

Evaluating Large Language Models' Responses to Common Queries on Retinal Surgery: Readability and Accuracy Analysis

Erkan Duman¹, Aydın Maçın², Ayşe Ergin³

¹Clinic of Ophthalmology, West Eye Hospital, Erbil, Iraq

²Clinic of Ophthalmology, West Hospital, Diyarbakır, Türkiye

³Clinic of Ophthalmology, University of Health Sciences Türkiye, Fatih Sultan Mehmet Training and Research Hospital, İstanbul, Türkiye

Abstract

BACKGROUND/AIMS: Large language models (LLMs) such as ChatGPT-4o mini, ChatGPT-4o, Gemini 1.5 Pro and Microsoft Copilot play an important role in improving health literacy and supporting patient decision making. This study aims to evaluate the clinical appropriateness and readability of responses provided by these LLMs to frequently asked questions about retinal detachment (RD), epiretinal membrane (ERM), and macular hole (MH).

MATERIALS AND METHODS: This comparative cross-sectional study did not involve human participants. A total of 121 questions related to RD, ERM and MH were compiled. These questions were presented to LLM interfaces and were categorised into different themes. Inter-rater agreement between two retinal specialists who scored the responses based on their clinical expertise was assessed using the intraclass correlation coefficient (ICC). In addition, the readability of the responses was analysed using five readability indices.

RESULTS: Response appropriateness rates were similar for ChatGPT-4o mini (100%), ChatGPT-4o (99.2%) and Gemini 1.5 Pro (99.2%), while Microsoft Copilot (90.9%) showed comparatively lower performance. A high level of inter-rater agreement was observed among clinicians. Regarding readability, ChatGPT-4o performed best, whereas ChatGPT-4o mini performed worst. Gemini 1.5 Pro and Microsoft Copilot were intermediate between these two models. The excellent inter-rater agreement (ICC: 0.891-0.938) further supports the reliability of the expert evaluation process.

CONCLUSION: Our study showed that the evaluated LLMs provided largely appropriate and comprehensive answers to questions related to vitreoretinal surgery. ChatGPT-4o showed the highest performance in both appropriateness and readability, while ChatGPT-4o mini, despite its high appropriateness, had the lowest readability score.

Keywords: LLMs, large language models, Gemini, ChatGPT, Microsoft Copilot

To cite this article: Duman E, Maçın A, Ergin A. Evaluating large language models' responses to common queries on retinal surgery: readability and accuracy analysis. Cyprus J Med Sci. 2026;11(5):369-377

ORCID IDs of the authors: E.D. 0000-0003-0739-536X; A.M. 0009-0003-8210-294X; A.E. 0000-0003-3593-0710.



Corresponding author: Erkan Duman

E-mail: drerkanduman@gmail.com

ORCID ID: orcid.org/0000-0003-0739-536X

Received: 25.03.2026

Accepted: 14.07.2026

Publication Date: 16.10.2026



Copyright© 2026 The Author(s). Published by Galenos Publishing House on behalf of Cyprus Turkish Medical Association.

This is an open access article under the Creative Commons AttributionNonCommercial 4.0 International (CC BY-NC 4.0) License.

INTRODUCTION

Natural language processing (NLP) techniques, particularly chatbots powered by large language models (LLMs), offer significant potential in healthcare by providing insights into disease prevention, diagnosis, and treatment.^{1,2} Trained on vast textual datasets, these models generate human-like responses through self-supervised learning.^{3,4}

LLMs process medical data contextually and perform semantic analysis, improving both information retrieval and patient communication.³ Unlike traditional rule-based chatbots with fixed keyword-based responses, LLMs generate more detailed and meaningful answers by analyzing queries in context.⁴ Their ability to retain prior interactions enhances response consistency, thereby benefiting patient education. Additionally, LLMs improve medical information accessibility and clarity, contributing to effective health communication. These models are now widely applied across healthcare settings.

This study evaluates the capabilities, applications, and readability levels of four prominent artificial intelligence (AI) models: OpenAI's ChatGPT-4o mini and ChatGPT-4o, Gemini 1.5 Pro, and Microsoft Copilot. ChatGPT-4o mini, a lightweight and cost-effective model, is optimized for basic NLP tasks, while ChatGPT-4o is an advanced multimodal system that processes text, images, audio, and video.^{5,6} Both models set a benchmark for LLM performance and enhance user interaction.

Google's Gemini project, the company's most advanced AI model, builds on the Bard AI programme's core technologies. Gemini 1.5 Pro aims to enhance AI capabilities and user interaction, making it the latest version chosen for this study.⁷ Microsoft's Copilot integrates NLP, retrieval-augmented generation (RAG), and LLMs to generate accurate responses. By retrieving diverse data, it improves clinical appropriateness and user experience, while its RAG framework enhances dynamic information processing in search-based models.^{8,9}

LLMs are trained on extensive datasets that incorporate content with varying levels of reliability. The presence of inconsistent online information gives rise to concerns regarding the reliability of generated responses, which may contain inaccuracies or omissions. This emphasises the necessity for rigorous evaluation and verification of medical information, particularly in healthcare settings where chatbot-generated content directly impacts patient safety.^{10,11} The challenge stems from LLMs' limited capacity to assess the reliability and clinical appropriateness of their training data.^{3,12} While these models continue to evolve, further in-depth assessment is required to determine their effectiveness in specific medical domains.

Vitreoretinal diseases are serious conditions that can lead to vision loss and often necessitate surgical intervention. It is imperative that patients have access to accurate, contemporary, and comprehensible information. However, ophthalmic surgical explanations frequently involve complex medical terminology, which can impede readability and comprehension, particularly for patients with limited medical knowledge. While LLMs enhance access to medical information, their clinical appropriateness and clarity are critical in facilitating effective patient communication.¹³ Therefore, this study evaluates both the clinical appropriateness and readability of LLM-generated responses to determine their accessibility for patient education.

Previous studies evaluating ophthalmology chatbots have primarily focused on individual retinal diseases, theoretical knowledge

assessments, or the clinical appropriateness of a limited number of LLMs. Moreover, readability and comprehensiveness have not been evaluated across multiple contemporary LLMs for common vitreoretinal surgical conditions. Therefore, this study aimed to compare the clinical appropriateness, comprehensiveness, and readability of responses generated by four widely used LLMs for frequently asked questions regarding retinal detachment (RD), epiretinal membrane (ERM), and macular hole (MH), thereby providing a broader assessment of their potential role in patient education. We hypothesized that although contemporary LLMs would generally provide clinically appropriate responses, there would be significant differences in their comprehensiveness and readability.

MATERIALS AND METHODS

Study Design

This study was conducted on February 5, 2025, using ChatGPT-4o mini, ChatGPT-4o (OpenAI, San Francisco, CA, USA), Gemini 1.5 Pro (Google LLC, Mountain View, CA, USA), and Microsoft Copilot (Microsoft Corporation, Redmond, WA, USA). These LLMs were selected based on their performance and accessibility.

Ethics Approval and Informed Consent

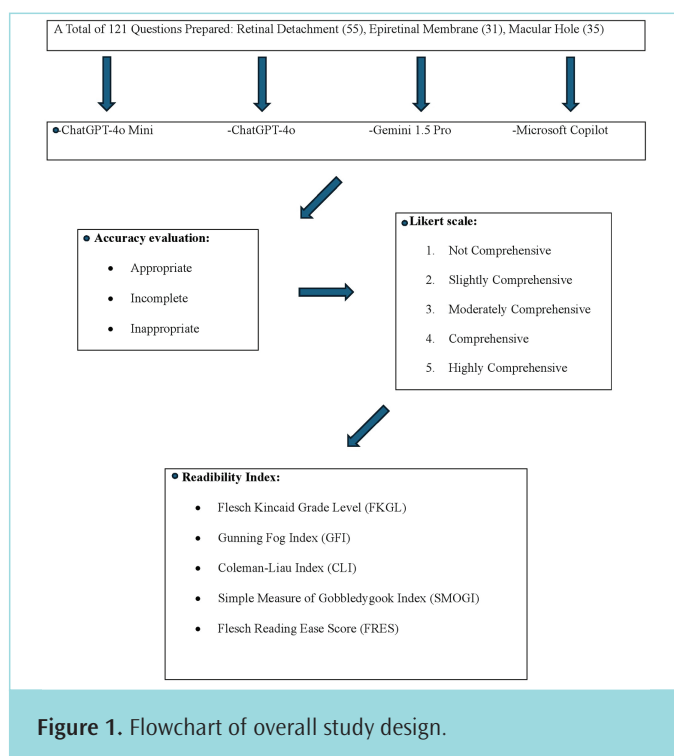
This study did not involve human participants or identifiable human data. The Institutional Review Board of our hospital determined that the study met the criteria for exemption from ethical review and that it was conducted in accordance with the tenets of the Declaration of Helsinki. Therefore, informed consent was not required.

ChatGPT-4o mini is optimized for efficiency, requiring less computational power and offering faster responses, while ChatGPT-4o is designed for high-precision tasks requiring more advanced processing. Both models were included to assess their respective strengths. Similarly, Gemini 1.5 Pro, an advanced multimodal model by Google, and Microsoft Copilot, an AI-powered assistant integrated with search-engine capabilities, were evaluated for their clinical appropriateness and readability in ophthalmic contexts.

Patient questions were sourced from reputable ophthalmology databases (e.g., the American Academy of Ophthalmology: <https://www.aao.org/eye-health>). Frequently asked questions were identified from online patient forums and clinical encounters. Each question was submitted to the LLMs three times to assess response variability. The three generated responses were assigned sequential numbers (1-3), and one response was selected using a computer-generated random number before expert evaluation to minimize selection bias. Only one randomly selected response was included in the analysis to simulate a typical real-world patient interaction and to avoid treating multiple responses to the same question as independent observations, which could have artificially inflated the sample size and influenced the comparative analyses. The questions were categorized into six domains: pathogenesis, risk factors, clinical presentation, diagnosis, treatment/prevention, and prognosis. To minimize bias, conversation history was cleared after each response, and chatbot-specific features were removed for a blinded assessment. Figure 1 illustrates the overall study design.

Clinical Appropriateness Assessment

All recorded responses were independently reviewed by two retina specialists, one attending physician (S.E.) and one associate professor



(M.K.), based on their clinical expertise. Each response was categorized as “appropriate,” “inappropriate,” or “incomplete”.

A response was classified as appropriate if it closely resembled the standard medical guidance that a clinician would provide and contained accurate information.

Responses containing incorrect or misleading information or significantly deviating from accepted clinical recommendations were considered inappropriate.

Incomplete: a response that was relevant and correct but lacked sufficient detail.

The final clinical appropriateness rating for each LLM-generated response was determined by consensus between the two evaluators. If consensus could not be reached after discussion, the corresponding question and response would have been excluded from the analysis according to the predefined study protocol. However, consensus was achieved for all evaluated responses; therefore, no responses were excluded.

Comprehensiveness Assessment

Responses that were rated as “appropriate” underwent further evaluation to assess their level of detail and completeness using a five-point Likert scale:¹⁴

Not comprehensive (1): The response failed to address the key aspects of the question or contained incorrect or misleading information, rendering it of no value to patients or clinicians.

Slightly comprehensive (2): The response provided only basic definitions with minimal detail and lacked sufficient explanation for patient education.

Moderately comprehensive (3): While covering fundamental concepts, the response remained superficial, omitting essential diagnostic or treatment details.

Comprehensive (4): The response addressed the most critical elements but lacked additional insights or alternative perspectives, making it generally useful but requiring further clarification for clinical decision-making.

Highly comprehensive (5): The response provided a detailed, scientifically accurate, and well-structured explanation, covering diagnosis, treatment, complications, and alternative approaches, making it optimal for patient education.

All responses were independently assessed by two retina specialists who were blinded to minimize bias. Inter-rater reliability was measured using the intraclass correlation coefficient (ICC) to assess consistency between evaluators. High ICC values indicated strong agreement, ensuring the reliability of the assessment process.

Readability Assessment

The readability of LLM-generated responses was assessed using the online readability tool Readable (<https://app.readable.com/text/>) to determine their accessibility for an average reader. Five well-established readability indices were used:^{15,16}

Flesch-Kincaid grade level: This metric estimates the United States (US) grade level required to understand a text by analyzing sentence length and syllables per word. Lower Flesch-Kincaid grade level (FKGL) scores indicate greater readability, suggesting that the content is easier to understand. This index is commonly used in medical education and patient information materials to assess whether a text aligns with the general population’s literacy level.¹⁵

The gunning fog index evaluates readability based on sentence complexity and the proportion of polysyllabic words with three or more syllables. It is particularly relevant for medical texts, as dense terminology can significantly impact comprehension. A gunning fog index (GFI) score above 12 indicates that the text requires university-level reading proficiency.^{15,17} In this study, GFI served as a key indicator of the ease with which LLM-generated content could be understood by general readers and patients.

Coleman-Liau index: Unlike traditional readability indices, Coleman-Liau index (CLI) assesses text complexity based on character count per word rather than word count per sentence. This makes it especially useful for medical documents, in which longer medical terms can distort traditional readability metrics. CLI provides a more precise measure of cognitive load and health literacy, offering insights into whether AI-generated medical texts align with the general public’s comprehension abilities.¹⁵

The simple measure of Gobbledygook index (SMOGI) is widely applied in public health and patient education; it evaluates readability by measuring the frequency of polysyllabic words. A score above 8 suggests that the text is difficult to read, requiring higher education levels for full comprehension.^{15,17} In healthcare communication, SMOGI is essential for assessing the readability of medical responses and their accessibility to individuals with lower health literacy.

Flesch reading ease score: Assesses readability on a 0-100 scale, with higher scores indicating greater ease of reading. Scores between 70 and 80 correspond to an 8th-grade reading level, making the text suitable for general patient education. A low flesch reading ease score (FRES) score suggests high text complexity, which may hinder effective communication in healthcare settings.¹⁷

Lower FKGL, GFI, CLI, and SMOGI scores indicate higher readability, making the text easier to comprehend. Conversely, a higher FRES score reflects improved accessibility. These indices correlate with the minimum education level required to understand the text in the US education system, providing a standardized measure for evaluating patient-directed medical information.¹⁸

Statistical Analysis

In this study, continuous variables were reported as mean $\pm$ standard deviation and minimum-maximum values, while descriptive statistics for categorical variables were presented as counts and percentages (%). The normality of the distributions of continuous variables was assessed using the Kolmogorov-Smirnov test. To determine statistically significant differences between readability metrics and LLM performance, a one-way analysis of variance was applied, followed by Tukey's honest significant difference post-hoc tests to identify pairwise differences. The results were analyzed to determine the best- and worst-performing LLMs for each disease group and readability metric. Additionally, Pearson correlation analysis was conducted to evaluate the relationships between LLM-generated responses and readability indices, with results presented in Supplementary Table 1. Statistical analyses were performed using the SPSS, version 28.0 (IBM Inc., Armonk, NY, USA). A p-value <0.05 was considered statistically significant, and 95% confidence intervals were used.

RESULTS

Response Appropriateness and Consistency

A total of 121 questions were asked of four LLMs. For RD, 55 questions were categorized into pathogenesis (n=10), risk factors (n=10), clinical findings (n=10), diagnosis (n=10), treatment/prevention (n=10), and prognosis (n=5). For ERM, 31 questions were categorized into pathogenesis (n=7), risk factors (n=3), clinical findings (n=4), diagnosis (n=3), treatment/prevention (n=8), and prognosis (n=6). For MH, 35

questions were categorized into pathogenesis (n=8), risk factors (n=7), clinical findings (n=6), diagnosis (n=5), treatment/prevention (n=5), and prognosis (n=4) (Supplementary Table 1).

Independent evaluations showed similar levels of response appropriateness for ChatGPT-4o mini (121/121, 100%), ChatGPT-4o (120/121, 99.2%), and Gemini 1.5 Pro (120/121, 99.2%), whereas Microsoft Copilot achieved 110/121 (90.9%) appropriate responses, with 10 incomplete and 1 inappropriate response (Table 1). Figure 2 illustrates the distribution of questions and response appropriateness ratings. Although other LLMs showed consistency across disease groups, Microsoft Copilot had lower reliability, particularly for treatment- and prognosis-related answers. Some diagnostic responses also lacked detail and were therefore clinically incomplete. ChatGPT-4o mini, despite its high response appropriateness, sometimes produced responses that were highly technical and difficult to read. In particular, Microsoft Copilot produced a treatment response that was scientifically questionable, further highlighting limitations in the appropriateness of its responses.

Response Comprehensiveness and Inter-Rater Reliability

The Likert scale evaluations of LLM responses are summarized in Table 2. ChatGPT-4o scored highest for RD, Gemini 1.5 Pro scored highest for ERM, and both models performed best for MH. However, no statistically significant differences in Likert scores were observed among LLMs in any disease group (RD: $p=0.063$, ERM: $p=0.532$, MH: $p=0.086$). Inter-rater agreement, assessed using ICC, demonstrated high consistency across evaluations (Table 3). This strong agreement confirms the reliability of the assessment and supports the objectivity of the evaluation process.

Readability Performance

The readability scores of LLM-generated responses are summarized in Table 2. ChatGPT-4o exhibited the highest readability across all disease groups, while ChatGPT-4o mini exhibited the lowest readability. ChatGPT-4o mini scored higher on the FKGL, GFI, CLI, and SMOG indices and lower on the FRES, indicating a more complex and technical language style ($p<0.001$). Conversely, ChatGPT-4o provided the most accessible responses, making it the most suitable model for patient education.

Table 1. Appropriateness of responses regarding vitreoretinal disease

Retinal detachment					Macular hole				
Questions	ChatGPT-4o mini (OpenAI)	ChatGPT-4o	Gemini 1.5 pro	Microsoft copilot	Questions	ChatGPT-4o mini (OpenAI)	ChatGPT-4o	Gemini 1.5 pro	Microsoft copilot
RD-23	Appropriate	Appropriate	Appropriate	Incomplete	MH-4	Appropriate	Incomplete	Appropriate	Appropriate
RD-27	Appropriate	Appropriate	Appropriate	Incomplete	MH-15	Appropriate	Appropriate	Appropriate	Incomplete
RD-30	Appropriate	Appropriate	Appropriate	Incomplete	MH-24	Appropriate	Appropriate	Appropriate	Incomplete
RD-31	Appropriate	Appropriate	Appropriate	Incomplete	Epiretinal membrane				
RD-38	Appropriate	Appropriate	Appropriate	Incomplete	Questions	ChatGPT-4o mini (OpenAI)	ChatGPT-4o	Gemini 1.5 pro	Microsoft copilot
RD-49	Appropriate	Appropriate	Appropriate	Inappropriate	ERM-20	Appropriate	Appropriate	Appropriate	Incomplete
RD-54	Appropriate	Appropriate	Incomplete	Appropriate	ERM-24	Appropriate	Appropriate	Appropriate	Incomplete
					ERM-25	Appropriate	Appropriate	Appropriate	Incomplete

AI: Artificial intelligence, RD: Retinal detachment, MD: Macular hole, ERM: Epiretinal membrane.

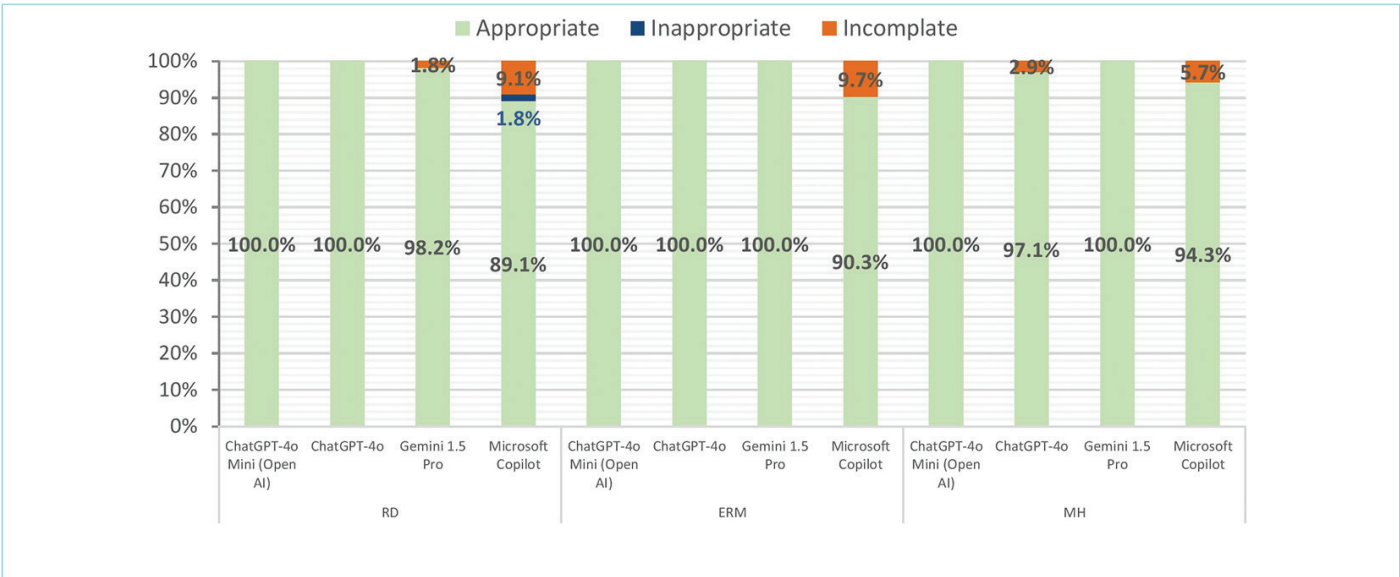


Figure 2. The suitability of responses from LLM-chatbots to questions related to retinal detachment, epiretinal membrane, and macular hole. LLM: Large language model, AI: Artificial intelligence, RD: Retinal detachment, MD: Macular hole, ERM: Epiretinal membrane.

Table 2. Readability indices of responses from LLM-chatbots to retinal detachment, epiretinal membrane, and macular hole							
		Likert	Flesch-Kincaid grade level	Gunning fog index	Coleman liau index	SMOG index	Flesch reading ease score
RD (n=54)	ChatGPT-4o mini (OpenAI)	4.9±0.3	16±5.8	19.8±3.4	16.2±2.6	17.8±2.2	23.8±7.5
	ChatGPT-4o	5±0.2	12±2.2	14.4±2	14.4±1.8	14.5±2	36.5±8.8
	Gemini 1.5 Pro	4.9±0.4	12.2±1.9	15.2±2.7	13.7±1.9	14.9±2.1	33.8±8.5
	Microsoft copilot	4.8±0.4	12.4±3.1	14.7±2.3	14±1.7	14.5±2.2	32.4±9.7
		0.063	<0.001	<0.001	<0.001	<0.001	<0.001
ERM (n=31)	ChatGPT-4o mini (OpenAI)	4.9±0.4	16.8±1.5	17.8±3	15.3±3.3	16.7±2.2	24±6.4
	ChatGPT-4o	4.9±0.4	11.9±1.9	14.6±2.2	13.8±1.9	15.4±2	34.8±8.8
	Gemini 1.5 pro	5±0.1	12.6±2.1	14.1±2.1	14.1±1.9	14.4±1.9	32.8±8.6
	Microsoft copilot	4.9±0.2	12.1±1.9	14.3±2	13.9±2	15±2.6	34.3±10.8
		0.532	<0.001	<0.001	0.036	<0.001	<0.001
MH (n=35)	ChatGPT-4o mini (OpenAI)	4.8±0.4	16.5±1.6	17.5±4.1	14.3±3	16.4±2.9	28.6±9.6
	ChatGPT-4o	5±0	11.9±1.7	14.7±2.1	14.1±1.8	14.9±2	32.6±9.6
	Gemini 1.5 pro	5±0.2	12.1±2.1	15.1±2.5	13.4±1.9	14.9±2.5	33.6±9.5
	Microsoft copilot	4.8±0.5	11.9±2.1	14.3±2.4	14.2±2.2	14.7±2.1	34.8±10.5
		0.086	<0.001	<0.001	0.372	0.013	0.052

*p<0.05. Values are presented as mean ± SD.
RD: Retinal detachment, ERM: Epiretinal membrane, MH: Macular hole, SD: Standard deviation, LLM: Large language model, SMOG: Simple measure of Gobbledygook index.

Analysis of the relationship between response appropriateness and readability metrics revealed that models with higher readability scores, such as ChatGPT-4o, tended to have higher appropriateness rates. Despite the appropriateness of ChatGPT-4o mini’s responses, it remained the least readable model. Gemini 1.5 Pro performed similarly to ChatGPT-4o in response appropriateness but had lower readability scores in some metrics.

Figure 3 presents scatter plots of readability indices for LLM-generated responses, illustrating variation across models. Correlation analyses by disease group revealed no strong associations, indicating variability

in readability metrics. Notably, ChatGPT-4o mini generated responses with higher GFI and SMOGI scores, reflecting longer and more complex sentences, whereas other models exhibited lower scores, suggesting simpler, more readable responses. These results demonstrate that LLMs do not consistently produce responses of equal readability; measured readability varies depending on the applied readability metric (Supplementary Table 2).

Response Length and Word Count

The word counts of LLM-generated responses are presented in Supplementary Table 3. Significant differences were observed

Table 3. Intraclass correlation coefficients among observers					
	Likert 1 × Likert 2	ICC	95% CI		p
			Lower bound	Upper bound	
RD	ChatGPT-4o mini	0.770	0.605	0.866	<0.001*
	ChatGPT-4o	0.885	0.803	0.933	<0.001*
	Gemini 1.5 pro	0.942	0.900	0.966	<0.001*
	Microsoft copilot	0.895	0.821	0.939	<0.001*
ERM	ChatGPT-4o mini	0.937	0.870	0.970	<0.001*
	ChatGPT-4o	0.950	0.897	0.976	<0.001*
	Gemini 1.5 pro	0.940	0.876	0.971	<0.001*
	Microsoft copilot	1.000	1.000	1.000	
MH	ChatGPT-4o mini	0.955	0.911	0.977	<0.001*
	ChatGPT-4o	1.000	1.000	1.000	
	Gemini 1.5 pro	1.000	1.000	1.000	
	Microsoft copilot	0.971	0.943	0.985	<0.001*
Total	ChatGPT-4o mini	0.891	0.843	0.924	<0.001*
	ChatGPT-4o	0.931	0.901	0.952	<0.001*
	Gemini 1.5 pro	0.926	0.894	0.949	<0.001*
	Microsoft copilot	0.938	0.911	0.957	<0.001*

*p<0.05.
ICC: Intraclass correlation coefficient, CI: Confidence interval, RD: Retinal detachment, ERM: Epiretinal membrane, MH: Macular hole.

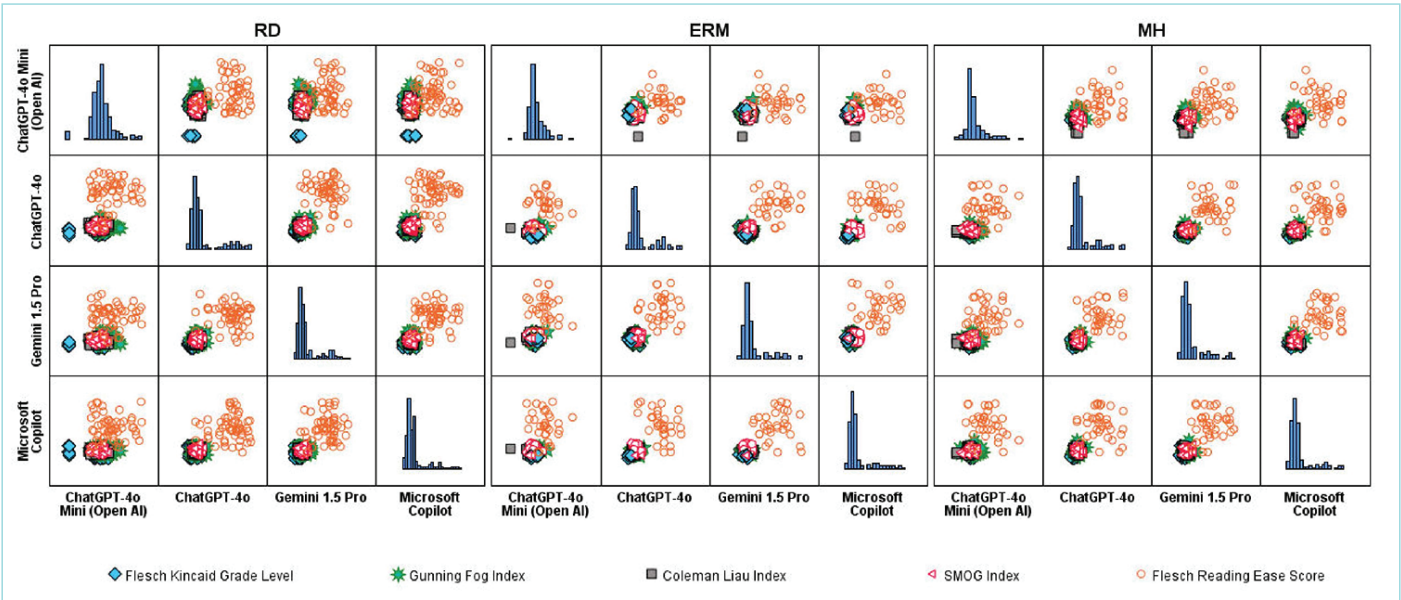


Figure 3. Scatter Plot Among LLM-chatbots to retinal detachment, epiretinal membrane, and macular hole. X-axis represents the readability scores obtained by each LLM-chatbot for a given metric. Y-axis represents the distribution of readability scores across different disease groups. The scatter plots indicate a lack of strong correlation among chatbot responses, highlighting the variability in readability metrics.
RD: Retinal detachment, ERM: Epiretinal membrane, MH: Macular hole, LLM: Large language model, SMOG: Simple measure of Gobbledygook index, AI: Artificial intelligence.

across disease groups ($p < 0.001$). ChatGPT-4o mini produced the shortest responses, likely due to its optimized architecture, which compresses information for efficiency. This design enables faster responses but results in more concise outputs compared to larger models.

DISCUSSION

Our findings indicate that LLMs provide comprehensive responses to vitreoretinal surgery-related questions, with comparable performance across models. However, Microsoft Copilot demonstrated the lowest appropriateness, particularly in diagnosis, treatment, and prevention, which was most pronounced in the RD group. This aligns with previous studies reporting Bing (now Microsoft Copilot) performing worse in similar categories.^{19,20}

Several studies have evaluated Gemini's appropriateness in ophthalmology, primarily focusing on theoretical knowledge in exam-based formats, without assessing readability or accessibility for patient education.²¹ Another study analyzed Gemini's ability to detect macular diseases in OCT images but did not evaluate chatbot-generated responses.²² Similarly, research on chatbot responses to age-related macular degeneration and diabetic retinopathy was limited to two disease groups.²³ Unlike previous studies, our research assesses both the appropriateness and readability of patient-centered chatbot responses across a broader spectrum of ophthalmic conditions, providing insights into those responses.

While LLM chatbots generate grammatically correct and human-readable texts, their suitability for patient education depends on meeting specific readability criteria. ChatGPT-4o mini exhibited the lowest readability despite its high appropriateness, likely due to its use of more technical, condensed language, complex sentence structures, and greater information density. This pattern aligns with findings from a study comparing ChatGPT-4o and ChatGPT-4o mini.²⁴

The American Medical Association recommends that health-related materials should be at a 5th to 6th-grade reading level.²⁵ However, our results indicate that LLM-generated responses are often complex, with low readability scores. This aligns with previous research demonstrating low readability in online RD and ERM resources, as well as findings that ChatGPT-4 responses for atrial fibrillation education had lower readability scores than other chatbots.²⁶⁻²⁸ In contrast, a study on melanoma found no significant readability differences among Bard, Bing, and ChatGPT-4.²⁹ These discrepancies suggest that some LLMs favor technical and academic language, potentially limiting accessibility for patient education. Improving readability could enhance public health communication and information dissemination.

Our readability analysis highlights that complex sentence structures may hinder patient comprehension.³⁰ Thus, LLM-generated responses should be evaluated not only for appropriateness but also for readability in medical education materials.^{31,32}

The integration of chatbots into ophthalmic surgery education and patient follow-up is becoming increasingly relevant.³³ Preoperatively, chatbot-assisted education can provide realistic expectations about surgery, potentially reducing patient anxiety.³⁴ Additionally, these systems can enhance patient-doctor interactions by enabling patients

to ask more informed questions.³⁰ In the postoperative phase, chatbots play a crucial role in improving patient adherence, monitoring for complications, tracking medication compliance, and encouraging follow-up visits.^{31,32} Such tools can alleviate clinical workload, facilitate more effective communication, and contribute to better clinical outcomes.³³

Our findings underscore that readability is not just a technical parameter but has a direct impact on patient understanding and decision-making.³⁴ Optimizing chatbot-generated responses for readability is essential not only for patient education but also for clinical decision-support systems.³³ Improving readability is essential for effective patient communication, as high clinical appropriateness does not necessarily translate into better patient comprehension.^{31,34}

Among all disease groups, ChatGPT-4o mini produced the shortest responses. While Gemini 1.5 Pro was previously reported to generate longer responses, our study found that ChatGPT-4o, Gemini 1.5 Pro, and Microsoft Copilot produced similar response lengths, all longer than ChatGPT-4o mini.³⁵ The shorter responses from ChatGPT-4o mini may result from its smaller, optimized architecture, designed to maximize efficiency and speed, leading to more concise output. Although shorter responses are generally expected to improve readability, ChatGPT-4o mini demonstrated the lowest readability scores despite producing the shortest responses. This finding suggests that readability was influenced more by linguistic complexity and technical terminology than by response length alone.

Study Limitations

This study has several limitations. First, we focused solely on three retinal diseases, thereby limiting the generalizability of the findings to other ophthalmologic or medical fields. Future research should be conducted across multiple specialties to enhance generalizability. Second, response appropriateness was assessed using a subjective rating system; however, 99.2% of responses were deemed appropriate or partially appropriate, supporting the reliability and validity of our evaluation method. While the subjectivity of the Likert scale is a concern, the use of two independent retina specialists and the attainment of a high ICC score enhanced test consistency. Additionally, our study design ensured that evaluators were blinded to the chatbot source, minimizing potential bias. However, in clinical practice, experts may interpret responses differently, highlighting the need for multidisciplinary validation of the appropriateness of LLMs. Another limitation is that LLMs do not provide references for their responses, making it impossible to validate the content using standardized quality assessment tools such as DISCERN or HONcode. The lack of citations poses challenges for patient education, as users cannot verify the reliability of chatbot-generated medical content. This is particularly concerning for patient decision-making and underscores the importance of reference-backed chatbot outputs in clinical applications. Furthermore, although established readability indices provide standardized measures of linguistic complexity, they do not directly assess patient comprehension or information retention. Therefore, readability scores should be interpreted as surrogate measures of text accessibility rather than direct indicators of patient comprehension. Lastly, LLMs are continuously updated, meaning that our findings reflect a specific time frame and may not remain consistent with future iterations. Consequently, the exact reproducibility of our

findings may be limited as model architectures, training data, and response generation strategies continue to evolve. Given the rapid evolution of AI models, periodic reassessment is necessary to ensure the ongoing validity and clinical relevance of comparative evaluations.

CONCLUSION

ChatGPT-4o, ChatGPT-4o mini, and Gemini 1.5 Pro provided highly clinically appropriate responses for RD, ERM, and MH. However, ChatGPT-4o demonstrated superior readability, whereas ChatGPT-4o mini, despite its appropriateness, exhibited the lowest readability owing to its condensed, technical responses. These findings highlight that readability is as essential as appropriateness in patient education. Future research should focus on optimizing LLM-generated content for better readability and comprehension, ensuring their effective integration into medical education and clinical decision-making. Nevertheless, LLM-generated information should be considered a complementary educational resource and not replace professional ophthalmologic consultation or clinical judgment.

MAIN POINTS

- Large language models (LLMs) demonstrated high clinical appropriateness when answering questions about vitreoretinal surgical diseases; ChatGPT-4o, ChatGPT-4o mini, and Gemini 1.5 Pro achieved near-perfect appropriateness rates.
- Despite comparable clinical appropriateness across models, Microsoft Copilot exhibited lower reliability, particularly for treatment- and prognosis-related responses, underscoring variability among AI systems.
- ChatGPT-4o produced the most readable responses across all disease groups, whereas ChatGPT-4o mini generated more technical and less accessible outputs despite high clinical appropriateness.
- Readability and clinical appropriateness were not always correlated, indicating that highly clinically appropriate responses may still be difficult for patients to understand.
- Optimizing LLM-generated medical content for readability is essential to improve patient education, communication, and clinical decision-making.

ETHICS

Ethics Committee Approval: This study did not involve human participants, patient data, or identifiable personal information. Therefore, it did not require formal ethical approval. The Institutional Review Board (IRB) of our institution reviewed the study protocol and confirmed that it met the criteria for exemption from ethical review. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Informed Consent: Informed consent was not required due to the absence of human subjects.

Footnotes

Authorship Contributions

Surgical and Medical Practices: E.D., A.M., A.E., Concept: E.D., A.M., A.E., Design: E.D., A.M., A.E., Data Collection and/or Processing: E.D., A.M.,

A.E., Analysis and/or Interpretation: E.D., A.M., A.E., Literature Search: E.D., A.M., A.E., Writing: E.D., A.M., A.E.

DISCLOSURES

Conflict of Interest: No conflict of interest was declared by the authors.

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

Declaration on the Use of Artificial Intelligence (AI): Artificial intelligence has been used to assist with “text editing”.

Supplementary Tables 1-3 Link: <https://d2v96fxpocvxx.cloudfront.net/c441b065-ac5d-471a-8435-98bcb2373414/content-images/c131dd2d-1800-4ee9-ae53-14c957b968b4.pdf>

REFERENCES

1. Xu L, Sanders L, Li K, Chow JCL. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. *JMIR Cancer*. 2021; 7(4): e27850.
2. Haupt CE, Marks M. AI-generated medical advice-GPT and beyond. *JAMA*. 2023; 329(16): 1349-50.
3. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023; 388(13): 1233-9.
4. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health*. 2023; 11: 1166120.
5. OpenAI. GPT-4o Mini: Advancing cost-efficient intelligence. Available from: <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence>.
6. OpenAI. Hello GPT-4o. Published May 13, 2024. Available from: <https://openai.com/index/hello-gpt-4o/>
7. Pichai S, Hassabis D. Our next-generation model: Gemini 1.5. Google. Published February 15, 2024. Available from: <https://blog.google/innovation-and-ai/products/google-gemini-next-generation-model-february-2024/>
8. Mehdi Y. Announcing Microsoft Copilot, your everyday AI companion. Microsoft. Published September 21, 2023. Available from: <https://blogs.microsoft.com/blog/2023/09/21/announcing-microsoft-copilot-your-everyday-ai-companion/>
9. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*. 2020; 33: 9459-74.
10. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. *Radiology*. 2023; 307(2): e230163.
11. Siddique S, Chow JCL. Machine learning in healthcare communication. *Encyclopedia*. 2021; 1(1): 220-39.
12. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature*. 2023; 614(7947): 224-6.
13. Momenai B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. *Ophthalmol Retina*. 2023; 7(10): 862-8.
14. Norman G. Likert scales, levels of measurement and the “laws” of statistics. *Adv Health Sci Educ Theory Pract*. 2010; 15(5): 625-32.
15. Readability is an essential content marketing tool. *Readable*. <https://readable.com/readability/#goodscore>.

16. Patel AJ, Kloosterboer A, Yannuzzi NA, Venkateswaran N, Sridhar J. Evaluation of the content, quality, and readability of patient accessible online resources regarding cataracts. *Semin Ophthalmol*. 2021; 36(5-6): 384-91.
17. Basch CH, Mohlman J, Hillyer GC, Garcia P. Public health communication in time of crisis: readability of on-line COVID-19 information. *Disaster Med Public Health Prep*. 2020; 14(5): 635-7.
18. Robinson E, McMenemy D. 'To be understood as to understand': a readability analysis of public library acceptable use policies. *J Librariansh Inf Sci*. 2020; 52(39): 713-25.
19. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun CH, Lam JSH, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine*. 2023; 95: 104770.
20. Rasmussen MLR, Larsen AC, Subhi Y, Potapenko I. Artificial intelligence-based ChatGPT chatbot responses for patient and parent questions on vernal keratoconjunctivitis. *Graefes Arch Clin Exp Ophthalmol*. 2023; 261(10): 3041-3.
21. Mihalache A, Grad J, Patil NS, Huang RS, Popovic MM, Mallipatna A, et al. Google Gemini and Bard artificial intelligence chatbot performance in ophthalmology knowledge assessment. *Eye (Lond)*. 2024; 38(13): 2530-5.
22. Antaki F, Chopra R, Keane PA. Vision-language models for feature detection of macular diseases on optical coherence tomography. *JAMA Ophthalmol*. 2024; 142(6): 573-6.
23. Cheong KX, Zhang C, Tan TE, Fenner BJ, Wong WM, Teo KY, et al. Comparing generative and retrieval-based chatbots in answering patient questions regarding age-related macular degeneration and diabetic retinopathy. *Br J Ophthalmol*. 2024; 108(10): 1443-9.
24. Wang S, Wang Y, Jiang L, Chang Y, Zhang S, Zhao K, et al. Assessing the clinical support capabilities of ChatGPT 4o and ChatGPT 4o mini in managing lumbar disc herniation. *Eur J Med Res*. 2025; 30(1): 45.
25. Weiss BD. Health literacy and patient safety: help patients understand. Manual for clinicians. 2nd ed. Chicago, IL: American Medical Association Foundation and American Medical Association; 2007.
26. Redick DW, Hwang JC, Kloosterboer A, Yannuzzi NA, Patel NA, Kuriyan AE, et al. Content, readability, and accountability of freely available online information for patients regarding epiretinal membranes. *Semin Ophthalmol*. 2022; 37(1): 67-70.
27. Ozturk Y, Gungor CB. Assessing the quality, reliability, and readability of online information on rhegmatogenous retinal detachment quality information retinal detachment. *Research Square (Preprint)*. 2022.
28. Hillmann HAK, Angelini E, Karfoul N, Feickert S, Mueller-Leisse J, Duncker D. Accuracy and comprehensibility of chat-based artificial intelligence for patient information on atrial fibrillation and cardiac implantable electronic devices. *Europace*. 2023; 26(1): eua369.
29. Mu X, Lim B, Seth I, Xie Y, Cevik J, Sofiadellis F, et al. Comparison of large language models in management advice for melanoma: Google's AI BARD, BingAI and ChatGPT. *Skin Health Dis*. 2023; 4(1): e313.
30. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. *J Med Internet Res*. 2023; 25: e40789.
31. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med Educ*. 2023; 9: e48291.
32. Li Z, Li F, Fu Q, Wang X, Liu H, Zhao Y, et al. Large language models and medical education: a paradigm shift in educator roles. *Smart Learn*. 2024; 11(1): 26.
33. Moldt JA, Festl-Wietek T, Madany Mamlouk A, Nieselt K, Fuhl W, Herrmann-Werner A. Chatbots for future docs: exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots. *Med Educ Online*. 2023; 28(1): 2182659.
34. Wah JNK. Revolutionizing e-health: the transformative role of AI-powered hybrid chatbots in healthcare solutions. *Front Public Health*. 2025; 13: 1530799.
35. Amin KS, Mayes LC, Khosla P, Doshi RH. Assessing the efficacy of large language models in health literacy: a comprehensive cross-sectional study. *Yale J Biol Med*. 2024; 97(1): 17-27.