Clinical Ophthalmology

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Accuracy and Reproducibility of Different Artificial Intelligence Chatbots' Responses to Patient-Based Vitreoretinal Questions: A Comparative Study

Motasem Al-latayfeh[1,2], Abdelwahab Aleshawi[3], Omar S El-Mulki[4], Mohammed Baker[5], Zaina Qaddoumi[6], Dalia Attar[7], Lina Alma'aitah[8], Elaf Z Jarrah[5], Zainah Abu Khalil[5], Walaa Awad[3], Mo'men Raed Dayeh[3], Seren Al Beiruti[3], Rami Al-Dwairi[3]

[1]Department of Special Surgery, Faculty of Medicine, the Hashemite University, Zarqa, Jordan; [2]Department of Ophthalmology, Prince Hamza Hospital, Amman, Jordan; [3]Ophthalmology Division, Department of Special Surgery, Faculty of Medicine, Jordan University of Science and Technology, Irbid, Jordan; [4]Department of Ophthalmology, Bascom Palmer Eye Institute, University of Miami Miller School of Medicine, Miami, FL, USA; [5]Faculty of Medicine, Jordan University of Science and Technology, Irbid, Jordan; [6]Faculty of Science, University of Jordan, Amman, Jordan; [7]Faculty of Medicine, Hashemite University, Zarqa, Jordan; [8]Faculty of Pharmacy, Hashemite University, Zarqa, Jordan

Correspondence: Motasem Al-latayfeh, Department of Special Surgery, Faculty of Medicine, the Hashemite University, Zarqa, Jordan, Email Motasem974@gmail.com; Rami Al-Dwairi, Ophthalmology Division, Department of Special Surgery, Faculty of Medicine, Jordan University of Science and Technology, Irbid, Jordan, Email ramialdwairi@yahoo.com

Background: Generative artificial intelligence (AI) chatbots are increasingly used by patients and their reliability in complex ophthalmic conditions remains uncertain. This study aimed to compare the accuracy, comprehensiveness, and reproducibility of five AI chatbots—ChatGPT-5.o, DeepSeek R1, Meta AI, Grok 3.0, and Google Gemini 2.5 Pro—in responding to patient-centered vitreoretinal questions.

Methods: A total of 135 questions covering diabetic retinopathy, floaters/flashes, age-related macular degeneration, retinal tear/detachment, and vitrectomy were sourced from the American Academy of Ophthalmology "Ask an Ophthalmologist" database. Each question was submitted twice to each chatbot under standardized instructions. Two board-certified vitreoretinal ophthalmologists independently graded responses for accuracy and reproducibility. Accuracy was calculated as the proportion of responses graded "Correct and Comprehensive" or "Accurate but incomplete"; reproducibility was defined as agreement between the two responses.

Results: ChatGPT-5.o achieved the highest overall accuracy (94%, n=127/135, 95% CI: 89.9%–98.1%) with a reproducibility rate of 96.3% (n=130/135, 95% CI: 93.1%–99.5%). DeepSeek R1 demonstrated the greatest reproducibility (98.5%, n=133/135, 95% CI: 96.5%–100.0%) and high accuracy (92.6%, n=125/135, 95% CI: 88.1%–97.1%). Meta AI showed 91% (95% CI: 86.1%–95.9%) accuracy and 94% (95% CI: 89.9%–98.1%) reproducibility, whereas Grok 3.0 yielded the lowest accuracy (49.6%, n=67/135, 95% CI: 41.2%–58.0%) despite moderate reproducibility (88.1%, n=119/135, 95% CI: 82.7%–93.5%). Google Gemini 2.5 Pro recorded 72.6% (95% CI: 65.1%–80.1%) accuracy and the lowest reproducibility (77%, 95% CI: 69.9%–84.1%). By category, "Vitrectomy" scored the highest across all chatbots (94%, 95% CI: 87.2%–100.0%), followed by "Macular degeneration" (90%, 95% CI: 85.0%–95.0%). However, the category "Diabetic retinopathy" scored the lowest accuracy rate (64.7%, 95% CI: 52.1%–77.3%).

Conclusion: ChatGPT-5.o and DeepSeek R1 approached high accuracy and reproducibility comparable to clinical standards, indicating potential as patient-education tools in vitreoretinal care. However, variability across models and disease categories highlights the need for cautious clinical adoption and continued optimization to ensure safe, reliable information delivery.

Keywords: vitreoretinal surgery, diabetic retinopathy, artificial intelligence, large language models

https://doi.org/10.2147/OPTH.S580133
Received: 9 November 2025
Accepted: 4 February 2026
Published: 14 February 2026

Introduction

In the current digital era, there has been a drastic transformation in the way information is processed and accessed through emerging technologies, including generative artificial intelligence (AI) and large language models (LLMs). These advanced AI systems are trained on vast datasets using deep learning architectures, enabling them to understand and generate human-like text. They form the foundation of modern AI-powered chatbots such as Chat Generative Pre-Trained Transformer −5 (ChatGPT-5.o), Grok 3.0, DeepSeek R1, Google Gemini 2.5 Pro, and Meta AI.

Due to their accessibility, ease of use, and informative nature, these chatbots have recently witnessed a significant surge in demand and reliance across various sectors, including healthcare. An increasing number of individuals are turning to these tools to seek medical advice and healthcare information, particularly as access to traditional medical services becomes more challenging.[1] A study conducted in India reported that at least one in four patients seek medical advice on the web.[2] In addition, another nationwide survey in Poland found that almost two-thirds of participants used the Internet for searching for health information (64.9%).[3] This trend highlights the critical importance of evaluating the accuracy of information patients may obtain from chatbots without a physician's supervision.

AI chatbots have shown major advancements in various clinical and educational medical fields, although obvious variability in their accuracy. For example, researchers found that GPT-4V default mode demonstrated the highest detection rate (97.1%), outperforming its data analyst mode (61.8%) and Google Gemini 2.5 Pro (41.2%).[4] Despite the relatively high detection rates, the quality of diagnostic descriptions was generally suboptimal.[4] On the other hand, one study found that ChatGPT performed poorly in internal consistency and accuracy of the indications generated compared to clinical practice guideline recommendations for lumbosacral radicular pain.[5] Several studies showed that AI chatbots have a promising role in improving the medical education process, such as gross anatomy and neuroscience courses.[6–8] However, Harrison et al found that Claude and ChatGPT-4 outperformed other chatbots and had higher scores compared to students in neurosciences.[9] The findings of these studies demonstrate the importance of incorporating newly developed LLMs in several aspects of clinical and educational medicine, while taking care of the changes in their performance due to their continuous updates and new versions.

Vitreoretinal diseases encompass a spectrum of conditions affecting the retina and vitreous that are significant contributors to visual morbidity worldwide, making their epidemiology a critical area of investigation for public health and clinical care. Diabetic retinopathy (DR), one of the most common vitreoretinal disorders, affects roughly one-third of people with diabetes globally and is a leading cause of vision loss in working-age adults, with vision-threatening forms including proliferative retinopathy and diabetic macular edema (DME) affecting a substantial subset of these patients.[10] Population-based studies have estimated global DR prevalence at approximately 22–35% among individuals with diabetes, with projections indicating an increasing burden through 2045 as diabetes prevalence rises.[11] In the United States alone, it is estimated that more than 26% of people with diabetes have some form of retinopathy, translating to millions of affected individuals, with significant variation by demographics and geography.[12] The public health impact of these conditions is profound: DR and other vitreoretinal diseases contribute substantially to visual impairment and preventable blindness, affecting quality of life, independence, and economic productivity.[13] Beyond DR, other vitreoretinal conditions such as retinal vein occlusion, age-related macular degeneration, and rhegmatogenous retinal detachment contribute additional morbidity across diverse populations.[13] Therefore, assessing the reliability of LLMs as a source of information regarding this specific topic of ophthalmology is considered a public health priority.

Several studies have attempted to address this topic. Subramanian et al evaluated the appropriateness and completeness of ChatGPT-4 responses to 20 diabetic retinopathy-related queries framed from the patient's perspective, reporting average scores of 4.84 and 4.38 (out of 5), respectively.[2] Similarly, Strzalkowski et al conducted a comparative study assessing the accuracy and readability of ChatGPT-4 and Google Gemini using 13 retinal detachment-related questions categorized into three difficulty levels (D1–D3).[1] Both tools required college-level comprehension across all levels; Gemini provided easier readability, while ChatGPT-4 produced more correct responses to difficult questions with fewer serious errors, outperforming Gemini in 8 of 13 questions.[1] In another comparative study, Shean et al analyzed the performance of first-generation reasoning models, including DeepSeek's R1 and R1 Lite, OpenAI's o1 Pro, and Grok3 on 493 ophthalmology questions from StatPearls and EyeQuiz.[14] Among these, o1 Pro achieved the highest accuracy

(83.4%), outperforming DeepSeek R1 (72.5%), DeepSeek R1-Lite (76.5%), and Grok 3.0 (69.2%).[14] However, ChatGPT-5.o is different than ChatGPT-4, as it's more accurate, it's reasoning ability is tailored toward mission-critical situations, and has greater ability to handle long documents and searching the internet for references.[15]

Despite these findings, data on the accuracy and reproducibility of chatbots' generated responses to patient-based vitreoretinal questions remain limited. For example, Alqudah et al examined the performance of ChatGPT-3.0 on several aspects of ophthalmology, which showed moderate accuracy across all fields.[16] However, they reported that the percentage of comprehensive responses provided to questions related to retinal diseases was the lowest, with only 50% of questions being answered comprehensively, and 30.5% of questions being answered by responses graded as "Correct but incomplete.".[16]

Despite growing literature on AI in medical education and clinical decision support, significant gaps remain in understanding how different LLM platforms perform in specialized medical fields such as vitreoretinal ophthalmology. Furthermore, while previous studies have evaluated single AI models, comprehensive comparative analyses across multiple platforms using real patient questions are lacking. Therefore, this study aims to conduct a comparative evaluation of the accuracy, comprehensiveness, and reproducibility of responses generated by five different AI chatbots: ChatGPT-5.o, Google Gemini 2.5 Pro, Grok 3.0, Meta AI, and DeepSeek R1, using patient-centered questions sourced from the American Academy of Ophthalmology's (AAO) "Ask an Ophthalmologist" page. We hypothesize that these AI chatbots will show variability in terms of accuracy and reproducibility of their responses, and ChatGPT-5.o and DeepSeek R1 will show the highest accuracy rates. The ultimate goal is to assess how effectively these tools address specific patient inquiries in vitreoretinal ophthalmology and to inform their potential role in patient education and clinical support.

Methods

We wrote this paper according to the "Strengthening the Reporting of Observational Studies in Epidemiology" (STROBE) checklist.[17] Institutional review board approval was not required for this type of article.

AI Chatbots

We examined the performance of 5 common and "publicly accessible"AI chatbots, which are: ChatGPT-5.o, DeepSeek R1, Google Gemini 2.5 Pro, Grok 3.0, and Meta AI - Llama 4. The details of the used AI chatbots are described in previous studies (1,14,18).

ChatGPT-5.o and DeepSeek R1 were selected because they were found to be among the most accurate and reliable chatbots in many studies.[2,14] We selected Google Gemini 2.5 Pro, Grok 3.0, and Meta AI because these chatbots are incorporated in social media platforms such as "Google.com", "X.com", and "Massenger" (which is under the umbrella of Meta.com), which became part of the daily interaction of many patients despite their special academic or cultural interests, which increase the exposure and utility of these chatbots by social media users.

Question Curation/Data Source

One author collected 172 questions on vitreoretinal diseases from the "Ask an Ophthalmologist" Page on the AAO official website. The questions spanned five key areas: diabetic retinopathy, floaters and flashes, age-related macular degeneration, retinal tear/detachment, and vitrectomy. Inclusion was limited to English-language questions directly relevant to vitreoretinal conditions, while repeated questions, unclear items (like: "Could a sudden scare cause cloudy vision?"), general questions, or questions outside this scope were excluded. Thereafter, the set of questions was reviewed by a board-certified ophthalmologist to approve their clarity and scientific validity.

Response Generation

We have input the questions to the AI chatbots during the period from August 15th to August 21st, 2025. Each investigator was assigned one AI chatbot and entered each question separately. Each approved question was prompted to every chatbot twice, with each entry conducted on separate occasions using the chatbot's "new chat" function. The goal was to generate two independent responses per question to assess the reproducibility of AI-generated answers. Prior to each entry, the standardized prompt "Please answer the following question as if you were an ophthalmologist who is answering his patients, accurately and

concisely" was included to ensure consistency in framing. All responses were recorded in a structured Excel sheet under "First Response" and "Second Response" fields, and this procedure was applied uniformly across all AI chatbots to provide a systematic dataset for subsequent evaluation of accuracy and consistency.

Question Grading

All AI-generated responses were independently evaluated by two board-certified vitreoretinal ophthalmologists, with at least 3 years of experience in the vitreoretinal field, for accuracy and reproducibility. For the purpose of calculating the overall accuracy of each chatbot, the grading of the first response to each question was considered. Reproducibility was assessed by comparing the two responses generated for each question by the same chatbot.

The accuracy of each response was assessed using a four-level grading system:

1. Correct and Comprehensive: The response was fully accurate and thorough, providing all necessary information such that an ophthalmologist would not need to add anything further if a patient asked the same question.
2. Correct but incomplete: The information presented was correct, but important details were missing; an ophthalmologist would have additional key points to include.
3. Partially correct: The response contained a mixture of accurate and inaccurate information.
4. Incorrect: The response was entirely inaccurate and did not reflect appropriate clinical knowledge.

Any discrepancies in grading or reproducibility between the two board-certified ophthalmologist reviewers were resolved through a meeting without the involvement of a third-party adjudicator, during which a consensus was reached. The finalized grades were then compiled and used to evaluate the overall performance of the AI chatbots in answering patients' ophthalmology questions.

Statistical Analysis

Statistical analysis was performed using the IBM SPSS statistical package for Windows v.26 (Armonk, NY). Descriptive data were presented as frequencies and percentages. Also, the accuracy of different AI chatbots was calculated as a percentage of questions that scored category 1 or 2 (according to the accuracy scale) of the overall number of questions. To assess reproducibility, responses were divided into two categories: grades 1 and 2 formed the first group, while grades 3 and 4 formed the second group. For each question, the two responses were considered non-reproducible (ie, significantly different) if their assigned grades belonged to different groups. Fisher's exact test was used to calculate p-values for differences in correct responses between different chatbots. Cohen's kappa equation was utilized to obtain the reproducibility.

Results

Accuracy and Comprehensiveness

In total, 135 questions were input into the five AI chatbots after resolving 15 discrepancies between the two reviewers. The chatbot with the highest accuracy was ChatGPT-5.o, which answered accurately to 94% of the questions (n=127, 95% CI: 89.9%–98.1%), followed by DeepSeek R1 with 92.6% accuracy rate (n=125,95% CI: 88.1%–97.1%). The chatbot with the lowest accuracy was Grok 3.0, which answered accurately to 67 questions only (49.6%, 95% CI: 41.2%–58.0%). By category, the category "Vitrectomy" scored the highest across all chatbots (94%, 95% CI: 87.2%–100.0%), followed by "Macular degeneration" (90%, 95% CI: 85.0%–95.0%). However, the category "Diabetic retinopathy" scored the lowest accuracy rate (64.7%, 95% CI: 52.1%–77.3%). In regard to inaccurate responses, 5 responses of ChatGPT-5.o were graded as "Completely incorrect – Grade 4". Also, Grok 3.0 and Google Gemini 2.5 Pro mistakenly answered 12 and 8 questions, respectively. However, DeepSeek R1 and Meta AI chatbots did not answer any question with "Completely incorrect – Grade 4" responses. Table 1 presents the proportions of accurate responses answered by different AI chatbots.

Due to their superiority in terms of accuracy compared to other chatbots, we have chosen to compare ChatGPT-5.o with DeepSeek R1. Table 2 shows a comparison between ChatGPT-5.o and DeepSeek R1 in terms of the number of responses graded as "Correct and Comprehensive – Grade 1". The performance of ChatGPT-5.o outperformed DeepSeek

Table 1 Accuracy of Responses Generated by Different AI Chatbots to Questions Related to Vitreoretinal Disorders Categorized by Category

	ChatGPT-5.o				Grok 3.0				Google Gemini 2.5 Pro				DeepSeek R1				Meta AI				Accurate responses
Grading score	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	
Diabetic retinopathy (n=17)	7	2	6	2	5	4	6	2	2	7	5	3	11	6	0	0	3	8	6	0	55 (64.7%)
Floaters and flashes (n=10)	9	1	0	0	3	3	2	2	5	1	3	1	6	2	0	0	4	4	2	0	38 (76%)
Macular degeneration (n=30)	21	9	0	0	18	7	4	1	15	11	3	1	15	11	4	0	11	17	2	0	135 (90%)
Retinal tear / detachment (n=68)	52	16	0	0	39	14	8	7	28	19	18	3	48	16	4	0	45	21	2	0	298 (87.6%)
Vitrectomy (n=10)	8	2	0	0	5	2	3	0	7	3	0	0	5	5	0	0	3	7	0	0	47 (94%)
Overall accuracy	127 (94%)				67 (49.6%)				98 (72.6%)				125 (92.6%)				123 (91.1%)				

Abbreviations: ChatGPT, Chat Generative Pre-Trained Transformer; N, number.

Table 2 Comparison Between Accuracy of ChatGPT 5.o versus DeepSeek R1 in Terms of the Number of Responses Graded as "Correct and Comprehensive – Grade 1"

	ChatGPT 5.o	DeepSeek R1	P-value
Diabetic retinopathy (n=17)	7 (41.1%)	11 (64.7%)	0.31
Floaters and flashes (n=10)	9 (90%)	6 (60%)	0.32
Macular degeneration (n=30)	21 (70%)	15 (50%)	0.22
Retinal tear / detachment (n=68)	52 (76.4%)	48 (70.6%)	0.39
Vitrectomy (n=10)	8 (80%)	5 (50%)	0.32
Overall accuracy	97 (71.85%)	85 (63%)	0.15

Abbreviations: ChatGPT, Chat Generative Pre-Trained Transformer; N, number.

R1 in all categories except for "Diabetic retinopathy", with overall accuracy of 71.85% (n=97) for ChatGPT-5.o versus 63% (n=85) for DeepSeek R1. However, no significant differences were shown. On the other hand, when comparing overall accuracy between the two chatbots in terms of responses graded as "Correct and comprehensive – Grade 1" and "Correct but incomplete – Grade 2" combined, the performance of DeepSeek R1 was significantly higher than ChatGPT-5.o when answering questions in the "Diabetic retinopathy" category (p-value = 0.0027), as shown in Table 3.

Table 3 Comparison Between Accuracy of ChatGPT 5.o versus DeepSeek R1 in Terms of Number of Responses of Grades 1+2

	ChatGPT 5.o	DeepSeek R1	P-value
Diabetic retinopathy (n=17)	9 (53%)	17 (100%)	0.0027
Floaters and flashes (n=10)	10 (100%)	8 (80%)	0.47

(*Continued*)

https://doi.org/10.2147/OPTH.S580133

Table 3 (Continued).

	ChatGPT 5.o	DeepSeek R1	P-value
Macular degeneration (n=30)	30 (100%)	26 (86.6%)	0.11
Retinal tear / detachment (n=68)	68 (100%)	64 (94.1%)	0.12
Vitrectomy (n=10)	10 (100%)	10 (100%)	1.00
Overall accuracy	127 (94%)	125 (91.1%)	0.81

Abbreviations: ChatGPT, Chat Generative Pre-Trained Transformer; N, number.

Response Reproducibility

Table 4 shows the percentages of reproducible responses generated by each chatbot and the kappa value for the agreement. DeepSeek R1 generated the responses with the highest reproducibility rate (n=133, 98.5%), followed by ChatGPT-5.o, which answered 130 (96.3%) questions with reproducible answers. On the other hand, Google Gemini 2.5 Pro scored the lowest reproducibility rate (n=104, 77%). When looked by category, reproducibility was highest among questions in "Vitrectomy" (n=49, 98%), followed by "Flashes & Floaters" and "Retinal detachment" categories (n=320, 94%). However, responses to questions in the "Diabetic retinopathy" category scored the lowest reproducibility rate (n=66, 77.6%).

Discussion

This study evaluated the ability of five different AI chatbots to provide accurate and reproducible answers to patients' concerns using patient-written questions from the AAO. We found that ChatGPT-5.o showed the highest accuracy (n=127, 94%), while DeepSeek R1 provided the greatest reproducibility (n=133, 98.5%). In contrast, Grok 3.0 provided the lowest accuracy (n=67, 49.6%), and Google Gemini 2.5 Pro scored the lowest reproducibility rate (n=104, 77%). These results address the performance variability between chatbot models and across clinical categories, highlighting the need for cautious clinical integration.

Our findings are consistent with the existing literature. Balas et al designed 130 questions covering a broad spectrum of topics within 12 AAO Preferred Practice Pattern (PPP) domains of retinal diseases, in which they found that ChatGPT achieved an overall average score of 4.9/5.0, suggesting high alignment with the AAO PPP guidelines.[18] Furthermore, Subramanian et al compared the responses of ChatGPT-4 with three retinal experts in answering 20 questions related to diabetic retinopathy. They found an overall 96.8% agreement among the experts for appropriateness and 87.6% for completeness regarding AI-generated answers.[2] Moreover, Strzalkowski et al compared the response of ChatGPT-4 and

Table 4 Overall Reproducibility of Responses Generated by Different AI Chatbots Categorized by Question Category

	ChatGPT-5.o	Grok 3.0	Google Gemini 2.5 Pro	DeepSeek R1	Meta AI	Reproducible Responses
Diabetic retinopathy (n=17)	12	13	11	17	13	66 (77.6%)
Floaters and flashes (n=10)	10	10	8	10	9	47 (94%)
Macular degeneration (n=30)	30	24	21	28	28	131 (87.3%)
Retinal tear / detachment (n=68)	68	63	54	68	67	320 (94%)
Vitrectomy (n=10)	10	9	10	10	10	49 (98%)
Overall Reproducibility	130 (96.3%)	119 (88.1%)	104 (77%)	133 (98.5%)	127 (94%)	
Kappa value	0.526	0.760	0.143	0.881	0.469	

Abbreviations: ChatGPT, Chat Generative Pre-Trained Transformer; N, number.

Google Gemini with ten vitreoretinal specialists in answering 13 questions about retinal detachment; their findings showed that ChatGPT-4 provided more correct answers for the more difficult questions (p=0.0005) with fewer serious errors.[1]

In addition, DeepSeek R1 showed the highest reproducibility rate (n=133, 98.5%) with 92.6% accuracy rate (n=125). This is due to the use of Group Relative Policy Optimization (GRPO) by the model to enhance its reasoning and chain-of -thought capabilities. This approach is thought to reduce variability across outputs and ensures more stable performance across different runs, thereby improving reproducibility.[19] Xu et al collected 130 multiple-choice questions from the Chinese ophthalmology senior professional title examination, and they took Responses from DeepSeek-R1, Gemini 2.0 Pro, OpenAI o1, and o3-min. As a result, DeepSeek-R1 demonstrated the highest overall accuracy.[14]

On the other hand, we observed that Grok 3.0 had the lowest accuracy, which answered accurately to 67 questions only (49.6%), in addition to answering 12 (8.8%) questions with incorrect responses. However, it showed somewhat high reproducibility (n=119, 88.1%). Referring to existing literature, we found agreement with our results. Shean et al found that Grok 3.0 had the lowest accuracy (69.2%) compared with the other three chatbots examined in their study.[20]

Google Gemini 2.5 Pro demonstrated the lowest overall accuracy and reproducibility among the evaluated chatbots, with an accuracy of 72.6% and a reproducibility rate of 77%, and generated eight completely incorrect responses. These findings are consistent with the results of Strzalkowski et al, who found that Google Gemini 2.5 Pro generates more serious errors and lower thematic accuracy than ChatGPT-4, especially at higher difficulty levels (D3).[1] It also suggests that Gemini prioritizes readability, potentially at the expense of factual precision, which may explain the higher rate of critical errors under complex queries.[1] Taken together, these findings highlight Google Gemini2.5 Pro's limited reliability for patient education. On the other hand, Meta AI achieved 91% accuracy and 94% reproducibility without producing any completely incorrect responses, indicating that Meta AI offers greater stability than Gemini. However, its accuracy and reproducibility remain insufficient to match leading models such as ChatGPT-5.o and DeepSeek R1. Differences in training data coverage and underlying reasoning processes can be the reason behind the variation in performance among AI chatbots.[20]

Performance also varied across disease categories; the category "Vitrectomy" scored the highest in both accuracy and reproducibility rate (94%, 98% respectively), which is similar to what was found by Strzalkowski et al.[1] On the contrary, the "Diabetic retinopathy" category scored the lowest in both accuracy and reproducibility rate (64.7%, 77.6% respectively). These results are in contradiction to Subramanian et al's findings, which showed an overall 96.8% agreement among the experts for appropriateness and 87.6% for completeness regarding ChatGPT-4 answers.[2] This may be due to differences in the formulation of the questions. Subramanian et al's questions were formulated by a retina specialist to simulate patients' perspectives. On the other hand, we have collected questions that were asked by the patients themselves, which are more representative of real-world input. Although the ChatGPT-5.o model showed consistent results in Diabetic Retinopathy similar to the ChatGPT-4 model in previous literature, other AI models evaluated in this study may have shown variable rates in the same category. However, in this specific category, DeepSeek R1 showed a significantly higher accuracy compared to ChatGPT 5.o (100% vs 53%, p=0.0027). This finding may reflect a category-specific defect of ChatGPT 5.o ability in this complex topic, which warrants further investigations to elucidate such a hypothesis. Taken together, these findings suggest further optimization is needed before AI chatbots can be reliably used in heterogeneous management and nuanced decision-making.[16]

There has been a marked shift toward the use of AI chatbots as a source of health-related information since AI can offer personalized and potentially accurate answers regarding health concerns. In addition, it gives direct and comprehensive responses in contrast to the overwhelming and contradictory information that conventional search engines may offer for the general population. Also, it is free of charge, which makes it more convenient and affordable for patients to use.[16] As the reliance on AI chatbots increases, the importance of evaluating chatbots' accuracy and reproducibility increases to ensure the safety of information and support their role in patient education.

Artificial intelligence is of great importance in many fields, including medicine and healthcare. Many studies are examining the effectiveness of AI in answering questions, providing information, diagnosing, and treating diseases. Nevertheless, we must maintain ethics, privacy, and the anonymity of patients during such research to ensure rights, deliver better results, and increase patient confidence in the provision of information. Implications of our findings include cautious use of different AI chatbots to avoid wrong answers regarding medical questions, which may be used harmfully by patients with vitreoretinal disorders in the wrong way. Our study highlights the significant potential of ChatGPT-5.o

and DeepSeek R1 in providing accurate and reliable answers, making them good choices for health information in the field of ophthalmology. Another implication of our results is that trainee resident physicians can use ChatGPT-5.o and DeepSeek R1 to aid their training process and help them respond to patients' queries regarding this sensitive branch of ophthalmology, which will indeed reduce the rate of false answers and decrease the threshold for health risks due to wrong answers in patients' counseling.

In the ethical aspects, we should emphasize that these AI chatbots cannot be used solely as the source of health information by patients or their physicians. However, these AI chatbots can be used as an adjunct in the healthcare services, as the physicians are the only ones responsible for the safety of their patients in legal terms.

Our study has some limitations. Firstly, we do not know the sources that such chatbots use to answer questions or provide information. Secondly, we investigated a small unequal set of questions that may influence the results. In addition, prompt engineering may not reflect actual patient interactions. Also, we used English-only questions. Furthermore, both evaluators were from the same institution. Moreover, we did not include data on the patient perspective on comprehension or satisfaction. In addition, the evaluators were not blinded to which chatbot generated responses which may introduce bias. Lastly, the results cannot be readily extrapolated to all countries, given that the medical standards adopted by national associations differ substantially.

We have used the latest versions of the selected chatbots in our study. However, the quality of chatbots' responses and their reliability may change due to continuous updating of these models. Also, although we have used a specific prompting while inputting the questions into different chatbots, bias and differences in response generation may occur due to interpersonal differences in handling these models. These factors may affect the reliability of our results over different points in time.

Conclusion

ChatGPT-5.o and DeepSeek R1 approached expert-level accuracy and reproducibility, indicating potential as patient-education tools in vitreoretinal care. However, variability across models and disease categories highlights the need for cautious clinical adoption and continued optimization to ensure safe, reliable information delivery.

Data Sharing Statement

The datasets generated and analyzed during the current study are available from the corresponding author (Dr Rami Al-Dwairi).

Funding

The authors have not declared any grant for this work from any funding authority.

Disclosure

The authors have no financial ties or conflicts of interest to disclose.

References

1. Strzalkowski P, Strzalkowska A, Chhablani J, et al. Evaluation of the accuracy and readability of ChatGPT-4 and Google Gemini in providing information on retinal detachment: a multicenter expert comparative study. *Int J Retin Vitr*. 2024;10(1):61. doi:10.1186/s40942-024-00579-9
2. Subramanian B, Rajalakshmi R, Sivaprasad S, Rao C, Raman R. Assessing the appropriateness and completeness of ChatGPT-4's AI-generated responses for queries related to diabetic retinopathy. *Indian J Ophthalmol*. 2024;72(Suppl 4):S684–S687. doi:10.4103/IJO.IJO_2510_23
3. Płaciszewski K, Wierzba W, Ostrowski J, Pinkas J, Jankowski M. Use of the internet for health purposes—a national web-based cross-sectional survey among adults in Poland. *Int J Environ Res Public Health*. 2022;19(23):16315. doi:10.3390/ijerph192316315
4. Srinivasan S, Ji H, Chen DZ, et al. Can off-the-shelf visual large language models detect and diagnose ocular diseases from retinal photographs? *BMJ Open Ophthalmol*. 2025;10(1):1–7. doi:10.1136/bmjophth-2024-002076
5. Gianola S, Bargeri S, Castellini G, et al. Performance of ChatGPT compared to clinical practice guidelines in making informed decisions for lumbosacral radicular pain: a cross-sectional study. *J Orthop Sports Phys Ther*. 2024;54(3):222–228. doi:10.2519/jospt.2024.12151
6. Bolgova O, Shypilova I, Sankova L, Mavrych V. How Well Did ChatGPT Perform in Answering Questions on Different Topics in Gross Anatomy? *Eur J Med Heal Sci*. 2023;5(6):94–100. doi:10.24018/ejmed.2023.5.6.1989
7. Mavrych V, Ganguly P, Bolgova O. Using large language models (ChatGPT, Copilot, PaLM, Bard, and Gemini) in gross anatomy course: comparative analysis. *Clin. Anat*. 2025;38(2):200–210. doi:10.1002/ca.24244
8. Mavrych V, Yaqinuddin A, Bolgova O. Claude, ChatGPT, Copilot, and Gemini performance versus students in different topics of neuroscience. *Adv Physiol Educ*. 2025;49(2):430–437. doi:10.1152/advan.00093.2024

9. Harrison DW. *Brain Asymmetry and Neural Systems: Foundations in Clinical Neuroscience and Neuropsychology*. Springer; 2015. doi:10.1007/978-3-319-13069-9
10. Lee R, Wong TY, Sabanayagam C. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye Vis*. 2015;1–25. doi:10.1186/s40662-015-0026-2
11. Teo ZL, Tham Y-C, Yu M, et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. *Ophthalmology*. 2021;128(11):1580–1591. doi:10.1016/j.ophtha.2021.04.027
12. Lundeen EA, et al. Prevalence of diabetic retinopathy in the US in 2021. *JAMA Ophthalmol*. 2023;30341(8):747–754. doi:10.1001/jamaophthalmol.2023.2289
13. Zhou C, Li S, Ye L, et al. Visual impairment and blindness caused by retinal diseases: a nationwide register-based study. *J Glob Health*. 2023;13:4126. doi:10.7189/jogh.13.04126
14. Shean R, Shah T, Pandiarajan A, et al. A comparative analysis of DeepSeek R1, DeepSeek-R1-Lite, OpenAi o1 Pro, and Grok 3 performance on ophthalmology board-style questions. *Sci Rep*. 2025;15(1):23101. doi:10.1038/s41598-025-08601-2
15. Baker M, et al. Ocular manifestations in congenital insensitivity to pain with anhidrosis: a window into a rare syndrome. *Vision*. 2025;9(3)16.
16. Alqudah AA, Aleshawi AJ, Baker M, et al. Evaluating accuracy and reproducibility of ChatGPT responses to patient-based questions in Ophthalmology: an observational study. *Medicine*. 2024;103(32):e39120. doi:10.1097/MD.0000000000039120
17. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Int J Surg*. 2014;12(12):1495–1499. doi:10.1016/j.ijsu.2014.07.013
18. Balas M, Mandelcorn ED, Yan P, Ing EB, Crawford SA, Arjmand P. ChatGPT and retinal disease: a cross-sectional study on AI comprehension of clinical guidelines. *Can J Ophthalmol*. 2025;60(1):e117–e123. doi:10.1016/j.jcjo.2024.06.001
19. Xu P, Wu Y, Jin K, Chen X, He M, Shi D. DeepSeek-R1 outperforms Gemini 2.0 Pro, OpenAI o1, and o3-mini in bilingual complex ophthalmology reasoning. *Adv Ophthalmol Pract Res*. 2025;5(3):189–195. doi:10.1016/j.aopr.2025.05.001
20. Bellanda VCF, Santos MLD, Ferraz DA, Jorge R, Melo GB. Applications of ChatGPT in the diagnosis, management, education, and research of retinal diseases: a scoping review. *Int J Retin Vitr*. 2024;10(1):79. doi:10.1186/s40942-024-00595-9