Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Expert evaluation and readability of conversational AI responses to common ophthalmic patient questions: an exploratory generational cross-sectional study

Javier Gismero Rodriguez
University of Malaga
Carlos Ruiz Nuñez
University of Malaga
Antonio Jose Garcia Ruiz
ajgr@uma.es

University of Malaga

Research Article

Keywords: Artificial intelligence, Health literacy, Patient education, Ophthalmology, Readability, Large language models

Posted Date: April 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9242640/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Purpose:

To evaluate expert-rated informational quality and linguistic accessibility of responses generated by a contemporary large language model to common ophthalmic patient questions, and to explore generational evolution in conversational artificial intelligence (AI) performance.

Methods:

In this cross-sectional exploratory study, 12 frequently asked ophthalmology questions were used to generate patient-oriented responses from ChatGPT-5 using a standardized specialist-role prompt. Responses were evaluated by ophthalmologists using validated instruments assessing Global Quality Score (GQS), Reliability Score (RS), and Usefulness Score (US). Readability was analyzed using the Flesch–Szigriszt Index (FSI) and categorized according to the INFLESZ scale. Descriptive analyses were performed, and findings were contextually interpreted against previously reported data obtained using identical questions and evaluation methods. Domain-specific variability and correlations among evaluation constructs were explored.

Results:

Responses were rated favorably across expert-assessed domains (mean GQS 4.01, RS 5.40, US 5.68), with contextual comparisons suggesting modest generational improvements. In contrast, readability showed a substantial increase (mean FSI 67.7 vs 53.9), corresponding to a shift from “somewhat difficult” to “fairly easy” patient comprehension. Performance changes were domain-dependent, with greater gains in explanatory topics than in context-sensitive counselling. Readability demonstrated minimal correlation with expert-rated quality constructs.

Conclusions:

Generational development of conversational AI in ophthalmology suggests a tendency to improve linguistic accessibility more consistently than expert-perceived informational quality. Conversational AI may therefore support patient education by improving communicative clarity, which would be useful in settings with limited consultation time. Careful clinical integration and further evaluation of real-world educational impact remain necessary.

1. Introduction

Effective communication represents a fundamental part of ophthalmology yet conveying complex visual and surgical concepts to patients remains challenging in routine clinical practice. Many ophthalmic conditions involve elaborate anatomical and optical concepts, technical terminology, and complex decisions that can be difficult for patients to fully understand, particularly in the context of visual impairment or limited consultation time. In ophthalmology, limited patient understanding and suboptimal communication have been linked to poorer adherence, follow-up, and care experience, highlighting the importance of strategies aimed at improving patient understanding and engagement[1, 2].

In recent years, digital technologies have been increasingly explored as tools to support patient education beyond clinical encounter. Among these, large language model (LLM)-based conversational agents have attracted growing attention due to their ability to generate natural-language responses to health-related queries[3]. Early ophthalmology literature suggests that these systems can generate structured and potentially understandable information, although concerns about reliability and clinical appropriateness remain[4–7]. As conversational AI systems continue to evolve, understanding the nature and clinical relevance of generational performance changes has become an emerging research priority.

Previous investigations have primarily focused on benchmarking informational accuracy or comparing chatbot responses with those of clinicians[8–10]. However, the potential role of conversational AI in enhancing health literacy may depend not only on informational correctness but also on linguistic accessibility and communicative clarity. The relationship between these dimensions remains insufficiently characterized, particularly in ophthalmology, where patient education needs are substantial and topic-dependent variability may influence the perceived usefulness of AI-generated information[10, 11].

The present study was therefore designed to evaluate expert-rated quality metrics and readability of patient-oriented responses generated by a contemporary LLM in ophthalmology. While previous studies have primarily assessed the accuracy or overall performance of individual models, few investigations have systematically examined how these dimensions evolve across successive model generations using consistent evaluation frameworks. By using an identical set of frequently asked clinical questions and comparable evaluation instruments to those previously reported, this exploratory analysis also aimed to contextualize generational evolution in conversational AI performance. In addition, domain-specific variability and conceptual relationships between readability and expert-perceived informational quality were examined to better understand the potential educational role of these systems.

2. Materials and Methods

2.1. Study Design and Objectives

This cross-sectional exploratory study evaluated the quality, reliability, usefulness, and readability of ophthalmology-related patient information generated by a contemporary large language model (LLM), ChatGPT-5. Model selection was informed by an independent benchmarking exercise using a standardized ophthalmology examination; details are provided in Supplementary File 1. The study focused on patient-centered responses to frequently asked ophthalmology questions and was conducted in accordance with STROBE reporting recommendations for observational studies.

The primary objective was to characterize expert-rated quality metrics and linguistic accessibility of ChatGPT-5 responses to common ophthalmic patient queries. A secondary objective was to contextualize these findings against previously published data obtained using ChatGPT-3.5, thereby examining generational evolution in model performance. Additional analyses explored interrelationships among evaluation scales and domain-specific variability across questions.

2.2. Curation of Questions and Expert Evaluation Framework

Patient-oriented ophthalmology questions were derived from a previously published dataset developed by Ruiz-Núñez et al.[12], in which ophthalmologists from multiple medical institutions in Spain contributed frequently asked patient questions, ensuring diversity in subspecialty representation and geographical distribution.

From the original list of 36 questions, 12 were curated for the present study based on clinical relevance, clarity, and representativeness of routine outpatient concerns across ophthalmology subspecialties. These questions reflect common patient inquiries encountered in daily practice:

- Q1. Will my cataracts be operated on with a laser?
- Q2. Is eye pressure related to blood pressure?
- Q3. How many drops of eye drops should I use each time?
- Q4. If I have surgery, will I stop being myopic?
- Q5. Why don’t ophthalmologists get surgery to get rid of their glasses?
- Q6. What is considered normal eye pressure?
- Q7. Until when will my myopia increase?
- Q8. Is vitreous detachment the same as retinal detachment?
- Q9. When will I regain my vision after retinal surgery?

- Q10. Do I have macula, doctor?
- Q11. Will it hurt a lot if I get an injection in my eye?
- Q12. Is astigmatism for distance or near vision?

These questions were used to generate ChatGPT-5 responses, which were subsequently evaluated using the same validated instruments previously applied in the ChatGPT-3.5 study. This methodological continuity enabled contextual comparison while maintaining the exploratory focus of the present investigation.

2.3. Response Generation Protocol

Each of the 12 curated questions was submitted independently to ChatGPT-5 using a standardized specialist-role prompt: “As an expert ophthalmologist, how would you respond to this patient’s question?”. All responses were generated in Spanish to reflect real-world patient communication contexts. No additional contextual information or follow-up prompts were provided. Each question was submitted once (single-shot generation), and no post-editing, correction, or manual refinement of outputs was performed. Model responses were generated within a defined timeframe to reduce potential version drift. Default system parameters were maintained throughout the generation process to ensure internal consistency across responses. The resulting texts were anonymized and randomized prior to expert evaluation. The prompting strategy and evaluation framework were intentionally aligned with those used in the previously published ChatGPT-3.5 study to enable contextual comparison while maintaining the exploratory focus of the present investigation.

2.4. Expert evaluation of generated responses

The ophthalmologists evaluated ChatGPT-5’s responses using three validated scales. The Global Quality Scale (GQS), on a five-point scale (where 1 indicates the worst information and 5 the best), was used to assess the quality and ease of understanding of the information. The Reliability Score (RS) scale evaluated the reliability of the medical sources, and the Usefulness Score (US) scale assessed the usefulness of the response for the patient, both on a scale of 0 to 7, where 7 indicates the highest reliability and usefulness, respectively.

A readability analysis was performed of both ChatGPT-3.5 and ChatGPT-5’ responses using the Flesch–Szigriszt Index (FSI)[13], a validated readability formula widely used for Spanish-language health texts. The FSI was calculated for each response using the following equation: $FSI = 206.835 - (62.3 \times \text{syllables/word}) - (\text{words/sentence})$.

The score ranges from 0 to 100, higher FSI values indicating easier readability. For interpretability, numerical FSI scores were categorized according to the validated INFLESZ scale[14], which classifies texts into qualitative readability levels tailored to a

Spanish reader. Although the INFLESZ categories were originally defined in Spanish, they were translated into English for reporting purposes as follows: *muy difícil* (very difficult, 0–40), *algo difícil* (somewhat difficult, 40–55), *normal* (standard, 55–65), *bastante fácil* (fairly easy, 65–80), and *muy fácil* (very easy, 80–100).

2.5. Statistical Analysis

Expert evaluation data for ChatGPT-5 responses (n = 85 evaluations) were analyzed and descriptively contextualized against previously published ChatGPT-3.5 data (n = 21 evaluations) obtained using the same questions and evaluation instruments. The combined dataset comprised 1,272 individual assessments across questions and outcome measures.

Statistical analyses were performed using Jamovi (version 2.6.44). Readability analyses were conducted using Python (OS, Pandas, and Textstat libraries).

2.5.1. Descriptive Analysis

Primary analyses focused on the descriptive characterization of ChatGPT-5 performance. Means, standard deviations (SD), medians, and 95% confidence intervals (CI) were calculated for all quantitative variables, including GQS, RS, US, and FSI.

Normality of continuous variables (GQS, RS, US, and FSI) was assessed using the Shapiro–Wilk test and visual inspection of histograms and Q–Q plots. As several variables deviated from normal distribution, non-parametric methods were applied for inferential comparisons.

2.5.2. Contextual Comparison with Historical Data

To explore generational evolution, ChatGPT-5 descriptive metrics were contextualized against previously reported ChatGPT-3.5 results derived using identical instruments and question sets. Given the non-contemporaneous nature of the datasets and unequal evaluator sample sizes, these comparisons were considered exploratory and interpreted cautiously.

Where appropriate, non-parametric tests (Mann–Whitney U) were applied to examine directional differences between cohorts, given deviations from normality. Effect sizes were estimated using rank-biserial correlation (r). These analyses were intended to provide contextual insight rather than establish formal superiority.

2.5.3. Variability Across Questions

To assess domain-specific variability in ChatGPT-5 performance, analyses were conducted across the 12 questions. Variability was examined with two-way ANOVA as an exploratory pattern-detection analysis to examine interaction structure across questions and cohorts

and were interpreted descriptively rather than as confirmatory inferential tests. Interaction patterns between question domain and score distributions were explored to identify areas of relative strength or decline compared with historical benchmarks.

2.5.4. Item-Level Analysis

Item-level analyses were performed to characterize question-specific patterns in ChatGPT-5 responses. Differences relative to previously reported ChatGPT-3.5 values were described to highlight potential domain-dependent evolution. Because evaluators assessed multiple responses, residual within-rater dependence may have been present and was not explicitly modelled in this exploratory analysis.

2.5.5. Correlation Analysis

Spearman's rank correlation coefficient (ρ) was used to assess associations among GQS, RS, US, and FSI within the ChatGPT-5 dataset. This analysis explored conceptual overlaps between expert-rated constructs and the relationship between readability and perceived quality.

2.5.6. Readability Analysis

FSI values were calculated for each response. Absolute differences in FSI between ChatGPT-5 and previously reported ChatGPT-3.5 values (ΔFSI) were computed descriptively to quantify shifts in linguistic accessibility. Categorical changes across INFLESZ levels were summarized.

All statistical tests were two-sided, and a p-value < 0.05 was considered statistically significant. Given the exploratory and contextual nature of between-cohort comparisons, results were interpreted with caution, emphasizing magnitude and clinical relevance rather than statistical significance alone.

3. Results

3.1. Study Sample and Evaluations

Expert ratings of ChatGPT-5 responses were obtained from 85 ophthalmologists using the same validated instruments previously applied in a published ChatGPT-3.5 study ($n = 21$ evaluators). Across the 12 curated ophthalmology questions, a total of 1,272 individual assessments were analyzed when combining the historical and current datasets for contextual interpretation.

Primary analyses focused on the descriptive characterization of ChatGPT-5 performance, with previously reported ChatGPT-3.5 data presented for contextual reference.

3.2. Overall Expert-Rated Performance of ChatGPT-5

3.2.1. Global Quality Score (GQS)

ChatGPT-5 achieved a mean GQS of 4.01 (95% CI 3.97–4.06), with a median of 4 and a standard deviation of 0.777 (Table 1). Scores clustered toward the upper end of the 5-point scale, indicating generally high perceived quality and clarity.

When contextualized against previously reported ChatGPT-3.5 results (mean 3.86; 95% CI 3.73–3.99), the observed difference (+ 0.15 points) suggests a directional improvement in expert-perceived quality. However, the absolute magnitude of change was modest.

Table 1
Descriptive analysis of main variables

				95% Confidence Interval			
Variable	Model	Mean	Δ Mean	Lower	Upper	Median	SD
GQS (Quality) Score 1–5	**ChatGPT-3.5**	3.86	+ 0.15	3.73	3.99	4	1.058
	ChatGPT-5	4.01		3.97	4.06	4	0.777
RS (Reliability) Score 1–7	**ChatGPT-3.5**	5.04	+ 0.36	4.83	5.24	5	1.680
	ChatGPT-5	5.40		5.32	5.49	6	1.382
US (Usefulness) Score 1–7	**ChatGPT-3.5**	5.15	+ 0.53	4.94	5.36	6	1.692
	ChatGPT-5	5.68		5.60	5.75	6	1.230
FSI (Readability) Score 0-100	**ChatGPT-3.5**	53.9	+ 13.8	49.06	58.7	53.3	7.62
	ChatGPT-5	67.7		61.37	74.1	66.6	10.03

3.2.2. Reliability Score (RS)

The mean RS for ChatGPT-5 was 5.40 (95% CI 5.32–5.49), with a median of 6 and SD of 1.382. Ratings were consistently above the midpoint of the 7-point scale, reflecting generally favorable perceptions of trustworthiness.

Compared with prior ChatGPT-3.5 findings (mean 5.04; 95% CI 4.83–5.24), ChatGPT-5 demonstrated a directional increase (+ 0.36 points). Reduced dispersion in the ChatGPT-5 cohort suggests greater consistency in reliability assessments.

3.2.3. Usefulness Score (US)

ChatGPT-5 achieved a mean US of 5.68 (95% CI 5.60–5.75), median 6, SD of 1.230. This domain demonstrated the highest absolute values among the expert-rated constructs.

Relative to previously reported ChatGPT-3.5 performance (mean 5.15; 95% CI 4.94–5.36), the observed increase (+ 0.53 points) indicates a directional gain in perceived practical utility.

Across all three expert-rated domains, ChatGPT-5 responses were evaluated favorably, with contextual comparisons suggesting incremental generational refinement rather than marked performance shifts.

3.2.4. Readability Analysis

Readability analysis revealed a substantial increase in linguistic accessibility. ChatGPT-5 responses were significantly shorter than GPT-3.5 responses (median 85.5 vs 257.5 words, $p < 0.001$), although sentence length was similar in both models (23 vs 24.4 words, $p < 0.319$).

The mean FSI for ChatGPT-5 responses was 67.7 (95% CI 61.4–74.1), compared with 53.9 (95% CI 49.1–58.7) in the previously reported ChatGPT-3.5 dataset. This represents an absolute increase of + 13.8 points. According to INFLESZ categorization, this shift corresponds to a transition from “Somewhat difficult” readability to “Fairly easy” readability. Unlike the modest directional changes observed in expert-rated quality metrics, the readability improvement was substantial in magnitude. This suggests that generational model evolution was associated with a marked enhancement in linguistic simplicity and accessibility.

3.3. Variability Across Questions

Exploratory two-factor analyses incorporating question domain and model cohort (historical ChatGPT-3.5 vs current ChatGPT-5 dataset) were performed to examine whether score distributions varied across clinical topics.

A significant main effect of question was observed across GQS, RS, and US ($p < 0.001$), indicating that expert ratings differed according to topic area. In addition, a significant model × question interaction was detected for all three outcomes ($p < 0.001$) (Table 2), suggesting that directional changes between generational datasets were not uniform across domains.

Table 2 *Variability analysis of main variables, two-way ANOVA (GQS: Global Quality Scale, RS: Reliability scale. US: Usefulness Scale, FSI: Flesz-Szigriszt Index*

Outcome	Effect	F(df1, df2)	p-value
GQS	Model	$F(1,1248) = 7.70$	0.006
	Question	$F(11,1248) = 3.79$	$< .001$
	Model × Question	$F(11,1248) = 4.82$	$< .001$
RS	Model	$F(1,1248) = 13.53$	$< .001$
	Question	$F(11,1248) = 3.27$	$< .001$
	Model × Question	$F(11,1248) = 3.08$	$< .001$
US	Model	$F(1,1248) = 32.46$	$< .001$
	Question	$F(11,1248) = 2.72$	0.002
	Model × Question	$F(11,1248) = 3.44$	$< .001$

These findings indicate that performance evolution appears domain-dependent rather than globally consistent. Improvements were more pronounced in selected topics, whereas other areas demonstrated minimal change or isolated regressions. Given the non-contemporaneous nature of the cohorts and unequal evaluator sample sizes, these analyses were considered exploratory pattern-detection tools rather than confirmatory inferential testing.

3.4. Item-Level Analysis

Item-level analyses were conducted to characterize domain-specific patterns in ChatGPT-5 performance and to contextualize findings relative to previously reported ChatGPT-3.5 results (Fig. 1).

Across the 12 questions, expert ratings of ChatGPT-5 responses were generally high, though variability was observed depending on topic. Directional increases in GQS were most notable in questions addressing intraocular pressure (Q2), treatment explanations (Q3), refractive outcomes (Q4), and macular terminology (Q10). Similar patterns were observed for RS, particularly in Q2 and Q3, and for US, where several items demonstrated increases exceeding one point relative to historical values.

Conversely, isolated declines were observed in selected items, including Q5 (rationale for refractive surgery decisions among ophthalmologists) and Q11 (intravitreal injection discomfort). These findings indicate that generational evolution in model performance was not uniform across domains.

Effect sizes at the item level ranged from small to moderate. Given the multiple comparisons performed and the non-contemporaneous nature of the cohorts, these analyses should be interpreted as exploratory. Rather than indicating consistent superiority, the results suggest domain-dependent refinement, with certain explanatory topics benefiting more clearly from generational model updates.

Overall, item-level variability underscores that improvements in LLM-generated ophthalmic information appear context-sensitive and may depend on question framing and subspecialty complexity. However, these item-level comparisons were exploratory and not adjusted for multiplicity. Detailed item-level statistics are provided in Supplementary File 2.

3.5. Correlation Analysis

Spearman correlation analysis (Table 3) demonstrated strong positive associations among the three expert-rated constructs. These findings suggest substantial conceptual overlap between perceived quality, reliability, and practical utility as evaluated by specialist raters.

In contrast, readability as measured by FSI showed no significant correlation with GQS ($\rho = 0.050$) and only weak, non-significant associations with RS ($\rho = 0.203$) and US ($\rho = 0.259$). This pattern indicates that linguistic accessibility operates as a distinct dimension from expert-perceived informational quality. Improvements in readability therefore appear to reflect changes in communication style rather than parallel shifts in perceived medical reliability or usefulness.

Table 3 *Correlation analysis of main variables (GQS: Global Quality Scale, RS: Reliability scale. US: Usefulness Scale, FSI: Flesz-Szigriszt Index)*

Compared metrics	Spearman Correlation Coefficient		Compared metrics	Spearman Correlation Coefficient	
GQS vs RS	0,911	p < 0,001	**GQS vs FSI**	0,050	p > 0,05
GQS vs US	0,901	p < 0,001	**RS vs FSI**	0,203	p > 0,05
RS vs US	0,900	p < 0,001	**US vs FSI**	0,259	p > 0,05

4. Discussion

In this cross-sectional exploratory study, responses generated by ChatGPT-5 to common ophthalmic patient questions were rated favorably by expert ophthalmologists across domains of quality, reliability, and usefulness. Readability was assessed as a proxy for communicative accessibility rather than a direct measure of patient comprehension. When

contextualized against previously reported ChatGPT-3.5 findings obtained using identical questions and evaluation instruments, generational differences in expert-rated performance were present but modest. In contrast, a more pronounced improvement was observed in linguistic accessibility, with readability analyses indicating a shift towards clearer and more comprehensible patient communication. These results suggest that recent advances in LLMs may be reflected more strongly in communicative refinement than in clinically perceived informational gains. Overall, successive model development appears to influence the health-literacy dimension of AI-mediated patient education to a greater extent than the expert-assessed quality of medical content.

Consistent with this selective pattern, previous studies assessing LLM chatbots in ophthalmology have reported generally acceptable clarity and practical usefulness of AI-generated responses[3, 15], alongside persistent variability in reliability and accuracy[6, 7]. Comparative investigations involving ophthalmologist responses have similarly demonstrated the capacity of these systems to provide structured and understandable explanations while revealing limitations in nuanced clinical reasoning and contextual guidance[10, 16]. The present findings extend this literature by suggesting that apparent performance gains across newer model versions may be unevenly distributed, reinforcing the view that conversational AI development in ophthalmology is likely to be incremental and domain-specific.

This non-uniform pattern prompted further exploration of topic-specific performance. Item-level analyses indicated that changes in expert ratings were not consistent across clinical domains. Directional improvements were more evident in questions requiring structured explanations of disease mechanisms, treatment rationale, or general clinical concepts, whereas limited change or isolated declines were observed in areas involving experiential symptoms, behavioral decision-making, or context-sensitive reassurance. These findings are consistent with prior work suggesting that chatbot performance may be stronger for standardized informational tasks than for nuanced, context-dependent counselling[17–19]. Such domain dependence highlights the importance of evaluating chatbot performance at the level of specific clinical topics rather than assuming uniform improvement across successive model iterations[20].

Among the domains in which generational differences appeared most pronounced was linguistic accessibility. This observation has potentially important implications for patient education in ophthalmology, where effective communication is often challenged by complex terminology, visual impairment, and the chronic course of many conditions[1, 21]. Health literacy is increasingly recognized as a determinant of treatment adherence, patient satisfaction, and clinical outcomes, and conversational AI tools may represent a scalable means of supporting patient understanding beyond the clinical encounter[22]. In this context, improved readability may contribute to greater patient engagement and

perceived usefulness of educational information. However, the absence of strong correlation between readability and expert-rated informational quality in the present analysis indicates that clearer language does not necessarily equate to greater medical reliability. Accessibility and accuracy should therefore be considered complementary but distinct dimensions when evaluating the educational role of LLMs.

These observations raise broader questions regarding the trajectory of conversational AI development in medicine. Improvements in discourse coherence, simplicity, and the simulation of empathic tone may arise from advances in model scaling and training approaches, leading to outputs that are more conversational and accessible to patients[23]. In contrast, gains in clinically nuanced reasoning, uncertainty calibration, and context-sensitive guidance may evolve more gradually, as they depend not only on language fluency but also on structured knowledge representation and decision framing[24].

This differential pattern may help explain findings from emerging real-world implementation studies. Pilot investigations evaluating chatbot use in ophthalmic patient-facing contexts have reported encouraging early feasibility and user-facing performance signals, indicating that patients are willing to engage with conversational AI for educational support and perioperative guidance[23, 25–29]. The present results provide a potential explanatory framework for such observations, suggesting that increasing communicative clarity may be a key factor underlying patient receptiveness. Chatbots may therefore be perceived as valuable not necessarily because they achieve clinician-level informational precision, but because they deliver information in a way that patients experience as accessible, understandable, and immediately relevant.

Taken together, these considerations have potential implications for clinical educational practice. LLM–based chatbots may have a role as adjunct tools to reinforce information provided during consultations, support comprehension of postoperative instructions, and facilitate preparation for clinical encounters[30–32]. Such applications may be particularly relevant in high-volume ophthalmic settings or telemedicine environments, where time constraints often limit detailed counselling[33]. Nonetheless, the modest magnitude of generational improvement in expert-perceived informational quality underscores that these systems should currently be regarded as complementary rather than substitutive to clinician-led education. There is a lack of standardized health-literacy outcomes assessment, which would allow a proper evaluation and comparison of these interventions beyond self-reported acceptance or quality. Careful integration strategies, including expert oversight and clear communication of limitations, will be essential to maximize potential benefits while minimizing risks related to incomplete or context-inappropriate guidance.

This study has several strengths. The use of validated expert-rating instruments enabled structured evaluation of multiple dimensions of informational quality. Employing an identical question set and comparable methodology to previously reported data allowed contextual exploration of performance trends across successive model iterations. In addition, inclusion of a readability assessment tailored to Spanish-language health communication provided a complementary perspective on linguistic accessibility. Finally, the item-level analytical approach offered insight into topic-specific variability, contributing to a more nuanced understanding of conversational AI performance in ophthalmic patient education.

The findings should be interpreted considering several limitations. The comparison with earlier model outputs was based on non-contemporaneous cohorts with unequal evaluator sample sizes, limiting the strength of generational inferences. Responses were generated using a single-shot prompting strategy without iterative clarification, which may differ from real-world patient interactions. The exploratory design and absence of adjustment for multiple comparisons further restrict causal interpretation. Expert perception of informational quality does not necessarily reflect patient comprehension or behavioral outcomes, and the study did not include direct measures of educational effectiveness. Linguistic and cultural specificity may also restrict generalizability to other healthcare contexts. Furthermore, the rapidly evolving nature of LLMs means that performance characteristics may change over time, highlighting the need for ongoing evaluation. Future studies should incorporate patient-centered outcomes, including comprehension, adherence, and satisfaction, to determine whether improvements in readability translate into clinically meaningful benefits.

5. Conclusion

Responses generated by ChatGPT-5 to common ophthalmic patient questions were generally rated favorably by expert ophthalmologists, with generational improvements in expert-perceived quality, reliability, and usefulness appearing modest. In contrast, a more marked enhancement was observed in communicative accessibility, indicating that recent model development may primarily influence the health-literacy dimension of patient education. These findings suggest that conversational AI systems may serve as supportive adjuncts to clinician-led counselling by facilitating clearer patient understanding, while underscoring the ongoing need for careful validation, expert oversight, and topic-specific performance assessment. Further prospective research is required to determine whether improved linguistic clarity translates into measure benefits in standardized patient health-literacy outcome metrics in ophthalmology.

Abbreviations

AI - Artificial Inteligence, LLM - large language model, GQS - Global Quality Score, RS - Reliability Score, US - Usefulness Score, FSI - Flesch-Szigriszt Index

Declarations

Author Contribution

Conceptualization: J.G.R. and C.R.N. Formal analysis: A.J.G.R. Data curation: J.G.R. and C.R.N. Investigation: J.G.R., C.R.N. and A.J.G.R Writing—original draft: J.G.R. Methodology: J.G.R. and C.R.N. Writing—review and editing: J.G.R and C.R.N. Supervision: C.R.N. and A.J.G.R. Validation and project administration: A.J.G.R. All authors have read and agreed to the published version of the manuscript.

Acknowledgement

The authors thank all the ophthalmologists who contributed to the evaluation of the questions.

Data Availability

Data are available from the corresponding author upon reasonable request.

Funding

Funding for open access charge: Universidad de Málaga / CBUA. No further involvement in the research.

Generative AI statement

During the preparation of this work the authors used ChatGPT-5.3 to translate from Spanish and draft certain sections of the paper. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Sumodhee D, Ioannidou E, Ramessur R, Prashar J, Abbas M, Balaskas K, et al. Interventions to improve patients’ knowledge in ophthalmology: A systematic review and narrative synthesis. Surv Ophthalmol. 2026;71:749–58. https://doi.org/10.1016/j.survophthal.2025.09.006

2. Ziemssen F, Loewenstein A, Aslam TM. Effective communication strategies in intravitreal therapy: How to listen and how to talk to people who need injections in their eyes. Survey of Ophthalmology [Internet]. Elsevier; 2026 [cited 2026 Mar 12];0. https://doi.org/10.1016/j.survophthal.2026.02.007
3. Bernstein IA, Zhang Y, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. JAMA Netw Open. American Medical Association; 2023;6. https://doi.org/10.1001/jamanetworkopen.2023.30320
4. 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-PACIFIC JOURNAL OF OPHTHALMOLOGY. 2024;13. https://doi.org/10.1016/j.apjo.2024.100085
5. Zheng H, Dong H, Zhao H. Trends and advances in ChatGPT applications in ophthalmology. J Fr Ophtalmol. 2025;48:104622. https://doi.org/10.1016/j.jfo.2025.104622
6. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a Large Language Model's Responses to Questions and Cases About Glaucoma and Retina Management. JAMA Ophthalmol. 2024;142:371–5. https://doi.org/10.1001/jamaophthalmol.2023.6917
7. Al-latayfeh M, Aleshawi A, El-Mulki OS, Baker M, Qaddoumi Z, Attar D, et al. Accuracy and Reproducibility of Different Artificial Intelligence Chatbots' Responses to Patient-Based Vitreoretinal Questions: A Comparative Study. OPTH. Dove Press; 2026;20:1–9.https://doi.org/10.2147/OPTH.S580133
8. Bahir D, Rostov A, Busool Abu Eta Y, Hamed Azzam S, Lockington D, Teichman JC, et al. Artificial intelligence versus ophthalmology experts: Comparative analysis of responses to blepharitis patient queries. Eur J Ophthalmol. 2025;35:1958–66. https://doi.org/10.1177/11206721251350809
9. Gürbostan Soysal G, Mercanlı M, Özer Özcan Z, Yılmaz İE, Berhuni M. Evaluating the effectiveness of chatbots and traditional resources in patient education on dry eye disease. Clin Exp Optom [Internet]. Taylor and Francis Ltd.; 2025; https://doi.org/10.1080/08164622.2025.2517750
10. Srinivasan S, Ai X, Zou M, Zou K, Kim H, Lo TWS, et al. Ophthalmological Question Answering and Reasoning Using OpenAI o1 vs Other Large Language Models. JAMA Ophthalmol. 2025;143:740–8. https://doi.org/10.1001/jamaophthalmol.2025.2413
11. Nasra M, Jaffri R, Pavlin-Premrl D, Kok HK, Khabaza A, Barras C, et al. Can artificial intelligence improve patient educational material readability? A systematic review and narrative synthesis. Internal Medicine Journal. 2025;55:20–34. https://doi.org/10.1111/imj.16607

12. Ruiz-Núñez C, Gismero Rodríguez J, Garcia Ruiz AJ, Gismero Moreno SM, Cañizal Santos MS, Herrera-Peco I. Can Generative AI Contribute to Health Literacy? A Study in the Field of Ophthalmology. Multimodal Technologies and Interaction. Multidisciplinary Digital Publishing Institute; 2024;8:79. https://doi.org/10.3390/mti8090079
13. Szigriszt Pazos F. Sistemas predictivos de legilibilidad del mensaje escrito : fórmula de perspicuidad [Internet]. Universidad Complutense de Madrid, Servicio de Publicaciones; 2001 [cited 2026 Mar 12]. https://hdl.handle.net/20.500.14352/62699. Accessed 12 Mar 2026
14. Barrio-Cantalejo IM, Simón-Lorda P, Melguizo M, Escalona I, Marijuán MI, Hernando P. [Validation of the INFLESZ scale to evaluate readability of texts aimed at the patient]. An Sist Sanit Navar. 2008;31:135–52. https://doi.org/10.4321/s1137-66272008000300004
15. Bondok M, Selvakumar R, Law C, Ing EB, Bakshi NK, Felfeli T. Comparing Ophthalmologist and Artificial Intelligence Chatbot Responses to Patient Questions. Clin Ophthalmol. 2025;19:4293–300. https://doi.org/10.2147/OPTH.S549820
16. Inooka T, Ota H, Taki Y, Yasuda S, Sajiki AF, Suzumura A, et al. Evolving Consultation: Enhancing Ophthalmic Diagnostic Performance Using Large Language Model. Ophthalmology Science [Internet]. Elsevier; 2026 [cited 2026 Mar 12];6. https://doi.org/10.1016/j.xops.2025.101004
17. Shiferaw MW, Zheng T, Winter A, Mike LA, Chan L-N. Assessing the accuracy and quality of artificial intelligence (AI) chatbot-generated responses in making patient-specific drug-therapy and healthcare-related decisions. BMC Med Inform Decis Mak. 2024;24:404. https://doi.org/10.1186/s12911-024-02824-5
18. Yau JY-S, Saadat S, Hsu E, Murphy LS-L, Roh JS, Suchard J, et al. Accuracy of Prospective Assessments of 4 Large Language Model Chatbot Responses to Patient Questions About Emergency Care: Experimental Comparative Study. J Med Internet Res. 2024;26:e60291. https://doi.org/10.2196/60291
19. Schuss P, Gonschorek AS, Kämper M, Lemcke J, Meisel H-J, Rogge W, et al. Artificial Intelligence Chatbot Responses to Patient Queries on Traumatic Brain Injury: An Expert Assessment of Reliability and Accuracy. J Neurotrauma. 2025; https://doi.org/10.1177/08977151251401539
20. Goodman RS, Patrinely JR, Stone CA Jr, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Netw Open. 2023;6:e2336483. https://doi.org/10.1001/jamanetworkopen.2023.36483
21. Wang E, Kalloniatis M, Ly A. Effective health communication for age-related macular degeneration: An exploratory qualitative study. Ophthalmic Physiol Opt. Optometrists; 2023;43:1278–93. https://doi.org/10.1111/opo.13168

22. Capó H, Edmond JC, Alabiad CR, Ross AG, Williams BK, Briceño CA. The Importance of Health Literacy in Addressing Eye Health and Eye Care Disparities. Ophthalmology. Elsevier; 2022;129:e137–45. https://doi.org/10.1016/j.ophtha.2022.06.034
23. Xompero C, Benettayeb W, Souied EH, Mehanna C-J. Pilot study evaluating the usability of MonŒil, a ChatGPT-based education tool in ophthalmology. AJO International [Internet]. Elsevier B.V.; 2024;1. https://doi.org/10.1016/j.ajoint.2024.100032
24. Wei M-Y, Li Y-L, Liu S-Y, Li G-Y. Evaluating the competence of large language models in ophthalmology clinical practice: a multi-scenario quantitative study. Front Cell Dev Biol [Internet]. Frontiers; 2025 [cited 2026 Mar 12];13. https://doi.org/10.3389/fcell.2025.1704762
25. Shi R, Liu S, Xu X, Ye Z, Yang J, Le Q, et al. Benchmarking four large language models' performance of addressing Chinese patients' inquiries about dry eye disease: A two-phase study. Heliyon [Internet]. Elsevier Ltd; 2024;10. https://doi.org/10.1016/j.heliyon.2024.e34391
26. Wu Y, Chen X, Zhang W, Liu S, Sum WMR, Wu X, et al. ChatMyopia: An AI agent for myopia-related consultation in primary eye care settings. iScience [Internet]. Elsevier Inc.; 2025;28. https://doi.org/10.1016/j.isci.2025.113768
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. https://doi.org/10.1136/bjo-2024-325599
28. Wei B, Yao L, Hu X, Hu Y, Rao J, Ji Y, et al. Evaluating the Effectiveness of Large Language Models in Providing Patient Education for Chinese Patients With Ocular Myasthenia Gravis: Mixed Methods Study. J Med Internet Res. 2025;27:e67883. https://doi.org/10.2196/67883
29. Sachdeva B, Ramjee P, Sharma R, Thulasidas M, Raveendra Murthy S, Fulari G, et al. Utility of an LLM-powered experts-in-the-loop chatbot for pre- and post-operative care of cataract surgery patients. Eur J Ophthalmol. SAGE Publications Ltd; 2025; https://doi.org/10.1177/11206721251396664
30. Özer Özcan Z, Doğan L, Yilmaz IE. Artificial Doctors: Performance of Chatbots as a Tool for Patient Education on Keratoconus. Eye Contact Lens. 2025;51:e112–6. https://doi.org/10.1097/ICL.0000000000001160
31. Esposito EP, Cardakli N, Christoff A, Kraus CL. Diagnostic Accuracy and Counseling Quality of GPT-4o for Strabismus and Pseudostrabismus in Patient-Generated Mobile Photographs: A Preliminary Evaluation. Clin Ophthalmol. Dove Medical Press Ltd; 2025;19:4077–84. https://doi.org/10.2147/OPTH.S556186
32. Wang X, Liu Y, Song L, Wen Y, Peng S, Ren R, et al. Transforming cataract care through artificial intelligence: an evaluation of large language models' performance in

addressing cataract-related queries. Frontier Artif Intell [Internet]. Frontiers Media SA; 2025;8. https://doi.org/10.3389/frai.2025.1639221

33. Hamzeh N, Lidder AK, Feder RS, Sarmiento EA, Mirza RG, Thau AJ, et al. Accuracy and Readability of Chat Generative Pre-Trained Transformer-4 Omni in Answering Ophthalmology Patient Questions. Ophthalmol Sci. Elsevier Inc.; 2026;6. https://doi.org/10.1016/j.xops.2025.101007

Figures

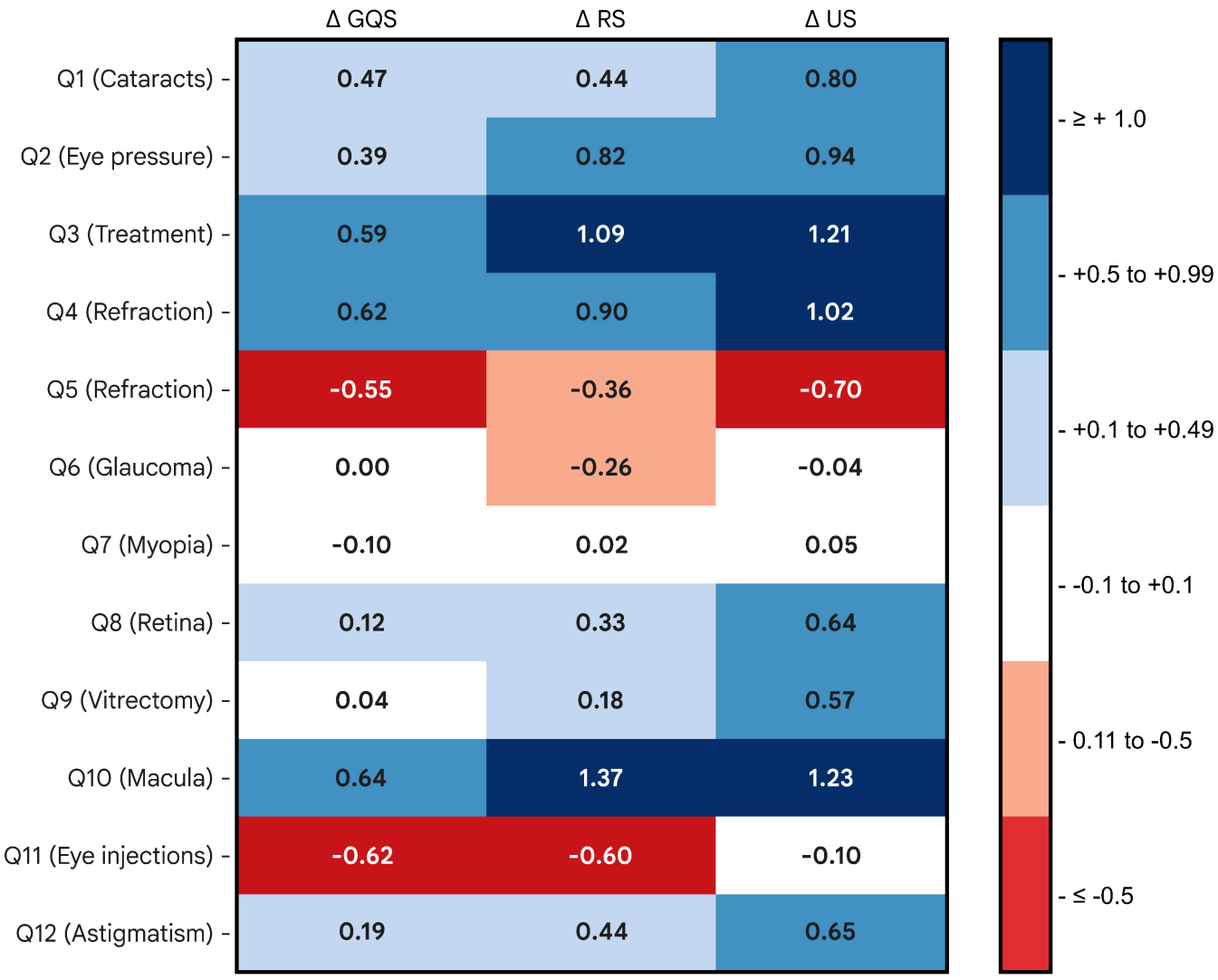


Figure 1

Heatmap of delta values (ChatGPT 5 – ChatGPT 3.5) by question and variable means

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile1.docx
- SupplementaryFile2.docx