9–10th grade level, whereas the responses from the DeepSeek model were rated at a 7–8th grade level. In the GPT-4.0 model, the simplification process (Q1 and Q2) did not produce a significant change in readability. On the other hand, a significant increase in readability scores was detected in Q1 and Q2 formats in Gemini and DeepSeek models. In the Gemini model, the readability score of the IRs was notably lower than that of the simplified formats (Q1 and Q2). Likewise, in the DeepSeek model, the readability scores following the simplification process demonstrated a significant improvement compared to the original responses (Table 2).

Bezirci-Yilmaz Readability Scores

The one-way ANOVA revealed a statistically significant difference in Bezirci-Yilmaz readability scores among the

Table 1. Ateşman and Bezirci-Yilmaz readability scores of the initial responses from TOA, GPT-4.0, Gemini, and DeepSeek models

Outcome	TOA (Mean±SD)	GPT-4.0 (Mean±SD)	Gemini (Mean±SD)	DeepSeek (Mean±SD)	Comparison	Mean difference	95% CI (Lower–upper)	*p*	Effect size (Cohen's d)
Ateşman readability score	58.61±12.41	62.85±8.84	68.91±5.78	72.11±3.38	TOA versus GPT-4.0	4.24	−3.2–11.6	0.758	0.18
					TOA versus Gemini	10.30	2.1–18.5	0.101	0.45
					TOA versus DeepSeek	13.50	4.3–21.7	0.082	0.51
					GPT-4.0 versus Gemini	6.06	−1.2–13.8	0.503	0.30
					GPT-4.0 versus DeepSeek	9.26	1.0–17.5	0.106	0.39
					Gemini versus DeepSeek	3.20	−2.6–9.0	0.711	0.15
Bezirci-Yilmaz Score	13.02±3.27	8.64±4.10	6.83±0.87	6.92±1.24	TOA versus GPT-4.0	−4.38	−9.1–0.3	0.077	0.42
					TOA versus Gemini	−6.19	−10.2–−2.1	**0.007**	0.71
					TOA versus DeepSeek	−6.10	−10.0–−2.2	**0.033**	0.68
					GPT-4.0 versus Gemini	−1.81	−4.3–0.7	0.724	0.15
					GPT-4.0 versus DeepSeek	−1.72	−4.5–0.9	0.098	0.14
					Gemini versus DeepSeek	0.09	−1.0–1.2	0.911	0.02

Statistically significant results ($p<0.05$) are indicated in bold. TOA: Turkish Ophthalmology Association's, CI: Confidence interval, SD: Standard deviation. Effect sizes are reported as Cohen's d. *Post hoc* comparisons adjusted with Bonferroni correction

Table 2. Comparison of Ateşman readability scores and education levels among the IR, Q1, and Q2 formats of GPT-4.0, Gemini, and DeepSeek

Model	Format	Ateşman score (Mean±SD)	Education level	Comparison	Mean difference	95% CI (Lower–upper)	*p*	Effect size (Cohen's d)
GPT-4.0	IR	62.85±8.83	9–10th Grade	IR versus Q1	0.53	−1.2–2.3	0.365	0.07
	Q1	63.38±7.90	9–10th Grade	IR versus Q2	0.32	−1.5–2.1	0.866	0.04
	Q2	63.18±8.84	9–10th Grade	Q1 versus Q2	−0.21	−2.1–1.7	0.912	0.02
Gemini	IR	68.91±5.78	9–10th Grade	IR versus Q1	4.19	1.1–7.2	**0.035**	0.61
	Q1	73.10±3.22	7–8th Grade	IR versus Q2	8.12	3.9–12.3	**0.004**	0.95
	Q2	77.03±2.73	7–8th Grade	Q1 versus Q2	3.93	1.8–6.0	**0.007**	0.72
DeepSeek	IR	72.11±3.38	7–8th Grade	IR versus Q1	3.53	0.9–6.2	**0.014**	0.58
	Q1	75.64±3.56	7–8th Grade	IR versus Q2	5.54	2.1–8.9	**0.004**	0.82
	Q2	77.65±6.23	7–8th Grade	Q1 versus Q2	2.01	0.5–3.7	**0.020**	0.44

Statistically significant results ($p<0.05$) are indicated in bold. IR: Initial response; CI: Confidence interval; SD: Standard deviation. Effect sizes are reported as Cohen's d. *Post hoc* comparisons adjusted with the Tukey honestly significant difference test

groups (p=0.007). Tukey HSD *post hoc* analysis showed that the IRs generated by Gemini and DeepSeek had significantly lower readability scores compared to the TOA brochure (p=0.007 and p=0.033, respectively). In contrast, the difference between GPT-4.0 and the TOA brochure was not statistically significant (p=0.077). Furthermore, no significant differences were found among the LLMs themselves, including GPT-4.0 versus Gemini (p=0.724), GPT-4.0 versus DeepSeek (P=0.098), and Gemini versus DeepSeek (p=0.911) (Table 1).

When comparing the IRs of the LLMs (GPT-4.0, Gemini, and DeepSeek) with the answers formatted in Q1 and Q2, a statistically significant enhancement in readability was observed exclusively for the Gemini model (IR and Q1: p=0.013, IR and Q2: p=0.005). In contrast, the GPT-4.0 and DeepSeek models did not exhibit a significant difference in readability scores following the simplification process (p>0.05). Furthermore, no significant differences were identified between the responses in Q1 and Q2 formats across any of the models (p>0.05) (Table 3).

Comprehensiveness and Accuracy Results

The evaluation of the accuracy and comprehensiveness scores of the responses generated by the LLMs revealed variability among the models, particularly in comprehensiveness. Specifically, when assessing accuracy, no statistically significant differences were identified between the GPT-4.0, Gemini, and DeepSeek models across all three formats (Table 4). Each question was posed on a separate chat page. Inter-rater reliability analysis demonstrated excellent agreement for accuracy ratings (ICC = 0.84, 95% CI: 0.79–0.88) and good agreement for comprehensiveness ratings (ICC=0.76, 95% CI: 0.69–0.82), confirming consistency between evaluators.

The analysis of comprehensiveness scores revealed that the Gemini model achieved the highest score in its IRs, demonstrating a statistically significant difference when compared to GPT-4.0 (p=0.042). However, no significant difference was found between the Gemini and DeepSeek models (p=0.103). In addition, the comparison between GPT-4.0 and DeepSeek models showed no significant difference in comprehensiveness (p=0.077).

Upon evaluating the Q1 and Q2 formats after simplification, the comprehensiveness scores for GPT-4.0 and DeepSeek remained largely unchanged. In contrast, a decline in the comprehensiveness score of the Gemini model was noted. A comparison of the responses from the LLM models in the Q1 format revealed a statistically significant difference between Gemini and both GPT-4.0 and DeepSeek, with P=0.025 for each comparison. However, no statistically significant difference was observed in the responses within the Q2 format (Tables 4 and 5).

In our study, we compared the responses of the Gemini, DeepSeek, and GPT-4.0 models across different formats (IR, Q1, and Q2) to assess their comprehensiveness and identify any statistically significant differences (Fig.1). The analysis of the Gemini model revealed a significant difference in comprehensiveness between the first response and the simplified Q1 format (p=0.011). However, no significant difference was observed between the Q1 and Q2 formats (p=0.465). For the DeepSeek model, the

Table 3. Comparison of Bezirci-Yilmaz readability scores and education levels among the IR, Q1, and Q2 formats of GPT-4.0, Gemini, and DeepSeek

Model	Format	Bezirci-Yilmaz Score (Mean±SD)	Education level	Comparison	Mean difference	95% CI (Lower–upper)	*p*	Effect size (Cohen's d)
GPT-4.0	IR	8.64±4.10	Primary school	IR versus Q1	0.14	−1.0–1.3	0.848	0.03
	Q1	8.78±3.27	Primary school	IR versus Q2	0.10	−0.9–1.2	0.888	0.02
	Q2	8.74±3.28	Primary school	Q1 versus Q2	−0.04	−1.0–0.9	0.825	0.01
Gemini	IR	6.83±0.87	Primary school	IR versus Q1	−1.01	−1.8–−0.3	**0.013**	0.78
	Q1	5.82±0.49	Primary school	IR versus Q2	−1.18	−2.0–−0.4	**0.005**	0.81
	Q2	5.65±0.58	Primary school	Q1 versus Q2	−0.17	−0.7–0.3	0.252	0.25
DeepSeek	IR	6.92±1.24	Primary school	IR versus Q1	−0.61	−1.6–0.4	0.222	0.28
	Q1	6.31±1.71	Primary school	IR versus Q2	−0.01	−1.2–1.1	0.261	0.02
	Q2	6.91±1.92	Primary school	Q1 versus Q2	0.60	−0.7–1.9	0.779	0.19

Statistically significant results ($P<0.05$) are indicated in bold. IR: Initial response; CI: Confidence interval; SD: Standard deviation. Effect sizes are reported as Cohen's d. *Post hoc* comparisons adjusted with the Tukey honestly significant difference test

Table 4. Accuracy scores of GPT-4.0, Gemini, and DeepSeek across IR, Q1, and Q2 formats

Format	GPT-4.0 (Mean±SD)	Gemini (Mean±SD)	DeepSeek (Mean±SD)	Comparison	Mean difference	95% CI (Lower–upper)	*p*	Effect size (Cohen's d)
IR	2.88±0.35	3.00±0.25	2.75±0.34	GPT-4.0 versus Gemini	−0.12	−0.4–0.1	0.317	0.35
				GPT-4.0 versus DeepSeek	0.13	−0.2–0.4	1.000	0.20
				Gemini versus DeepSeek	0.25	−0.1–0.6	0.317	0.38
Q1	2.51±0.23	2.63±0.34	2.61±0.40	GPT-4.0 versus Gemini	−0.12	−0.3–0.1	0.317	0.36
				GPT-4.0 versus DeepSeek	−0.10	−0.3–0.2	1.000	0.28
				Gemini versus DeepSeek	0.02	−0.3–0.3	0.317	0.04
Q2	2.39±0.37	2.72±0.40	2.42±0.36	GPT-4.0 versus Gemini	−0.33	−0.6–−0.1	0.317	0.55
				GPT-4.0 versus DeepSeek	−0.03	−0.3–0.2	1.000	0.08
				Gemini versus DeepSeek	0.30	0.0–0.6	0.317	0.50

Statistically significant results ($p<0.05$) are indicated in bold. IR: Initial response; CI: Confidence interval; SD: Standard deviation. Effect sizes are reported as Cohen's d. *Post hoc* comparisons adjusted with the Tukey honestly significant difference test

Table 5. Comprehensiveness scores of GPT-4.0, Gemini, and DeepSeek across IR, Q1, and Q2 formats

Format	GPT-4.0 (Mean±SD)	Gemini (Mean±SD)	DeepSeek (Mean±SD)	Comparison	Mean difference	95% CI (Lower–upper)	*p*	Effect size (Cohen's d)
IR	3.25±0.71	3.88±0.35	3.50±0.84	GPT-4.0 versus Gemini	−0.63	−1.2–0.1	**0.042**	0.72
				GPT-4.0 versus DeepSeek	−0.25	−0.8–0.3	0.077	0.30
				Gemini versus DeepSeek	0.38	−0.1–0.9	0.103	0.46
Q1	2.00±0.51	2.50±0.53	2.00±0.42	GPT-4.0 versus Gemini	−0.50	−0.9–0.1	**0.025**	0.82
				GPT-4.0 versus DeepSeek	0.00	−0.4–0.4	1.000	0.00
				Gemini versus DeepSeek	0.50	0.1–0.9	0.025	0.81
Q2	2.00±0.48	2.13±0.64	2.00±0.56	GPT-4.0 versus Gemini	−0.13	−0.5–0.3	0.537	0.22
				GPT-4.0 versus DeepSeek	0.00	−0.4–0.4	1.000	0.01
				Gemini versus DeepSeek	0.13	−0.3–0.6	0.537	0.20

Statistically significant results ($p<0.05$) are indicated in bold. IR: Initial response; CI: Confidence interval; SD: Standard deviation. Effect sizes are reported as Cohen's d. *Post hoc* comparisons adjusted with the Tukey honestly significant difference test

comparisons indicated no statistically significant difference in comprehensiveness between the IR and the Q1 format (p=0.076). Similarly, the analysis of the GPT-4 model showed no significant difference between the IR and the Q1 format (p=0.092). Furthermore, when examining accuracy rates, no significant differences were found between the IR and Q1

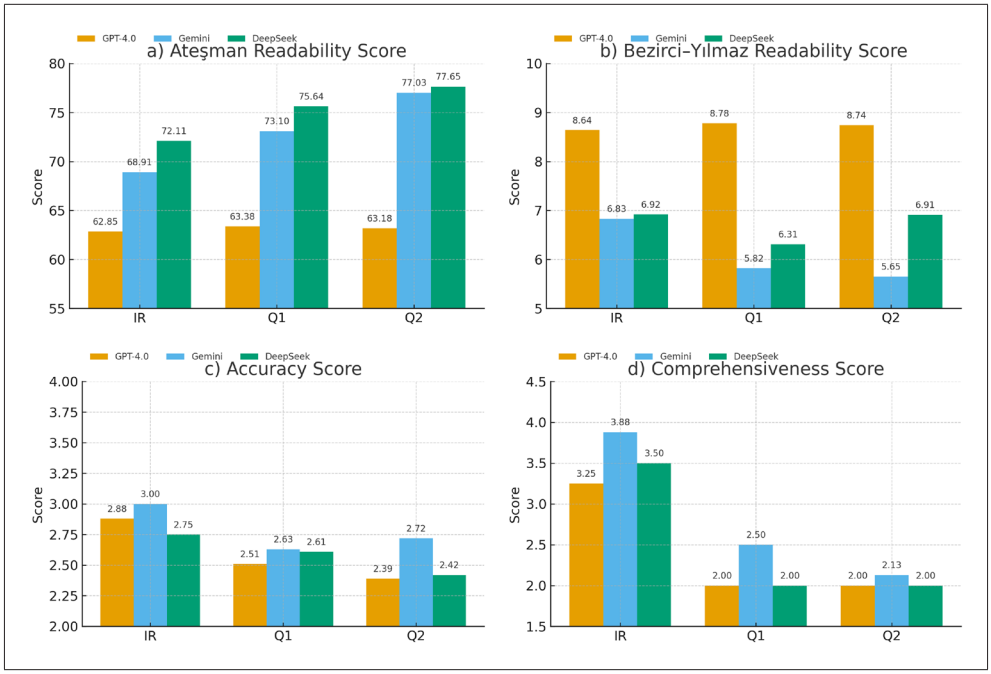

Fig. 1. Summary of main outcomes across GPT-4.0, Gemini, and DeepSeek in three response formats (IR, Q1, and Q2). **(a)** Ateşman readability scores, **(b)** Bezirci-Yilmaz readability scores, **(c)** Accuracy scores, and **(d)** Comprehensiveness scores. IR: Initial response; Q1: First simplified version, Q2: Second simplified version.

formats across all three LLM models (GPT-4.0, Gemini, and DeepSeek) (with p=0.141, 0.097, and 0.102, respectively).

Discussion

This study evaluated the readability of the glaucoma patient information brochure created by TOA, as well as the effectiveness and reliability of LLMs in delivering Turkish information to glaucoma patients. The findings allow for a direct comparison between traditional brochures prepared for Turkish-speaking patients and the content generated by LLMs. The results suggest that AI-based models have the potential to produce more accessible materials that may enhance patient understanding and adherence.

Effective patient information brochures should be easily understood by individuals with a low level of education while providing comprehensive and accurate information.[21] The American Medical Association advises that patient education materials should be written at a readability level of 6th grade or lower.[8] A recent study indicated that the patient information materials from the American Academy of Ophthalmology were developed at an 8th-grade reading level, which exceeds the recommended patient education level of 6th grade.[16] In our study, we found that the readability level of the brochures prepared by TOA was assessed to be at the 11–12th grade according to the Ateşman formula and at the undergraduate level according to the Bezirci-Yilmaz formula, both exceeding the recommended level. These results suggest that a higher level of education is necessary for a better understanding of patient information materials in ophthalmology. This situation should be considered to enhance clarity in patient education. However, when readability is increased, there is a risk of reduced comprehensiveness. In our study, readability was primarily enhanced by reducing sentence length and simplifying word choices, which allowed the content to become more accessible without altering its scope. Nevertheless, to ensure that simplified brochures maintain their comprehensiveness, additional strategies such as supplementary information layers and visual aids may be incorporated to preserve the depth and accuracy of patient education materials.

While AI-based language models can enhance patient education by increasing knowledge, it is essential to thoroughly assess the accuracy, comprehensiveness, and readability of the content they generate.[14] The results of our study showed that the responses produced by GPT-4.0 had a similar level of readability as the TOA brochure, but the responses produced by Gemini and DeepSeek required a lower level of training to be meaningfully understood compared to the Bezirci-Yilmaz formula. This finding shows that some LLMs can produce more understandable Turkish content. Previous studies have suggested that AI-supported approaches to reduce the level of education required to understand patient information materials may be beneficial for patient education.[20] Our research substantiates this perspective.

In the research conducted by Yalla *et al*.,[16] it was noted that ChatGPT-4.0 demonstrated a higher accuracy level compared to other AI models such as Bing and Bard. However, our study revealed no significant differences in accuracy among the ChatGPT-4.0, Gemini, and DeepSeek models. Regarding comprehensiveness, while the previous study[22] indicated that ChatGPT provided the most thorough answers, our findings showed that the Gemini model achieved the highest comprehensiveness score, with GPT-4.0 and DeepSeek offering lower levels of comprehensiveness. Although the accuracy scores were similar among the models, differences in understandability and comprehensiveness may reflect variations in training datasets, linguistic adaptation, and the ability of each model to process Turkish-specific language structures. This discrepancy highlights the challenge of objectively comparing models and underscores the need for standardized benchmarks and larger expert panels in future studies to ensure consistent and unbiased evaluation across multiple dimensions. Prior research has explored the capacity of LLMs to customize patient education materials to align with health literacy, thereby enhancing the accessibility of health information. These studies have demonstrated that LLMs can effectively adjust content to accommodate lower reading levels.[23] Our research aligns with these findings, demonstrating that structuring patient education materials based on individual health literacy levels can enhance the understanding of the disease and promote informed health decisions.[22,24]

The utilization of AI -driven models for patient information materials presents certain risks. While the content generated by these models appears realistic and fluent, it relies solely on statistical word associations and lacks a genuine reasoning process.[25] The reliability and contextual coherence of the information generated by LLMs can be called into question. It is important to note that the evaluations in this study were performed in May 2025. Given the rapid evolution of AI models, future updates may alter the performance and content generation capabilities of these systems. Models such as ChatGPT from OpenAI are notable for their propensity to generate false or fabricated content, often referred to as "hallucinations" or "confabulations," which significantly restricts their applicability in patient education.[26] In our

study, no hallucinations or major factual inaccuracies were observed in the responses generated by the LLMs. Patients are progressively utilizing the internet and AI-driven models to obtain health-related information.[27] Research indicates that the accuracy of information retrieved from Google searches related to glaucoma is notably low, with considerable discrepancies in the reliability of such information. In this context, the utilization of LLMs as patient information resources may offer a distinct advantage compared to conventional online information sources; however, it is imperative that they are meticulously assessed for accuracy and reliability. Our research indicates that LLMs deliver precise and comprehensive information for educating glaucoma patients in Turkish.

When interpreting the results of our study, it is essential to acknowledge several limitations. First, the questions assessed were derived from glaucoma patient information brochures created by TOA, which may not fully encompass the inquiries or needs most frequently expressed by patients. Nonetheless, the thoroughness of LLMs' responses was evaluated against the TOA brochures, based on the assumption that these materials offer the most accurate and comprehensive information available. The study was conducted exclusively in Turkish, which may limit its applicability to other languages and cultural contexts. In addition, reliance on TOA materials alone may not capture the full spectrum of ophthalmology patient resources, and the generalizability of findings across different populations and healthcare settings remains limited. Furthermore, assessments of accuracy and comprehensiveness were conducted utilizing subjective rating scales. Nevertheless, this limitation is mitigated as comprehensiveness is directly evaluated against TOA brochures, whereas accuracy scores are derived from numerical values rather than a binary "true or false" system. Major language models undergo continuous updates, with new information being integrated periodically. The responses generated by the LLMs in our study were derived from a specific timeframe, indicating that the results may evolve over time. Therefore, longitudinal analyses and repeated evaluations are necessary to ensure consistency of model performance over time. In addition, the readability formulas used in this study (Bezirci-Yilmaz and Ateşman) only measure sentence and word complexity and do not assess the factual accuracy of the content. The good-to-excellent inter-rater agreement supports the robustness of our evaluation process, although future studies with larger expert panels could provide further validation. Furthermore, the default configurations of the language models were employed; variations in prompts or customized settings could potentially alter the outcomes. Future research should consider conducting longitudinal analyses to assess the consistency of these models over time and to evaluate how their accuracy and comprehensiveness are affected by information updates.

Conclusion

In summary, this study demonstrates that LLMs may serve as valuable tools in patient information processes. The comparison between the TOA brochure and the content generated by the LLMs in terms of readability, accuracy, and comprehensiveness provides preliminary insights into the possible usability of AI-based systems as patient education materials, while highlighting the need for further validation in broader contexts. The advancement of AI-supported patient education models holds the potential to transform patient education and information processes. Using AI-generated materials without expert oversight carries risks, including potential hallucinations, incomplete explanations, or oversimplification of complex medical information. Therefore, LLMs should be regarded as supportive tools rather than standalone resources, and all outputs must be reviewed and validated by specialists before being shared with patients. However, further comprehensive studies are necessary to ensure accuracy, reliability, and patient safety before these models can be effectively integrated into clinical practice.

Ethics Committee Approval: Ethics committee approval was not required.

Informed Consent: Written informed consent was obtained.

Conflict of Interest: None declared.

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

Use of AI for Writing Assistance: None declared.

Authorship Contributions: Concept: A.D.; Design: A.D.; Supervision: A.D., B.K.; Resource: A.D., B.D.; Data collection and/or processing: A.D., B.D.; Analysis and/or interpretation: A.D., B.K.; Writing: A.D., B.D.; Critical review: M.E., B.K.

Peer-review: Externally peer-reviewed.

References

1. Bourne RRA, Steinmetz JD, Saylan M. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight. An analysis for the Global Burden of Disease Study. Lancet Glob Health 2021;9:e144–60.
2. Barkana Y, Dorairaj S. Re: Tham: Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systemat-

ic review and meta-analysis. Ophthalmology 2015;122:e40–1. [CrossRef]
3. Mitchell P, Smith W, Attebo K, Healey PR. Prevalence of open-angle glaucoma in Australia. The Blue Mountains Eye Study. Ophthalmology 1996;103:1661-9. [CrossRef]
4. Sleath B, Blalock S, Covert D, Stone JL, Skinner AC, Muir K, et al. The relationship between glaucoma medication adherence, eye drop technique, and visual field defect severity. Ophthalmology 2011;118:2398-402. [CrossRef]
5. Fu DJ, Ademisoye E, Shih V, McNaught AI, Khawaja A. Survival of medical treatment success in primary open-angle glaucoma and ocular hypertension. Br J Ophthalmol 2024;108:1701-7. [CrossRef]
6. Tapply I, Broadway DC. Improving adherence to topical medication in patients with glaucoma. Patient Prefer Adherence 2021;15:1477–89. [CrossRef]
7. McDonald S, Ferguson E, Hagger MS, Foss AJE, King AJ. A theory-driven qualitative study exploring issues relating to adherence to topical glaucoma medications. Patient Prefer Adherence 2019;13:819-28. [CrossRef]
8. Lee GG, Goodman D, Chang TCP. Impact of demographic modifiers on readability of myopia education materials generated by large language models. Clin Ophthalmol 2024;18:3591-604. [CrossRef]
9. Kharod BV, Johnson PB, Nesti HA, Rhee DJ. Effect of written instructions on accuracy of self-reporting medication regimen in glaucoma patients. J Glaucoma 2006;15:244-7. [CrossRef]
10. Weiss BD. Help patients understand: manual for clinicians. 2nd ed. Chicago: American Medical Association; 2007.
11. Yang Z, Wang D, Zhou F, Song D, Zhang Y, Jiang J, et al. Understanding natural language: Potential application of large language models to ophthalmology. Asia Pac J Ophthalmol (Phila) 2024;13:100085. [CrossRef]
12. Haynes RB, McDonald HP, Garg AX. Helping patients follow prescribed treatment: clinical applications. JAMA 2002;288:2880-3. [CrossRef]
13. Cohen SA, Brant A, Fisher AC, Pershing S, Do D, Pan C. Dr. Google vs. Dr. ChatGPT: exploring the use of artificial intelligence in ophthalmology by comparing the accuracy, safety, and readability of responses to frequently asked patient questions regarding cataracts and cataract surgery. Semin Ophthalmol 2024;39:472-9. [CrossRef]
14. Srinivasan N, Samaan JS, Rajeev ND, Kanu M U, Yeo YH, et al. Large language models and bariatric surgery patient education: a comparative readability analysis of GPT-3.5, GPT-4, Bard, and online institutional resources. Surg Endosc 2024;38:2522-32. [CrossRef]
15. Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Sam S, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open 2023;6:e2336483. [CrossRef]
16. Yalla GR, Hyman N, Hock LE, Zhang Q, Shukla AG, Kolomeyer NN. Performance of artificial intelligence chatbots on glaucoma questions adapted from patient Brochures. Cureus 2024;16:e56766. [CrossRef]
17. Ateşman E. Türkçede okunabilirliğin ölçülmesi. Dil Dergisi 1997;58:71-4. [Article in Turkish]
18. Bezirci B, Yılmaz AE. A software library for measurement of readability of texts and a new readability metric for Turkish. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi 2010;12:17-25.
19. Kianian R, Sun D, Crowell EL, Tsui E. The use of large language models to generate education materials about uveitis. Ophthalmol Retina 2024;8:195-201. [CrossRef]
20. Postacı SA, Dal A. The ability of large language models to generate patient information materials for retinopathy of prematurity: evaluation of readability, accuracy, and comprehensiveness. Turk J Ophthalmol. 2024;54:330-6. [CrossRef]
21. Cohen SA, Fisher AC, Xu BY, Song BJ. Comparing the accuracy and readability of glaucoma-related question responses and educational materials by Google and ChatGPT. J Curr Glaucoma Pract 2024;18:110-6. [CrossRef]
22. Killeen OJ, Niziol LM, Cho J, Heisler M, Resnicow K, Darnley-Fisch D, et al. Glaucoma medication adherence 1 year after the support, educate, empower personalized glaucoma coaching program. Ophthalmol Glaucoma 2023;6:23-8. [CrossRef]
23. Newman-Casey PA, Niziol LM, Lee PP, Musch DC, Resnicow K, Heisler M. The impact of the Support, Educate, Empower personalized glaucoma coaching pilot study on glaucoma medication adherence. Ophthalmol Glaucoma 2020;3:228–37. [CrossRef]
24. Vaishya R, Misra A, Vaish A. ChatGPT: Is this version good for healthcare and research? Diabetes Metab Syndr 2023;17:102744. [CrossRef]
25. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, Shue A, Chou JC, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open 2023;6:e2330320. [CrossRef]
26. Quaranta L, Novella A, Tettamanti M, Pasina L, Weinreb RN, Nobili A. Adherence and persistence to medical therapy in glaucoma: an overview. Ophthalmol Ther 2023;12:2227–40. [CrossRef]
27. Wang J, Shi R, Le Q, Shan K, Chen Z, Zhou X, et al. Evaluating the effectiveness of large language models in patient education for conjunctivitis. Br J Ophthalmol 2025;109:185–91. [CrossRef]