DOI: 10.14744/eer.2025.27247
Eur Eye Res 2025;5(2):95–102

EUROPEAN EYE RESEARCH

ORIGINAL ARTICLE

Can artificial intelligence become a board-certified ophthalmologist? Assessing machine intelligence in ophthalmology education

Pelin Kiyat, Hazan Gul Kahraman

Department of Ophthalmology, İzmir Democracy University, Buca Seyfi Demirsoy Training and Research Hospital, İzmir, Türkiye

Abstract

Purpose: To evaluate and compare the performance of three leading artificial intelligence (AI) models (ChatGPT 4o, ChatGPT o1, Claude 3.5 Sonnet, and Gemini 2.0 Flash Experimental) in answering ophthalmology questions from two different, popular board preparation question resources and to analyze performance variations across subspecialties and resources.

Methods: From the 398 available questions in the ebodtraining.com question bank, 344 text-based questions were selected and organized to include 35 questions per subspecialty. The same number of questions per subspecialty was randomly selected from eyedocs.co.uk to match those from ebodtraining.com. ChatGPT 4o, ChatGPT o1, Claude 3.5 Sonnet, and Gemini were tested on these questions, with responses evaluated as either correct or wrong, allowing calculation of both overall and subspecialty-specific performance metrics.

Results: Various AI models were evaluated on two ophthalmology question banks: Ebodtraining.com (344 questions) and eyedocs.co.uk (345 questions). For ebodtraining.com, ChatGPT o1 achieved 88.0% accuracy, followed by Claude 3.5 Sonnet (84.7%), Gemini (81.7%), and ChatGPT 4o (81.2%), with all models showing weaker performance in the Neuro-ophthalmology section. Similarly, on eyedocs.co.uk, ChatGPT o1 led with 88.4%, while Claude 3.5 Sonnet reached 84.6%, Gemini 79.2%, and ChatGPT 4o 73.4%. ChatGPT o1 significantly outperformed ChatGPT 4o on both platforms and demonstrated higher accuracy across multiple subspecialties compared to Claude 3.5 Sonnet and Gemini.

Conclusion: In the modern world, time is getting more precious every day and with the help of AI models, students can receive information and explanations rapidly. In addition, with the advantage of asking further questions, students can access personalized answers, reduce time consumption, and get a tailored learning experience. However, it should be taken into consideration that although AI models demonstrate promising capabilities in ophthalmology board examination preparation, their performance varies significantly across subspecialties and question types. These tools can serve as valuable supplementary resources for exam preparation, but cannot replace comprehensive clinical training and expertise

Keywords: Artificial intelligence; ChatGPT; ophthalmology board examinations.

Cite this article as: Kiyat P, Kahraman HZ. Can artificial intelligence become a board-certified ophthalmologist? Assessing machine intelligence in ophthalmology education. Eur Eye Res 2025;5(2):95–102.

Correspondence: Pelin Kiyat, M.D. Department of Ophthalmology, İzmir Democracy University, Buca Seyfi Demirsoy Training and Research Hospital, İzmir, Türkiye
E-mail: pelinkiyat@hotmail.com

Submitted Date: 05.01.2025 **Revised Date:** 22.02.2025 **Accepted Date:** 24.03.2025 **Available Online Date:** 26.08.2025

Generative pre-trained transformer (GPT) series have been generated interest and excitement among various areas of science, including medicine. ChatGPT is an artificial intelligence (AI) chatbot platform developed by OpenAI (San Francisco, CA, USA) that consists of a conversation-based technology that allows users to receive appropriate responses as texts for their questions.[1] The program had several version updates since its inception and GPT 4o version was presented on May 2024 and the novel GPT o1 on December 2024. This latter version was defined as a more robust, concise, and intelligent model.[2] Due to the fact that GPT models, including ChatGPT, have been programmed on a textual database and have the ability to provide coherent and contextually appropriate responses,[3] they have been assumed to have promising potential for educational purposes in medicine.

Alongside ChatGPT, other novel AI language models have emerged as potential educational tools. Claude 3.5 Sonnet, developed by Anthropic and released in June 2024, has demonstrated promising capabilities in medical reasoning and problem-solving tasks.[4] Google's Gemini 2.0 Flash Experimental introduced in December 2024 as the successor to Bard, represents another significant advancement in AI technology. Gemini's multimodal capabilities and extensive training on scientific literature make it a promising tool for medical education.[5]

In our country, the examinations provided by internationally recognized qualifications, such as the European Board of Ophthalmology (EBO) and the International Council of Ophthalmology (ICO), have gained significant popularity among young ophthalmologists, especially among the ophthalmology residents recently. An important component of ophthalmology examinations and residents' education determination involves standardized multiple-choice questions.[3] In terms of preparing for these examinations, question banks about ophthalmology subspecialties are generally preferred among candidates as they resemble to the real examination format and content. Ophthalmologists who aim to succeed in these exams utilize different question banks and helpful websites, including https://www.eyedocs.co.uk/[6] and https://ebodtraining.com/[7] to practice and get prepared. One of the most popular question bank websites, https://www.eyedocs.co.uk/ is a useful source to get prepared to the ICO exams, especially due to the case-based multiple choice question format. On the other hand, https://ebodtraining.com/ can be considered more helpful in getting prepared to the EBO examinations due to the specific question format, including True/False-based questions. Both resources are accepted as effective and safe in getting prepared for the ophthalmology exams, including the EBO and ICO board exams. In https://ebodtraining.com/ web site, it is mentioned that the contents and educational resources have been created by an independent scientific committee composed by renowned experts with the aim of helping ophthalmologists achieve clinical excellence Furthermore, a statement is declared by the web site "We use AI-based functionalities and the latest technology available in the market to test your abilities through practical activities that aim to simulate the comprehensive EBO exam and other official tests methodology."[6] In addition, in https://www.eyedocs.co.uk/ web site shows featuring with the ICO.[7]

The aim of this present study is to evaluate and compare the performance of three leading AI models (ChatGPT o1, ChatGPT 4o, Claude 3.5 Sonnet, and Gemini 2.0 Flash Experimental) in answering ophthalmology questions from two different, popular board preparation question resources and to analyze performance variations across subspecialties and resources. As far as we know, while there are studies on other question banks, there is no study yet with these question banks, and comparing the performance of these AI models.

Materials and Methods

Multiple-choice questions from two common and popular resources for board certification examination preparation, https://www.eyedocs.co.uk/ and https://ebodtraining.com/, were used to assess the performance of ChatGPT o1, ChatGPT 4o, Claude 3.5 Sonnet, and Gemini 2.0 Flash Experimental. Text-based multiple-choice questions were included in the study. Questions were omitted from analysis if they included an image or table due to the lack of some AI models' ability of processing them.

From the https://ebodtraining.com/ platform, 344 questions (86.4%) were selected out of 398 available questions. These questions were distributed to include 35 questions for each subspecialty, while the "Optics and refraction" section contained 31 questions. All questions followed a multiple-choice format where each option required a True or False response. An equal number of multiple-choice questions were also randomly selected from https://www.eyedocs.co.uk/, matching the distribution of 35 questions per subspecialty and 30 questions in the "Optics" section.

The subspecialties were "Retina, vitreous and uvea," "Pediatric ophthalmology and strabismus,""External, corneal and adnexial diseases," "Glaucoma, cataract and refractive surgery," "Optics and refraction," "Neuro-ophthalmology," "Orbital disease and Oculoplastic surgery,""General medicine relevant to ophthalmology," "Ophthalmic pathology, microbiology," and "Pharmacology and therapeutics" in the https://ebodtraining.com/ question bank.

The subspecialties were "Retina and uveitis," "Pediatrics and strabismus," "External eye and cornea," "Glaucoma and cataract," "Optics," "Neurology and pupils," "Orbit and oculoplastics" "General medicine," "Pathology and microbiology," and "Pharmacology and therapeutics" in the https://www.eyedocs.co.uk/ question bank.

It is well known that the style and prompt of questions have an important impact on AI models' performance. To receive standardized answers, the process of asking questions was standardized across all models (ChatGPT o1, ChatGPT 4o, Claude 3.5 Sonnet, and Gemini 2.0 Flash Experimental). Questions were formatted in Microsoft Word following Gilson et al.'s[8] procedure – the stem in a paragraph, multiple choice options on separate lines, with two empty lines between the stem and choices. New accounts were created for each AI model to prevent conversation history bias. Each model's conversation history was cleared before new questions to avoid sequential influence. One researcher (H.G.K.) performed all question inputs consistently across models. Responses were manually reviewed and recorded as correct/wrong, with overall percentages and subspecialty-specific calculations for each model. Official answers from the question banks' websites determined correctness.

Since this study involved no human participants, institutional review board approval wasn't necessary.

The primary outcome measured how accurately ChatGPT o1, ChatGPT 4o, Claude 3.5 Sonnet, and Gemini 2.0 Flash Experimental answered ophthalmology board-style questions across all subspecialties. Secondary outcome focused on comparing these models' performance within individual subspecialties. The third outcome was the performance difference between ChatGPT versions (o1 and 4o), while the fourth outcome compared the capabilities of the latest AI models - ChatGPT 4o, Claude 3.5 Sonnet, and Gemini 2.0 Flash Experimental.

Statistical Analysis

"IBM The Statistical Package for the Social Sciences version 25" (SPSS Inc., Chicago, IL, USA) was used for statistical purposes. Categorical variables were expressed as frequency and percentage, and numeric variables as mean and standard deviation. Kolmogorov-Smirnov tests were used to determine whether the data were normally distributed. The correct response rates of each chatbot were calculated. To compare the correct response rates between two chatbots, an Independent T-Test was performed to determine the differences in the normality of the distribution, or Mann–Whitney U test was performed to determine differences in non-normal distribution. Analysis of variance (ANOVA) testing was performed to evaluate differences in normally distributed data to evaluate the performance differences of the three chatbots, and Kruskal–Wallis test was applied when the distribution was non-normal. When the ANOVA test gave significant results, pairwise comparisons were performed with the post hoc Dunn test or Mann–Whitney U test to determine which chatbots had differences between the groups. A P-value under 0.05 was considered statistically significant.

Results

Among the 344 text-based multiple choice questions with 1730 True/False options in https://ebodtraining.com/ question bank, ChatGPT o1 answered 1524 (88.0%), ChatGPT 4o answered 1407 (81.2%) questions correctly, Claude 3.5 Sonnet demonstrated 1468 (84.7%) correct answers, while Gemini 2.0 Flash Experimental achieved 1416 (81.7%) correct responses.

With respect to subspecialty performance in the https://ebodtraining.com/ question bank, all models showed similar patterns of strength and weakness. The lowest performance was consistently observed in "Neuro-ophthalmology" (ChatGPT 4o: 68.57%, ChatGPT o1: 74.85%, Claude 3.5 Sonnet: 76.57%, Gemini: 69.71%). Conversely, all models excelled in "General medicine relevant to ophthalmology" (ChatGPT 4o: 95.43%, ChatGPT o1:96.0%, Claude 3.5 Sonnet: 93.71%, Gemini: 92.57%) and "Ophthalmic pathology, microbiology" sections (ChatGPT 4o: 91.43%, ChatGPT o1:97.14%, Claude 3.5 Sonnet: 93.14%, Gemini: 92.57%) (Fig. 1).

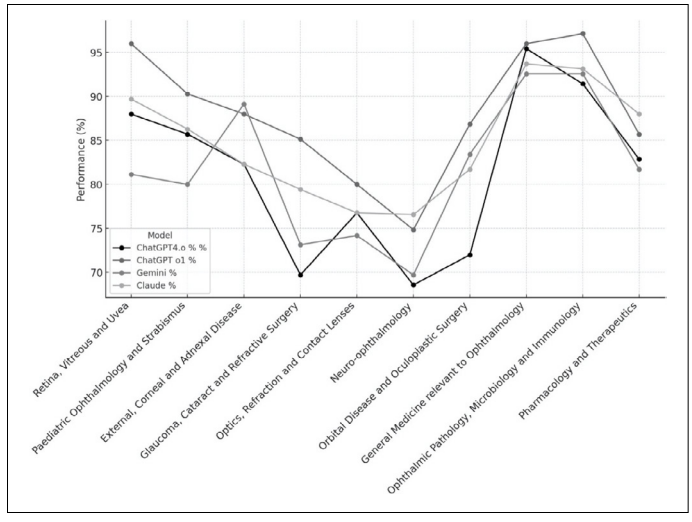

Fig. 1. Performance comparison of artificial intelligence models (ChatGPT 4o, ChatGPT o1, Claude 3.5 Sonnet, and Gemini 2.0 Flash Experimental) Across Subspecialties in ebodtraining.com Question Bank.

According to the analysis performed on https://www.eyedocs.co.uk/ question bank, AI models showed different levels of accuracy in answering the 345 multiple-choice questions. ChatGPT 4o correctly answered 253 questions (73.4%), ChatGPT o1 showed the highest accuracy with 304 correct answers (88.4%). Claude 3.5 Sonnet correctly answered 292 questions (84.6%) and Gemini 2.0 Flash Experimental achieved 274 correct answers (79.2%).

The analysis of AI models' performance regarding to the subspecialties in the https://www.eyedocs.co.uk/ question bank revealed variations contrasting with the more consistent rankings observed in the former question bank. All AI models achieved their highest success rates in "Pathology and microbiology" section, with ChatGPT o1 leading at 97.14%, followed by ChatGPT 4o and Claude 3.5 Sonnet both at 94.29%, and Gemini at 91.43%. However, each model showed different weaknesses in specific subspecialties: ChatGPT 4o performed weakest in "Glaucoma and cataract" questions, ChatGPT o1 was weakest in "Pharmacology and therapeutics." Claude 3.5 Sonnet had its lowest scores in Orbit and oculoplastics,' while Gemini showed the weakest performance in "Pediatrics and strabismus." (Fig. 2)

The comparative analysis revealed that ChatGPT o1 achieved higher scores than ChatGPT 4o in all subspecialties in https://ebodtraining.com/ question bank, with statistically significant differences in four sections: "Retina, vitreous and uvea" (p=0.01), "Glaucoma, cataract and refractive Surgery" (p<0.001), "Orbital disease and oculoplastic surgery" (p<0.001), and "Ophthalmic pathology, microbiology" (p=0.04) (Table 1).

In addition, when comparing ChatGPT o1 with other AI models, it showed significantly better performance than both Claude 3.5 Sonnet and Gemini 2.0 Flash Experimental in

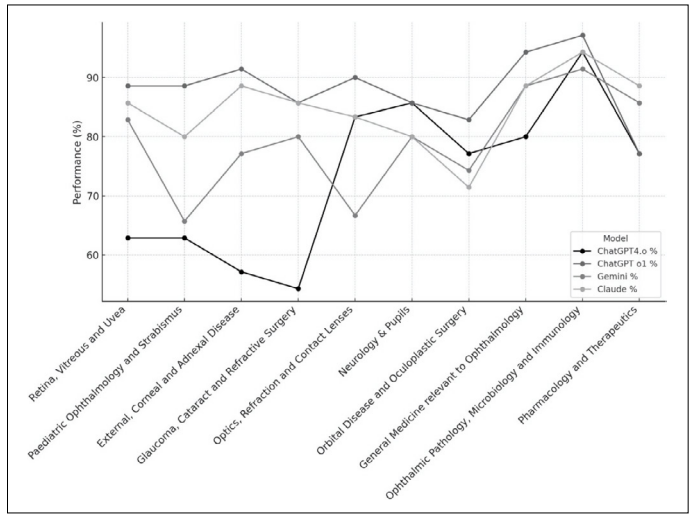

Fig. 2. Performance comparison of artificial intelligence models (ChatGPT 4o, ChatGPT o1, Claude 3.5 Sonnet, and Gemini 2.0 Flash Experimental) Across Subspecialties in eyedocs.co.uk/ Question Bank.

three subspecialties in https://ebodtraining.com/ question bank: "Retina, vitreous and uvea" (p<0.01), "Pediatric ophthalmology and strabismus" (p=0.02), and "Glaucoma, cataract and refractive surgery" (p=0.02) (Table 2).

ChatGPT's latest version (o1) presented substantial improvement in performance compared to its previous version (4o) in https://www.eyedocs.co.uk/ question bank. In categories where ChatGPT 4o achieved <70% accuracy showed remarkable enhancement, with ChatGPT o1 reaching rates above 85%. This improvement was found to be statistically significant for "Retina and uveitis" (p=0.03), "Pediatrics and strabismus" (p=0.03), External eye and cornea' (p<0.001), and Glaucoma and cataract' (p=0.01) (Table 3).

The comparison of newer models and ChatGPT o1 in answering https://www.eyedocs.co.uk/ questions, ChatGPT

Table 1. Performance Comparison of ChatGPT 4.o and ChatGPT o1 Across Subspecialties in ebodtraining.com Question Bank

Subspecialty	ChatGPT 4.o (accuracy rate %)	ChatGPT o1 (accuracy rate %)	p
Retina, vitreous, and uvea	88	96	0.01*
Pediatric ophthalmology and strabismus	85.71	90.29	0.253
External, corneal, and adnexal disease	82.29	88	0.181
Glaucoma, cataract, and refractive surgery	69.71	85.14	<0.001*
Optics and refraction	76.77	80	0.582
Neuro-ophthalmology	68.57	74.85	0.243
Orbital disease and oculoplastic surgery	72	86.86	<0.001*
General medicine relevant to ophthalmology	95.43	96	0.989
Ophthalmic pathology, microbiology	91.43	97.14	0.040*
Pharmacology and therapeutics	82.86	85.71	0.562

* P-value under 0.05 was considered statistically significant.

Table 2. Comparison of ChatGPT 4.o, Claude 3.5 Sonnet and Gemini 2.0 Flash Experimental Performance Across Subspecialties in ebodtraining.com Question Bank

Subspecialty	ChatGPT (accuracy rate %)	Gemini 2.0 Flash o1 experimental (accuracy rate %)	Claude 3.5 sonnet (accuracy rate %)	p
Retina, vitreous, and uvea	96	81.14	89.71	<0.01*
Pediatric ophthalmology and strabismus	90.29	80	86.29	0.02*
External, corneal, and adnexal disease	88	89.14	82.29	0.134
Glaucoma, cataract, and refractive surgery	85.14	73.14	79.43	0.02*
Optics and refraction	80	74.19	76.77	0.482
Neuro-ophthalmology	74.85	69.71	76.57	0.327
Orbital disease and oculoplastic surgery	86.86	83.43	81.71	0.419
General medicine relevant to ophthalmology	96	92.57	93.71	0.385
Ophthalmic pathology, microbiology	97.14	92.57	93.14	0.136
Pharmacology and therapeutics	85.71	81.71	88	0.257

* P-value under 0.05 was considered statistically significant.

Table 3. Performance Comparison of ChatGPT 4.o and ChatGPT o1 Across Subspecialties in eyedocs.co.uk/ Question Bank

Subspecialty	ChatGPT 4.o (accuracy rate %)	ChatGPT o1 (accuracy rate %)	p
Retina and uveitis	62.86	88.57	0.03*
Pediatrics and strabismus	62.86	88.57	0.03*
External eye and cornea	57.14	91.43	<0.001*
Glaucoma and cataract	54.29	85.71	0.01*
Optics	83.33	90	0.728
Neurology and pupils	85.71	85.71	0.977
Orbit and oculoplastics	77.14	82.86	0.774
General medicine	80	94.29	0.156
Pathology and microbiology	94.29	97.14	0.996
Pharmacology and therapeutics	77.14	77.14	0.997

*P-value under 0.05 was considered statistically significant.

o1 generated statistically significantly better performance in all categories except "Pharmacology and therapeutics" and "Glaucoma and cataract" sections. In "Pharmacology and therapeutics" section, Claude 3.5 Sonnet was found to be statistically significantly superior when compared to ChatGPT o1 and the performance of Gemini 2.0 Flash Experimental was found statistically significanly weaker than other models in "Glaucoma and cataract" section ($p<0.001$, $p<0.001$, respectively) (Table 4).

The analysis of question formats also revealed differences between the two question banks. The https://www.eyedocs.co.uk/ question bank consisted of multiple-choice questions, while the https://ebodtraining.com/ utilized a true/false format. The true/false format resulted in higher percentages of correct answers across most AI models, especially prominent in Chat GPT 4o compared to the multiple-choice format of https://www.eyedocs.co.uk/ questions.

Discussion

The comparative analysis of ChatGPT o1, ChatGPT 4o, Claude 3.5 Sonnet, and Gemini 2.0 Flash Experimental revealed both the promise and limitations of present AI technology in ophthalmology education. While all models demonstrated impressive capabilities in certain subspecialties, particularly in general medical knowledge and basic sciences, their performance varied remarkably among different question types and subspecialties.

In literature, although these specific question banks have not been analyzed yet, similar ophthalmology board examination preparation question banks were evaluated. In a study by Mihalache et al.,[9] it was revealed that ChatGPT correctly answered 46% of ophthalmic board certification preparation questions from the OphthoQuestions question bank. In another study that compared the performance of ChatGPT 3.5 with the 4o version in

Table 4. Comparison of ChatGPT 4.o, Claude 3.5 Sonnet and Gemini 2.0 Flash experimental performance Across Subspecialties in eyedocs.co.uk/ Question Bank

Subspecialty	ChatGPT o1 (accuracy rate %)	Gemini 2.0 Flash experimental (accuracy rate %)	Claude 3.5 Sonnet (accuracy rate %)	p
Retina and uveitis	88.57	82.86	85.71	<0.001*
Pediatrics and strabismus	88.57	65.71	80	0.01*
External eye and cornea	91.43	77.14	88.57	<0.001*
Glaucoma and cataract	85.71	80	85.71	<0.001*
Optics	90	66.67	83.33	0.008*
Neurology and pupils	85.71	80	80	0.001*
Orbit and oculoplastics	82.86	74.29	71.43	0.02*
General medicine	94.29	88.57	88.57	<0.001*
Pathology and microbiology	97.14	91.43	94.29	<0.001*
Pharmacology and therapeutics	77.14	85.71	88.57	<0.001*

*P-value under 0.05 was considered statistically significant.

answering ophthalmology board preparation questions from Ophthalmology Board Review Q&A question bank, reported that, of the total questions, the rate of correct answers for GPT-3.5 was detected 46.7% however it was detected 62.9% with the ChatGPT 4o version.[10] For reference, in EBO board examinations, students must achieve 60% of correct answers to pass.[11] On the other hand, the average Tshe Royal College of Ophthalmologists' examination pass mark has been reported as 60.2% for Part 1 and 63% for Part 2 since 2013.[12] The comparative analysis revealed particularly promising results for ChatGPT o1, which significantly outperformed other present AI models, including ChatGPT 4o, Claude 3.5 Sonnet, and Gemini 2.0 Flash Experimental in multiple ophthalmology subspecialties. When compared to earlier ChatGPT versions reported in the literature, ChatGPT o1 demonstrated remarkably higher accuracy rates in answering board-style questions. The high performance of ChatGPT o1 in ophthalmology board-style examinations, approaching EBO and ICO standards, showed promising potential for AI applications in ophthalmology education.

It is important to mention the difference in success rates of AI models between https://ebodtraining.com/ and https://www.eyedocs.co.uk/ question banks. In our study, most models performed better in https://ebodtraining.com/ question bank and the possible explanation might be the question structure. https://ebodtraining.com/ questions are usually in True/False option-based; however, https://www.eyedocs.co.uk/ questions are in a multiple-choice style. The True/False option-based questions might cause beneficial conditions for AI models by finding sentences more definite and broadly. Furthermore, the case-based questions were existed more in https://www.eyedocs.co.uk/ when compared to https://ebodtraining.com/ and the models might have less success in case-based questions because they require wisdom and deep evaluation. Chatbots, such as ChatGPT, might have difficulty in solving complex medical scenarios, such as case report-like questions, and it is well known that ChatGPT is considered more successful in single questions; however, case-questions require several answers to achieve the result.[13] Although significant performance variations were observed between different question formats in earlier versions of ChatGPT, the latest version demonstrated more consistent performance across diverse question formats, suggesting that this limitation appears to have been temporarily resolved.

The accuracy performance of AI models was evaluated in terms of subspecialty in both question banks. Although AI models showed inconsistent performance across subspecialties in https://www.eyedocs.co.uk/, most models demonstrated worse performance in "Neuro-ophthalmology" and Glaucoma, cataract, and refractive surgery' sections in https://ebodtraining.com/. The explanation in all models' insufficient performance in "Glaucoma, cataract and refractive surgery" section might be the necessity of specific knowledge, including surgical techniques in this section. In addition, the possible explanation of the low success rate in "Neuro-ophthalmology" section might be the high ratio of case-based questions. The highly specialized nature of these sections might also be challenging for AI models. This assumption can be supported by the fact that in subspecialties that require more clinical knowledge, such as "General medicine relevant to ophthalmology" and

"Ophthalmic pathology, microbiology," models performed much better in both question banks. Similar to this study, in Taloni et al.'s study,[14] it was reported that ChatGPT was found more successful on clinical questions when compared to surgical cases. Furthermore, in a study by Antaki et al.,[15] ChatGPT was found to be more successful in "General medicine related to ophthalmology" and the fundamentals. The worst section in that study was "Neuro-ophthalmology" which is consistent with our study.

Recent studies across various medical specialties also reported the evolving capabilities of AI models in medical education. Gencer and Gencer.[16] demonstrated ChatGPT-3.5's superior performance over medical graduates in specialization exams. Similarly, in oncology board examinations, Erdat and Kavak.[17] reported that advanced AI models, such as Claude 3.5 Sonnet and ChatGPT 4o achieved impressive scores of 77.6% and 67.8%, respectively. In medical neuroscience, Mavrych et al.[18] found that Claude 3.5 Sonnet and GPT-4 achieved 83% and 81.7% accuracy, respectively, surpassing average student performance. Regarding image-based questions, Vrindten et al.[19] found that GPT-4 achieved 78% accuracy on surgical image-based questions, outperforming both other AI models (Claude-3: 58%, Gemini-1.5: 57.3%) and medical students (67.4%). These findings across multiple medical disciplines highlight the potential of AI as a transformative educational tool, promising to enhance both teaching and learning across the broader spectrum of medical education.

AI models can be valuable tools in getting prepared for board examinations with their unique practice opportunities for students. In the modern world, time is getting more precious every day, and with the help of AI models, students can receive information and explanations rapidly. In addition, with the advantage of asking further questions, students can access personalized answers, which can reduce time consumption and result in a tailored learning experience.[10]

However, AI models have some disadvantages in the learning experience, especially in ophthalmology. First, "the lack of optimization and reliability of the information."[10] Access to recent and reliable ophthalmology literature and guidelines are essential in learning and accurate managing; however, restricted access to reliable databases can cause a serious limit in using AI for educational purposes.[20] Second, "the lack of ability to process images and videos." It is an important limitation because clinical practice in ophthalmology mainly relies on the slit-lamp examination, fundus evaluation, and additional visual imaging to diagnose, treat, and monitor patients. Some AI models, such as ChatGPT's inability to assess questions with visual features, cause an important limitation. However, novel technology, such as Claude 3.5 Sonnet, is capable of interpreting images and to the best of our knowledge, no study up to date has evaluated Claude 3.5 Sonnet's performance specifically in ophthalmology-related images; its capabilities in other medical departments have been recently analyzed. In a recent study by Kurokawa et al.,[21] Claude 3.5 Sonnet successfully diagnosed 30.1% of radiology image-based case questions. Another study reported that Claude 3.5 Sonnet achieved a 59% success rate in correctly diagnosing breast ultrasound images.[22] In addition, it is well known that the concept of multiple-choice questions does not capture the clinical decision and managing totally. These types of questions can be used to test theoretical knowledge; however, in the real world, managing a patient requires deep and complex evaluation.

Despite promising results, present AI models cannot fully replicate the comprehensive expertise required for ophthalmology board certifications like the fellowship of EBO or ICO. The significant performance drop observed across all models in complex surgical scenarios, case-based reasoning, and subspecialty-specific questions (particularly in neuro-ophthalmology and surgical subspecialties) highlights the present limitations of AI in replicating the decision-making required for clinical practice.

While present AI models show promise in educational settings, future research should investigate their potential in hybrid learning environments where AI assists both educators and residents. Specifically, studies could explore how AI tools might enhance surgical training through real-time feedback systems and anatomical visualization. Future studies should also focus on developing AI models specifically trained in comprehensive ophthalmology, which could lead to better performance in complex subspecialties, such as neuro-ophthalmology. Furthermore, investigating the impact of AI-assisted learning on board examination outcomes through controlled studies would provide valuable insights into their educational effectiveness. This could help establish evidence-based guidelines for integrating AI tools into ophthalmology residency programs while maintaining the essential aspects of clinical expertise and hands-on training.

Conclusion

AI models have the potential to be valuable tools for ophthalmologists with their accessible and rapid nature. However, it should be considered that in evaluating complex cases or visual-based management, these platforms

might be insufficient. These models can be valuable complementary resources in ophthalmology education and board examination preparation, significantly reducing time consumption.

While AI chatbots show impressive results in medical exams, matching EBO and ICO standards, they face important limitations in clinical decision-making and patient care. AI tools can be valuable in supporting roles, particularly in medical education, diagnosis, and patient education. As technology progresses, chatbots will likely serve as helpful tools for ophthalmologists rather than replacing them entirely.

Ethics Committee Approval: Since this study involved no human participants, institutional review board approval wasn't necessary.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept: P.K., H.G.K.; Design: P.K., H.G.K.; Supervision: P.K., H.G.K.; Resource: P.K., H.G.K.; Materials: P.K., H.G.K.; Data Collection and/or Processing: P.K., H.G.K.; Analysis and/or Interpretation: P.K., H.G.K.; Literature Search: P.K.; Writing: P.K., H.G.K.; Critical Reviews: P.K., H.G.K.;

Conflict of Interest: None declared.

Use of AI for Writing Assistance: Not declared.

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

References

1. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2:e0000198. [CrossRef]
2. Open AI. Introducing OpenAI o1-preview. https://openai.com/index/introducing-openai-o1-preview/. Accessed June 26, 2025.
3. Jiao C, Edupuganti NR, Patel PA, Bui T, Sheth V. Evaluating the artificial intelligence performance growth in ophthalmic knowledge. Cureus 2023;15:e45700. [CrossRef]
4. Anthropic. Meet Claude, your thinking partner. https://www.anthropic.com/claude. Accessed June 26, 2025.
5. Google AI for Developers. Gemini modelleri. https://ai.google.dev/gemini-api/docs/models/gemini-v2?hl=tr. Accessed June 26, 2025.
6. EBOD Training. Immersive European Ophthalmology Exams Preparatory Course. https://ebodtraining.com/. Accessed June 26, 2025.
7. Eyedocs. https://www.eyedocs.co.uk/. Accessed June 26, 2025.
8. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312. Erratum in: JMIR Med Educ 2024;10:e57594. [CrossRef]
9. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol 2023;141:589–97. [CrossRef]
10. Haddad F, Saade JS. Performance of ChatGPT on ophthalmology-related questions across various examination levels: Observational study. JMIR Med Educ 2024;10:e50842. [CrossRef]
11. European Board of Ophthalmology. https://www.ebo-online.org/. Accessed June 26, 2025.
12. The Royal College of Ophthalmologists. RCOphth exams. https://www.rcophth.ac.uk/examinations/rcophth-exams/. Accessed June 26, 2025.
13. Cai LZ, Shaheen A, Jin A, Fukui R, Yi JS, Yannuzzi N, et al. Performance of generative large language models on ophthalmology board-style questions. Am J Ophthalmol 2023;254:141–9. [CrossRef]
14. Taloni A, Borselli M, Scarsi V, Rossi C, Coco G, Scorcia V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. Sci Rep 2023;13:18562. [CrossRef]
15. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: An analysis of its successes and shortcomings. Ophthalmol Sci 2023;3:100324. [CrossRef]
16. Gencer G, Gencer K. A Comparative analysis of ChatGPT and medical faculty graduates in medical specialization exams: Uncovering the potential of artificial intelligence in medical education. Cureus 2024;16:e66517. [CrossRef]
17. Erdat EC, Kavak EE. Benchmarking LLM chatbots' oncological knowledge with the Turkish Society of Medical Oncology's annual board examination questions. BMC Cancer 2025;25:197. [CrossRef]
18. Mavrych V, Yaqinuddin A, Bolgova O. Claude, ChatGPT, Copilot, and Gemini performance versus students in different topics of neuroscience. Adv Physiol Educ 2025;49:430–7. [CrossRef]
19. Vrindten KL, Hsu M, Han Y, Rust B, Truumees H, Katt BM. Evaluating the performance of ChatGPT4.0 versus ChatGPT3.5 on the Hand Surgery Self-Assessment Exam: A comparative analysis of performance on image-based questions. Cureus 2025;17:e77550. [CrossRef]
20. Botross M, Mohammadi S, Montgomery K, Crawford C. Performance of Google's artificial intelligence chatbot "Bard" (now "Gemini") on ophthalmology board exam practice questions. Cureus 2024;16:e57348. [CrossRef]
21. Kurokawa R, Ohizumi Y, Kanzawa J, Kurokawa M, Sonoda Y, Nakamura Y, et al. Diagnostic performances of Claude 3 Opus and Claude 3.5 Sonnet from patient history and key images in Radiology's "Diagnosis Please" cases. Jpn J Radiol 2024;42:1399–402. [CrossRef]
22. Güneş YC, Cesur T, Çamur E, Günbey Karabekmez L. Evaluating text and visual diagnostic capabilities of large language models on questions related to the Breast Imaging Reporting and Data System Atlas 5th edition. Diagn Interv Radiol 2025;31:111–29. [CrossRef]