The Medical Bulletin of Haseki

E-ISSN: 2147-2688

2025
Volume 63
Issue 3
June

www.hasekidergisi.com

The Medical Bulletin of Haseki

Editorial Board

Reviewing the articles' conformity to the publishing standards of the Journal, typesetting, reviewing and editing the manuscripts and abstracts in English, creating links to source data, and publishing process are realized by Galenos.

galenos yayınevi

Publisher Contact
Address: Molla Gurani Mah. Kacamak Sk. No: 21/1
34093 Istanbul, Türkiye
Phone: +90 (530) 177 30 97 / +90 (539) 307 32 03
E-mail: info@galenos.com.tr/yayin@galenos.com.tr
Web: www.galenos.com.tr
Publisher Certificate Number: 14521

Online Publishing Date: June 2025
E-ISSN: 2147-2688
International scientific journal published quarterly.

The Medical Bulletin of Haseki

Scientific Advisory Board

The Medical Bulletin of Haseki

Scientific Advisory Board

The Medical Bulletin of Haseki

Scientific Advisory Board

The Medical Bulletin of Haseki

Please refer to the journal's webpage **(https://www.hasekidergisi.com/)** for "About Us", "Instructions to Authors" and "Peer Review & Ethic".

The editorial and publication process of the Medical Bulletin of Haseki are shaped in accordance with the guidelines of ICMJE, WAME, CSE, COPE, EASE, and NISO. The journal is in conformity with the Principles of Transparency and Best Practice in Scholarly Publishing.

The Medical Bulletin of Haseki is indexed in **Emerging Sources Citation Index (ESCI), EBSCO Database, Gale, Turkish Medline-National Citation Index, Excerpta Medica/EMBASE, SCOPUS, TÜBİTAK/ULAKBİM, CINAHL, DOAJ, Hinari, GOALI, ARDI, OARE, AGORA, ProQuest, J-Gate, IdealOnline** and **Türkiye Citation Index**.

Owner: Mehmet Mesut Sonmez on University of Health Sciences Türkiye, İstanbul Haseki Training and Research Hospital

Responsible Manager: Akif ERBİN

The Medical Bulletin of Haseki

Contents

Original Articles

Original Article

DOI: 10.4274/haseki.galenos.2025.79553
Med Bull Haseki 2025;63(3):128-134

Harnessing GPT Technology for Clinical Decision Support in Retinal Detachment

Abdullah Agin*, Yucel Ozturk**, Ulviye Kivrak***

*University of Health Sciences Türkiye, Istanbul Haseki Training and Research Hospital, Clinic of Ophthalmology, Istanbul, Türkiye

**Istanbul Health and Technology University Faculty of Medicine, Department of Ophthalmology, Istanbul, Türkiye

***University of Health Sciences Türkiye, Kartal Dr. Lutfi Kirdar Training and Research Hospital, Clinic of Ophthalmology, Istanbul, Türkiye

Abstract

Aim: Considering the increasing incorporation of artificial intelligence (AI) in healthcare, it is crucial to comprehend the advantages and constraints of these technologies within ophthalmologic settings for their secure and efficient clinical utilization. This study aims to comprehensively assess the efficacy of three leading Generative Pre-trained Transformer (GPT) -based platforms in providing clinical decision-support for retinal detachment (RD).

Methods: This cross-sectional comparative study was conducted between April 2024 and May 2024. Fifty questions were created based on the American Academy of Ophthalmology "Retina Book", specifically targeting RD. The answers were produced by three different platforms and assessed by three independent reviewers who used Likert scales to evaluate their comprehensiveness and accuracy. Six readability metrics, including the Flesch-Kincaid Grade Level (FKGL) and Flesch Reading Ease Score (FRES), average words per sentence, average syllables per word, total sentence count, and total word count, were assessed.

Results: Gemini earned the most outstanding results for comprehensiveness (4.11±0.72) and accuracy (1.49±0.61), followed by ChatGPT and Copilot. ChatGPT had superior readability metrics, achieving an FKGL of 15.62±2.85 and a FRES of 62.54±12.34, establishing it as the most accessible platform. ChatGPT demonstrated significantly higher performance compared to other platforms in the metrics of average syllables per word ($p=0.0421$) and total word count ($p=0.0115$). At the same time, no significant differences were found among the platforms in the metrics of average words per sentence ($p=0.0842$) and total sentence count ($p=0.1603$). Intraclass correlation coefficient (ICC) values indicated strong inter-rater agreement for comprehensiveness (ICC >0.74) and moderate-to-high agreement for accuracy (ICC >0.56).

Conclusion: Gemini's detailed and accurate responses position it as a robust tool for professional use, while ChatGPT's superior readability makes it suitable for patient education. These findings emphasize the synergistic advantages of AI platforms in research and development management and show the necessity for hybrid systems that integrate accessibility with accuracy.

Keywords: Artificial intelligence, readability, ophthalmology, retina, retinal detachment

Introduction

Retinal detachment (RD) is an urgent condition in ophthalmology that can lead to vision loss if not treated appropriately (1). This disorder, characterized by the detachment of the neurosensory retina from the retinal pigment epithelium, requires prompt clinical and surgical care. Clinical decision-making in RD generally entails synthesizing intricate information and performing comprehensive assessments. The significance of artificial intelligence (AI)-supported Generative Pre-trained Transformer (GPT) platforms in delivering information and facilitating decision-making has attracted growing interest.

Comprehensiveness denotes the degree to which a platform delivers a thorough answer to a clinical prompt. Accuracy, simultaneously, relates to the scientific and clinical alignment of the response with information.

Corresponding Author: Abdullah Agin, MD, University of Health Sciences Türkiye, Istanbul Haseki Training and Research Hospital, Clinic of Ophthalmology, Istanbul, Türkiye
E-mail: abdullahagin@gmail.com **ORCID:** orcid.org/0000-0001-7173-8617
Received: 05.02.2025 **Accepted:** 19.06.2025 **Epub:** 12.08.2025 **Publication Date:** 29.08.2025

Cite this article as: Agin A, Ozturk Y, Kivrak U. Harnessing GPT technology for clinical decision-support in retinal detachment. Med Bull Haseki. 2025;63(3):128-134

The readability levels of the responses were evaluated using metrics like the Flesch-Kincaid Grade Level (FKGL) and the Flesch Reading Ease Score (FRES). Measuring these parameters enables a multifaceted understanding of each platform's capability to deliver information.

The application of AI, supported by GPT platforms, in the medical field emerges as a novel and dynamic area of interest in the literature (2,3). This study provides critical insights into the potential impact of contemporary technologies on clinical decision-support systems and evaluates the role of these platforms in the dissemination of medical knowledge. In this work, we conducted a comparative evaluation of the performance of GPT platforms concerning the urgent condition of RD, employing a methodologically updated approach (4,5). We hypothesized that the performance of GPT -based platforms would vary significantly in terms of accuracy, depth, and readability when applied to clinical questions related to RD.

Materials and Methods

This study was designed as a cross-sectional comparative analysis. This study seeks to evaluate the performance of three distinct GPT platforms [ChatGPT (GPT-4, OpenAI, accessed April 2024), Microsoft Copilot (powered by GPT-4, *via* Edge-Browser, accessed April 2024), and Google Gemini (Gemini Advanced, based on Gemini 1.5 Pro, accessed April 2024)] using 50 questions sourced from chapter 13 (RD and other RD) of the American Academy of Ophthalmology's "Retina and Vitreous" textbook (BCSC Section 12, 2022-2023) (6) (Supplementary Document 1). Sample prompts included: (1) "What are the main surgical indications for RD?" and (2) "How is rhegmatogenous RD differentiated from exudative RD?" All prompts were manually input using the default interface of each platform. Responses were generated without follow-up questions or user interactions. Each prompt was submitted independently, and responses were collected in their default format without editing.

Three vitreoretinal surgeons (AA, YO, UK) evaluated these questions in terms of their ability to provide comprehensive information on RD, to answer accurately, and to ensure readability. The study represents a significant step toward better understanding the potential use of these platforms in medical information delivery and patient care. The answers to these inquiries were obtained by soliciting the most comprehensive responses from three GPT platforms. Three separate evaluators assessed the thoroughness and precision of these responses according to the scoring standards outlined below; then, the mean scores were computed. No ethics review committee approval was required, as this study did not access protected patient information.

Comprehensiveness

The definitions of the Likert scores were as follows: (1=very incomprehensible or very dissimilar to physician response; 2=incomprehensible or dissimilar to physician response; 3=somewhat comprehensive or somewhat like a physician response; 4=comprehensive or similar to physician response; 5=very comprehensive).

Accuracy

-Two or "Very poor": responses contain at least two pieces of incorrect information. -One or "Poor": responses contain one piece of incorrect information. Zero: no response. Generative Pre-trained Transformer platforms responded to every prompt, resulting in no scores of 0. 1 or "Good": responses are medically accurate but incomplete. Two or "Very good": responses are medically accurate. Furthermore, six readability metrics were evaluated using an online application.

The parameters comprised the FKGL, FRES, average words per sentence, average syllables per word, total sentence count, and total word count. Outcomes were compared across the three GPT platforms.

Statistical Analysis

The statistical studies were conducted using Statistical Package for the Social Sciences version 25 (IBM, Chicago, IL, US). Descriptive statistics were calculated, including the mean, standard deviation, minimum, and maximum values for each metric, as well as the median. A One-Way Analysis of Variance (ANOVA) was performed to evaluate differences among the groups (ChatGPT, Microsoft Copilot, and Google Gemini) for each statistic. The post-hoc Tukey's Honestly Significant Difference Test was applied to identify group differences after significant disparities were found. The intraclass correlation coefficient (ICC) was calculated to assess inter-observer agreement on comprehensiveness and accuracy ratings among three raters for each platform. A two-way random effects model and single-rater consistency were used for the ICC calculations, with results shown alongside 95% confidence intervals. A p-value of less than 0.05 was accepted as statistically significant.

Results

Comprehensiveness and Accuracy

Table 1 displays the descriptive data for comprehensiveness and accuracy scores across the various platforms. Gemini earned the highest average comprehensiveness score of 4.11±0.715, whereas ChatGPT recorded a score of 3.61±0.795. In terms of accuracy, Gemini again led with 1.49±0.607, whereas Copilot scored the lowest average of 1.11±0.647. These results indicate statistically significant differences (ANOVA, F=7.653, p=0.00069 for comprehensiveness; F=5.993, p=0.0031 for accuracy). The post-hoc analysis (Table 1) showed that the difference was due to Gemini's better performance. Figure 1 illustrates the comparative scores for comprehensiveness and accuracy among the ChatGPT, Copilot, and Gemini platforms.

Readability and Word Metrics

Table 2 outlines the readability and lexical metrics for the platforms, including FKGL, FRES, average words per sentence, syllables per word, and the total count of sentences and words. ChatGPT gained superior readability scores (FKGL=15.62±2.85, FRES=62.54±12.34), exceeding those of Copilot and Gemini. Analysis of Variance revealed significant differences in FKGL (F=6.87, p=0.0012) and FRES (F=4.32, p=0.0168), with post-hoc tests demonstrating that ChatGPT exhibited superior readability scores compared to Gemini and Copilot. Although Gemini had marginally inferior readability metrics, it delivered more comprehensive responses, as evidenced by its word and sentence counts (23.4±8.1 sentences and 330.5±68.4 words). Figure 2 illustrates the readability metrics (FKGL and FRES) for the three platforms.

Inter-rater Reliability

The ICC values for comprehensiveness were 0.823 for ChatGPT, 0.856 for Copilot, and 0.741 for Gemini. The ICC scores for accuracy were 0.569 for ChatGPT, 0.782 for Copilot, and 0.745 for Gemini. Figure 3 shows the ICC for both comprehensiveness and accuracy, highlighting significant agreement among the raters.

Overall Performance

The thorough examination highlights Gemini's superiority in both comprehensiveness and precision, but ChatGPT wins in readability measures. The data collectively underscore the performance diversity among GPT platforms and indicate that Gemini may be more appropriate for applications necessitating precise and comprehensive medical information, especially for RD.

Discussion

Our comparative analysis revealed distinct strengths and weaknesses across GPT -based platforms when applied to RD -specific clinical questions. These differences (particularly the trade-off between factual depth and linguistic clarity) highlight practical considerations for platform selection based on user type (specialist vs. patient). The research evaluates these platforms on their comprehensiveness, accuracy, and readability, emphasizing their potential roles in clinical decision-support. Each platform's unique profile suggests context-dependent utility, for example, Gemini for clinical precision and ChatGPT for public communication. Large language models (LLMs) have begun to reshape ophthalmology workflows, especially in patient communication and rapid information retrieval. Recent studies have emphasized the efficacy of several LLMs, such as ChatGPT, Microsoft Copilot, and Google Gemini, in delivering precise and thorough solutions to clinical concerns.

Table 1. Comprehensiveness and accuracy scores

Metric	ChatGPT (Mean ± SD)	Copilot (Mean ± SD)	Gemini (Mean ± SD)	ANOVA F-value	ANOVA p-value	Post-hoc significant differences
Comprehensiveness	3.61±0.795	3.53±0.891	4.11±0.715	7.653	0.00069*	ChatGPT<Gemini Copilot<Gemini
Accuracy	1.40±0.481	1.11±0.647	1.49±0.607	5.993	0.0031*	Copilot<Gemini

*:Statistically significant
GPT: Generative Pre-trained Transformer, SD: Standard deviation, ANOVA: Analysis of Varience

Table 2. Readability and word metrics

Metric	ChatGPT (Mean ± SD)	Copilot (Mean ± SD)	Gemini (Mean ± SD)	ANOVA F-value	ANOVA p-value	Post-hoc significant differences
FKGL (Flesch-Kincaid)	15.62±2.85	14.13±3.12	13.87±2.90	6.87	0.0012*	ChatGPT>Gemini ChatGPT>Copilot
FRES (Flesch score)	62.54±12.34	59.18±14.32	57.49±13.87	4.32	0.0168*	ChatGPT>Gemini
Average words per sentence	14.67±3.21	13.48±3.12	13.21±3.05	2.51	0.0842	None
Average syllables per word	1.89±0.34	1.76±0.32	1.72±0.29	3.22	0.0421*	ChatGPT>Gemini ChatGPT>Copilot
Total sentence count	25.3±8.4	22.7±7.9	23.4±8.1	1.87	0.1603	None
Total word count	375.6±65.3	340.2±72.8	330.5±68.4	4.78	0.0115*	ChatGPT > Gemini, ChatGPT > Copilot

*:Statistically significant
FKGL: Flesch-Kincaid Grade Level, FRES: Flesch Reading Ease Score, GPT: Generative Pre-trained Transformer, SD: Standard deviation, ANOVA: Analysis of Varience

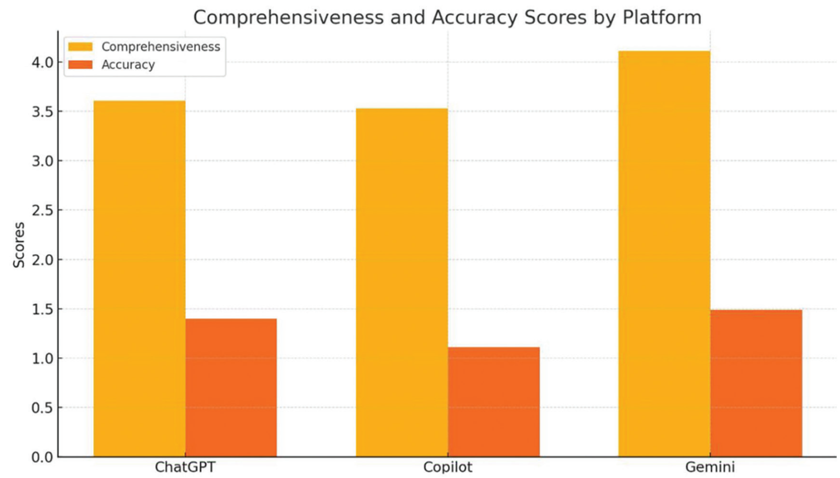

Figure 1. Comprehensiveness and accuracy scores by platform
Bar graph illustrating the mean comprehensiveness and accuracy scores assigned to responses generated by ChatGPT, Microsoft Copilot, and Google Gemini. Comprehensiveness was evaluated based on the breadth and depth of information, while accuracy reflected clinical correctness. Gemini demonstrated the highest comprehensiveness, while ChatGPT achieved relatively better accuracy than Copilot

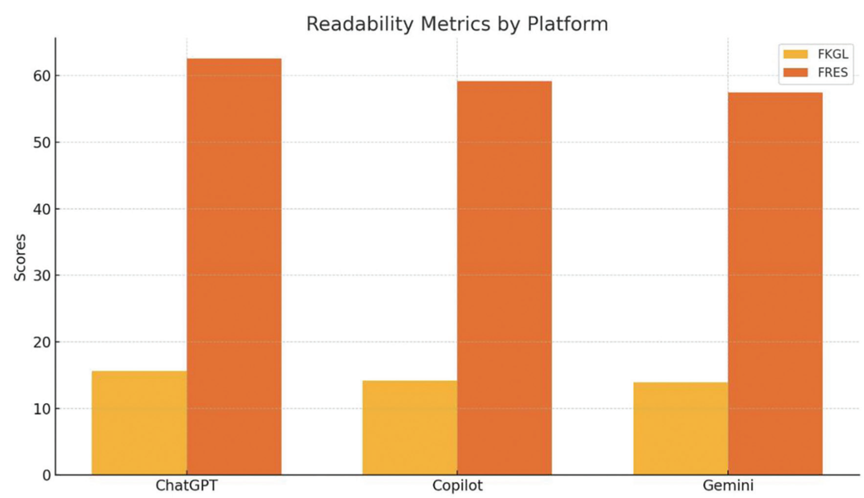

Figure 2. Readability metrics by platform
Bar graph comparing readability scores of responses from the three GPT -based platforms using Flesch-Kincaid Grade Level (FKGL) and Flesch Reading Ease Score (FRES). ChatGPT achieved the highest FRES (i.e., easiest to read), whereas Gemini's content was more complex linguistically, reflected in lower FRES and higher FKGL scores
GPT: Generative Pre-trained Transformer

The ramifications of these findings transcend simple performance measurements; they provoke essential inquiries regarding the application of LLMs in clinical environments. The implementation of AI-driven chatbots in patient triage has demonstrated potential. However, problems concerning the safe and effective adoption of measures, including ethical considerations, confidentiality, and physician responsibility, must be resolved (2,3). Moreover, the capacity of these models to aid in diagnosing disorders, as evidenced by research in neuro-ophthalmology and keratoconus, suggests their potential to enhance clinical practice, especially in regions with restricted access to specialists (3,7).

Furthermore, the efficacy of LLMs in educational settings, particularly in delivering preoperative information to patients undergoing ophthalmological procedures, has been examined. A comparison study showed that ChatGPT offers significantly more accurate responses than similar tools, highlighting its value as a reliable resource for patient education (4). This is especially important

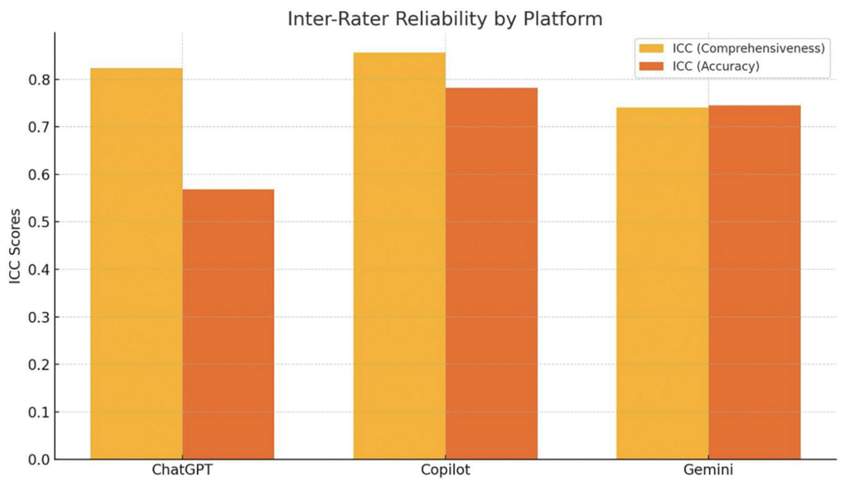

Figure 3. Inter-rater reliability by platform
Inter-rater reliability was measured using intraclass correlation coefficients for comprehensiveness and accuracy ratings across ChatGPT, Copilot, and Gemini. Intraclass correlation coefficient values >0.75 indicate good reliability. Copilot showed the highest agreement among reviewers for both parameters, while ChatGPT had strong agreement on comprehensiveness
ICC: Intraclass correlation coefficient

in ophthalmology, where a patient's grasp of intricate procedures can directly affect treatment results and compliance with medical recommendations.

The clarity of responses produced by these models is a crucial element that affects their efficacy. Research indicates that whereas specific models generate more precise information, they may also offer responses that are challenging for patients to comprehend (8). This underscores the imperative for continuous enhancement of LLMs to guarantee that they deliver correct information in a way that is comprehensible to a general audience. Readability measures like FKGL and FRES are essential tools that enable developers to customize responses for various patient populations. Alongside the technical capabilities of LLMs, the human aspect is crucial in the therapeutic environment. The capacity of AI to function as a reliable intermediary between physicians and patients has been investigated, although the details of these findings are not specified here. In contrast, LLMs can improve patient education, but they must not supplant the nuanced comprehension and empathy inherent in human clinicians (9). The equilibrium between utilizing AI for efficiency and preserving the human element in healthcare is essential for cultivating trust and guaranteeing patient satisfaction.

Gemini was identified as the most comprehensive platform, achieving an average score of 4.11, surpassing both ChatGPT and Copilot. This corroborates earlier research illustrating Gemini's efficacy in delivering thorough, contextually pertinent solutions, especially in medical fields necessitating accuracy (2,10). Its ability to incorporate nuanced details makes it a valuable tool for professionals requiring in-depth information. However, Gemini's high comprehensiveness often comes at the expense of readability, as observed in earlier evaluations of AI platforms in ophthalmology (8). Although Gemini's complexity benefits experts, it may pose challenges for lay users. For example, its use of advanced terminology and longer sentences may suit ophthalmologists preparing detailed reports but could overwhelm patients seeking simplified guidance. ChatGPT possibly reflects an emphasis on lower comprehensiveness (3.61). This is due to its emphasis on delivering more general responses. While this might seem like a limitation, it positions ChatGPT as a more efficient tool for scenarios requiring concise and user-friendly information (3). Copilot's lower performance across all metrics underscores its relative limitations in clinical contexts, reaffirming its secondary role compared to Gemini and ChatGPT.

Gemini also achieved the highest accuracy score (1.49), followed by ChatGPT (1.35) and Copilot (1.11). This is consistent with prior research indicating Gemini's ability to align responses with established clinical guidelines, particularly in specialized domains such as retinal conditions and glaucoma (4,5). Its exceptional precision renders it a dependable resource for healthcare experts. Nonetheless, in the context of thyroid eye disease, Gemini's comprehensive solutions may occasionally inundate users, especially patients or non-experts (11). This level of information is advantageous in professional contexts, highlighting the necessity of customizing responses according to the user's experience level. The high correlation between accuracy and comprehensiveness indicates that platforms excelling in one metric often perform well in the other. This relationship is particularly evident in Gemini's responses, which combine detail with adherence to clinical standards. ChatGPT, while slightly less accurate, compensates with its ability to simplify complex medical concepts, making it more accessible to non-specialists. ChatGPT outperformed its counterparts in readability, with the highest FRES (62.54), indicating easier readability, and FKGL (15.62), indicating more complex readability. However, this clashes with studies highlighting ChatGPT's capacity to simplify technical information without sacrificing essential details (12). Its user-friendly responses make it a preferred tool for patient education and public health communication. In contrast, Gemini scored lower on readability metrics, reflecting its focus on delivering detailed and technically precise information. This aligns with findings from evaluations of AI responses in refractive surgery and other ophthalmological contexts, where Gemini's advanced language constructs posed challenges for general comprehension (2,8). However, this trade-off, a key finding, could be more explicitly framed as a central consideration for potential users of these technologies. ChatGPT's superior readability makes it an ideal choice for patient-facing applications. Simplifying medical jargon and providing concise answers bridges the gap between complex medical information and patient understanding, which is a critical factor in improving health literacy (3). The study indicated that the average agreement among ophthalmologists for ChatGPT was 82.5% for both accuracy and comprehensiveness and 83.75% for clarity. Evaluations of Bard (on a prior version named Gemini) showed lower levels in agreement, with an average accuracy rate of 76.9%, comprehensiveness at 74.4%, and clarity reaching 83.8%. These findings suggest a strong consensus among evaluators, supporting the methodology and enhancing the reliability of the comparison results. Similar levels of agreement have been observed in studies comparing AI performance in ophthalmology, further strengthening the potential of these platforms as reliable decision-support tools (4,10). The distinct strengths of each platform highlight their complementary roles in clinical practice. Gemini's precision and comprehensiveness make it a valuable tool for specialists, particularly in drafting clinical reports or conducting in-depth analyses. On the other hand, ChatGPT's ability to simplify dense clinical content while maintaining factual reliability makes ChatGPT well-suited for patient education portals, informed consent preparation, and public health communication materials. The integration of AI platforms into clinical workflows aligns with broader trends in digital health, where AI is increasingly used for diagnostic support, patient communication, and personalized care (2,3). However, ethical concerns such as data security, bias in training datasets, and the potential for misinformation necessitate ongoing oversight and regulation (4).

The study's focus on RD-specific questions limits the generalizability of the findings to other ophthalmological conditions. Expanding the evaluation to include diverse subspecialties, such as cataract surgery or neuro-ophthalmology, could provide a more comprehensive understanding of these platforms' capabilities. Additionally, while Gemini excels in accuracy and detail, its limited readability highlights the need for hybrid models that combine its technical precision with ChatGPT's simplicity.

With the advancement of AI technologies in ophthalmology, it is crucial to undertake additional research to assess the long-term effects of these tools on clinical practice and patient outcomes. Research examining the comparative efficacy of various LLMs in diverse ophthalmological diseases would be essential for establishing optimal implementation procedures. Additionally, understanding the factors that influence patient adherence to treatment plans, particularly in the context of AI -assisted care, will be essential for improving therapeutic strategies.

The performance differences observed among the GPT -based platforms suggest that each may serve distinct roles depending on the clinical context. Gemini's structured, citation-rich responses indicate its potential utility in academic or professional settings such as assisting ophthalmology trainees or aiding in the drafting of consult letters where depth and precision are prioritized. In contrast, ChatGPT's balance between factual accuracy and linguistic simplicity makes it particularly suited for patient education, informed consent communication, and AI-powered triage systems. Meanwhile, Microsoft Copilot's inconsistent outputs, possibly due to integration constraints or model limitations, may currently hinder its reliability in scenarios requiring comprehensive clinical support. These findings underscore the importance of aligning platform selection with both the target user clinician versus patient and the cognitive complexity of the intended task. GPT based tools must be used with caution.

Study Limitations

Limitations include potential bias in training data, lack of individualized patient context, and the risk of misinformation. These models should augment rather than replace clinical judgment and must be used under physician oversight. While this study is limited to RD, the approach may be applicable to other ophthalmic or clinical domains. However, given the variability in complexity and terminology across subspecialties, further evaluations are warranted to assess generalizability. Despite these limitations, the study benefits from a robust comparative design and expert evaluation by subspecialists, which strengthens the validity and relevance of the findings.

Conclusion

This comparative analysis emphasizes the combined strengths of ChatGPT, Gemini, and Copilot in addressing RD-related questions. Gemini's accuracy and comprehensiveness make it a preferred choice for professional use, while ChatGPT's readability positions it as a valuable tool. Future research should explore adaptive AI systems capable of tailoring responses to the user's expertise level, ensuring both accessibility and accuracy. Longitudinal studies assessing the impact of these platforms on clinician workflows would provide valuable insights into their practical utility. Incorporating AI-generated solutions with human supervision is essential to reduce risks and guarantee the provision of high-quality treatment.

Ethics

Ethics Committee Approval: Ethics committee approval is not required, as the study does not involve the analysis of patient data.

Informed Consent: As the study does not include the use or analysis of patient data, neither informed consent is required.

Footnotes

Authorship Contributions

Concept: A.A., Y.O., U.K., Design: A.A., Data Collection or Processing: A.A., Y.O., U.K., Analysis or Interpretation: A.A., Literature Search: A.A., Writing: A.A.

Conflict of Interest: No conflicts of interest were declared by the authors.

Financial Disclosure: This study received no financial support.

References

1. D'Amico DJ. Clinical practice. Primary retinal detachment. N Engl J Med. 2008;359:2346-54.
2. Sabaner MC, Anguita R, Antaki F, et al. Opportunities and challenges of chatbots in ophthalmology: a narrative review. J Pers Med. 2024;14:1165.
3. David D, Zloto O, Katz G, et al. The use of artificial intelligence based chat bots in ophthalmology triage. Eye (Lond). 2025;39:785-9.
4. Patil NS, Huang R, Mihalache A, et al. The ability of artificial intelligence chatbots ChatGPT and Google bard to accurately convey preoperative information for patients undergoing ophthalmic surgeries. Retina. 2024;44:950-3.
5. 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.
6. American Academy of Ophthalmology. Basic and clinical science course (BCSC). Section 12: Retina and vitreous. Chapter 13: Retinal detachment and other retinal disorders. San Francisco (CA): American Academy of Ophthalmology; 2022.
7. Madadi Y, Delsoz M, Lao PA, et al. ChatGPT assisting diagnosis of neuro-ophthalmology diseases based on case reports. J Neuroophthalmol. 2024:2023.09.13.23295508.
8. Aydin FO, Aksoy BK, Ceylan A, et al. Readability and appropriateness of responses generated by ChatGPT 3.5, ChatGPT 4.0, Gemini, and Microsoft Copilot for FAQs in refractive surgery. Turk J Ophthalmol. 2024;54:313-7.
9. Güler MS, Baydemir EE. Evaluation of ChatGPT-4 responses to glaucoma patients' questions: can artificial intelligence become a trusted advisor between doctor and patient?. Clin Exp Ophthalmol. 2024;52:1016-9.
10. Bahir D, Zur O, Attal L, et al. Gemini AI vs. ChatGPT: a comprehensive examination alongside ophthalmology residents in medical knowledge. Graefes Arch Clin Exp Ophthalmol. 2025;263:527-36.
11. Bahir D, Hartstein M, Zloto O, Burkat C, Uddin J, Hamed Azzam S. Thyroid eye disease and artificial intelligence: a comparative study of ChatGPT-3.5, ChatGPT-4o, and Gemini in patient information delivery. Ophthalmic Plast Reconstr Surg. 2025;41:439-44.
12. Balci AS, Yazar Z, Ozturk BT, Altan C. Performance of Chatgpt in ophthalmology exam; human versus AI. Int Ophthalmol. 2024;44:413.

Supplementary Document 1: https://d2v96fxpocvxx.cloudfront.net/a2440bda-5c5c-4e7b-8a75-abf1691c9260/content-images/ccde405f-4b3f-4195-90fd-73ef0a0a587f.pdf