Original Article

Evolving Minds: Natural Learning vs. Artificial Learning in Ophthalmology Training

Ali Safa Balcı[1], Zeliha Yazar[2], Banu Turgut Öztürk[3], Çiğdem Altan[4]

[1]University of Health Sciences Türkiye, Sancaktepe Şehit Prof. Dr. İlhan Varank Training and Research Hospital, Clinic of Ophthalmology, İstanbul, Türkiye
[2]Ankara Bilkent City Hospital, Clinic of Ophthalmology, Ankara, Türkiye
[3]Selçuk University Faculty of Medicine, Department of Ophthalmology, Konya, Türkiye
[4]University of Health Sciences Türkiye, Beyoğlu Eye Training and Research Hospital, Clinic of Ophthalmology, İstanbul, Türkiye

Abstract

Objectives: This study aimed to compare year-over-year change in ChatGPT's performance on nationwide ophthalmology exams with the performance change among residents over the same period.

Materials and Methods: This observational study included ophthalmology residents in Türkiye who participated in both the 2023 and 2024 Resident Training Development Exams organized by the Turkish Ophthalmological Association Qualifications Committee. The 2023 examination consisted of 69 single-best-answer multiple-choice questions and was administered to ChatGPT-3.5. The 2024 version, containing 72 questions, was administered to ChatGPT-4o. The success rates of ChatGPT and the residents who participated in both exams were compared.

Results: ChatGPT's accuracy improved from 53.6% in 2023 to 84.7% in 2024. Among the 501 residents who participated in both years, the average score increased from 48.2% to 53.1%. ChatGPT ranked 292nd among residents in 2023 but achieved the top score in 2024. Based on percentage improvement in scores, ChatGPT-4o ranked 8th overall. The most notable performance gains for ChatGPT were seen in the areas of strabismus (+75%), neuro-ophthalmology (+40%), and optics (+40%). Among residents, the largest improvement occurred in oculoplastics (+33.5%), while a decrease was observed in cornea and ocular surface (-4.1%).

Conclusion: ChatGPT-4o showed a marked improvement in answering ophthalmology questions compared to its predecessor, whereas resident learning progressed more gradually. This rapid advancement in ChatGPT highlights the potential speed with which artificial learning can progress within defined boundaries. In contrast, human learning remains a deeper and more time-intensive process. Results suggest that evolving large language models will play an increasingly significant role in medical education and clinical support.

Keywords: Education, generative artificial intelligence, resident training

Cite this article as: Balcı AS, Yazar Z, Turgut Öztürk B, Altan Ç. Evolving Minds: Natural Learning vs. Artificial Learning in Ophthalmology Training. Turk J Ophthalmol. 2026;56:1-7

Address for Correspondence: Ali Safa Balcı, University of Health Sciences Türkiye, Sancaktepe Şehit Prof. Dr. İlhan Varank Training and Research Hospital, Clinic of Ophthalmology, İstanbul, Türkiye
E-mail: alisafabalci@gmail.com
ORCID-ID: orcid.org/0000-0001-9161-7839

Received: 02.08.2025
Revision Requested: 25.09.2025
Last Revision Received: 27.10.2025
Accepted: 11.12.2025
Publication Date: 18.02.2026

DOI: 10.4274/tjo.galenos.2025.16517

Introduction

Large language models such as OpenAI's ChatGPT are advanced artificial intelligence systems that operate based on natural language processing techniques and are capable of generating human-like responses. Built on the Generative Pre-trained Transformer (GPT) architecture, these models have reached the capacity to produce contextually consistent and meaningful responses by training on vast and diverse text datasets. While earlier versions such as ChatGPT-3.5 and ChatGPT-4.0 made significant advances in natural language comprehension, ChatGPT-4o (released on May 13, 2024) showed marked improvement over previous versions in terms of linguistic accuracy and interactional performance.[1] With increasing competencies in medical, educational, and academic contexts, ChatGPT exhibits high levels of accuracy and responsiveness.[2] Nevertheless, in the medical field especially, their responses must be

continuously evaluated to ensure clinical reliability for both healthcare professionals and patients.

The rapidly increasing popularity of ChatGPT in the medical field has led to growing interest in assessing its functionality in various health-related tasks. Research in ophthalmology has examined the accuracy of ChatGPT's responses to questions about various subspecialties, providing evidence that this AI model can be used as a complementary educational tool.[3,4,5] Furthermore, the ability to interpret and manage ocular conditions in clinical scenarios such as corneal ulcer, cataract management, and retinal pathologies has also been explored.[6,7,8] ChatGPT has emerged as a tool that can be helpful not only in diagnosis, but also in documentation processes such as the preparation of medical reports, as well as turning complex ophthalmological information into more understandable and accessible educational content.[9]

In this study, we aimed to evaluate the performance of both the ophthalmology residents who took a national resident training exam in two consecutive years and the ChatGPT models current in those years, and compare their year-over-year changes in performance.

Materials and Methods

The Turkish Ophthalmological Association Qualifications Committee held the third annual Resident Training Development Exam on May 26, 2023. In our previous study, we posed the questions from this exam to ChatGPT-3.5, the previous version of ChatGPT, and compared its performance with the results of the ophthalmology residents who took the exam nationwide.[10] The following year, on May 31, 2024, a total of 1,013 ophthalmology residents from 80 training centers across Türkiye took the fourth Resident Training Development Exam. In the present study, we administered the questions from this exam to ChatGPT-4o, the most current version of ChatGPT. Only residents who took the exam in both years were included in the study to allow a year-over-year analysis. The residents were grouped according to their year of training as of the 2024 exam date.

Both exams were prepared by the Turkish Ophthalmological Association Qualifications Committee to cover the same subspecialties, at a similar difficulty level. Each exam included a total of 75 questions. However, questions in a format other than single-best-answer multiple-choice questions were excluded from the study. Therefore, the analysis included 69 eligible questions from 2023 and 72 eligible questions from 2024. Distributions of the questions asked in 2023 and 2024 by subspecialty are given in Table 1 and Table 2, respectively.

After translating into English, the 2024 exam questions were posed to ChatGPT-4o (model identifier: gpt-4o-2024-05-13) using the official web interface on the website (https://chat.openai.com) in separate chat sessions on March 21, 2025. The system history was cleared before each question. As none of the questions contained visual or graphic content, no additional transcription or image description was required. To avoid the impact of subsequent updates to the language model, each question was accompanied by the prompt, "Answer the following question using the knowledge available as of May 31, 2024." The answers and explanations given by ChatGPT-4o for each question were recorded, and each response was evaluated as correct or incorrect according to the predetermined answer key.

Residents and ChatGPT-4o were scored out of 100 based on the number of correct answers. Additionally, a ranking was created based on these scores, calculated according to the number of examinees in the relevant year. Changes in performance were analyzed overall and by subspecialty for both the residents and ChatGPT. Year-over-year change in resident performance was determined from the average accuracy rates of the 501 residents who participated in both exams.

Ethics committee approval was not required because participant information was anonymized and no personal data were used.

Statistical Analysis

Statistical analyses were performed using SPSS version 26 (IBM, Armonk, NY, USA). The Kolmogorov-Smirnov test was used to evaluate the normality of data distributions. Resident data were not evaluated individually but averaged across the 501 residents in the sample. Comparisons between ChatGPT and the resident group were made descriptively, not with statistical tests. As ChatGPT provides a single model output, variation was not calculated and differences were compared using accuracy rates alone. Continuous variables were presented as mean ± standard deviation and range. The Wilcoxon signed rank test was used to analyze the change in resident accuracy rate overall and by subspecialty between the 2023 and 2024 exams. The 95% confidence intervals (CI) were calculated for the accuracy rates of the resident participant group and ChatGPT models. For comparisons of subspecialty, Bonferroni correction was performed to reduce the probability of type I error and the significance level was determined as $p<0.005$.

Table 1. Mean number of correct answers by residents and ChatGPT-3.5 on the 2023 exam, by subspecialty					
Subspecialty (number of questions)	**All residents (n=501)**	**First-year residents (n=249)**	**Second-year residents (n=132)**	**Third-year residents (n=120)**	**ChatGPT-3.5**
Lens and cataract (n=9)	3.82±1.64	3.16±1.36	4.21±1.63	4.77±1.61	7
Cornea/ocular surface/anterior segment (n=9)	4.55±1.24	4.4±1.29	4.57±1.24	4.83±1.09	4
Glaucoma (n=8)	3.34±1.41	2.92±1.34	3.54±1.4	4.01±1.29	4
Neuro-ophthalmology (n=5)	2.21±0.946	2.12±0.93	2.2±0.94	2.41±0.97	3
Oculoplasty (n=4)	1.30±0.83	1.11±0.78	1.35±0.76	1.63±0.89	2
Pediatric ophthalmology and strabismus (n=8)	3.96±1.51	3.62±1.46	4.27±1.49	4.32±1.52	0
Optics (n=5)	2.07±1.13	1.91±1.06	2.12±1.16	2.36±1.19	3
Retina (n=16)	8.98±2.53	7.85±2.27	9.68±2.33	10.56±2.15	11
Uveitis (n=5)	3.0±1.14	2.79±1.11	3.01±1.16	3.45±1.08	3
Total (n=69)	33.24±7.32	29.88±6.31	34.95±6.34	38.33±6.74	37

Table 2. Mean number of correct answers by residents and ChatGPT-4o on the 2024 exam, by subspecialty					
Subspecialty (number of questions)	**All residents (n=501)**	**Second-year residents (n=249)**	**Third-year residents (n=132)**	**≥Fourth-year residents (n=120)**	**ChatGPT-4o**
Lens and cataract (n=9)	4.14±1.59	3.81±1.50	4.21±1.46	4.73±1.71	8
Cornea/ocular surface/anterior segment (n=11)	5.11±1.90	4.59±1.81	5.39±1.89	5.87±1.76	9
Glaucoma (n=7)	3.34±1.28	3.11±1.21	3.47±1.25	3.66±1.36	5
Neuro-ophthalmology (n=4)	1.84±1.05	1.78±1.05	1.87±1.06	1.95±1.03	4
Oculoplasty (n=7)	4.61±1.32	4.20±1.28	4.74±1.30	5.32±1.07	6
Pediatric ophthalmology and strabismus (n=8)	4.34±1.59	3.96±1.49	4.52±1.64	4.93±1.52	6
Optics (n=5)	2.21±0.98	2.04±0.92	2.26±0.97	2.49±1.06	5
Retina (n=16)	9.30±2.46	8.62±2.33	9.52±2.52	10.45±2.20	14
Uveitis (n=5)	3.20±1.17	2.96±1.07	3.24±1.21	3.66±1.16	4
Total (n=72)	38.20±8.47	35.12±7.07	39.42±8.99	43.25±7.82	61

Results

A total of 501 ophthalmology residents took the exam in both years. When categorized by months of training in 2024, there were 249 second-year residents (12-23 months of experience), 132 third-year residents (24-35 months of experience), and 120 fourth-year or higher residents (≥36 months of experience). The mean training duration of the residents was 28.4±10.6 months (range, 13-64 months). Residents who took the exam in both years correctly answered a mean of 38.2±8.5 of the 72 questions in the 2024 exam, achieving a success rate of 53.1% (95% CI: 52.2%-54.0%). Second-year residents achieved a success rate of 48.8% (95% CI: 47.5%-50.1%), with a mean of 35.1±7.1 correct answers; third-year residents had a success rate of 54.8% (95% CI: 53.3%-56.3%), with a mean of 39.4±8.9 correct answers; and fourth-year or higher residents reached a success rate of 60.1% (95% CI: 58.7%-61.5%), with a mean of 43.3±7.8 correct answers. In contrast, ChatGPT-4o answered 61 of the 72 questions correctly, for an accuracy rate of 84.7% (95% CI: 74.7%-91.3%). ChatGPT-3.5 ranked 292nd among residents in the 2023 exam, whereas ChatGPT-4o achieved the top score in 2024. The mean numbers of questions (overall and by subspecialty) answered correctly by residents and ChatGPT-3.5 in 2023 are presented in Table 1, and the means of the same residents and ChatGPT-4o for 2024 are presented in Table 2.

Overall, the residents' accuracy rates in most subspecialties improved compared to the previous year, although this increase did not reach statistical significance in the field of neuro-ophthalmology (p=0.655). Corneal and ocular surface diseases was the only subspecialty in which residents' performance declined, and this decrease

was statistically significant (p<0.001). In contrast, ChatGPT showed major improvements across all subspecialties, with a 30.4% increase in overall accuracy rate. When the residents and ChatGPT were ranked according to the percentage increase in accuracy rate, ChatGPT-4o ranked 8th. Changes in accuracy rates between the two exams are summarized in Table 3.

Discussion

This study aimed to evaluate the year-over-year change in performance of Turkish ophthalmology residents and a large language model based on a nationwide resident training exam held over two consecutive years, thereby presenting a comparison of natural versus artificial learning. Our findings revealed that residents' average performance in most subspecialties improved between 2023 and 2024, whereas ChatGPT-4o showed consistent improvement over its predecessor ChatGPT-3.5 in all areas and outperformed all human examinees in 2024.

The widespread adoption of AI in the healthcare field has led to an increasing trend among both patients and healthcare professionals toward using these tools to obtain medical information and provide educational support.[11,12] As their use becomes increasingly widespread, particularly through advanced large language models like ChatGPT-4o, it is becoming more important than ever to evaluate the reliability and scientific accuracy of the responses produced by these systems. Despite the advantage of providing rapid and accessible information, their potential impact on clinical decision-making processes and medical education makes it imperative to rigorously assess their responses to domain-specific, evidence-based questions.

Artificial intelligence systems are constantly evolving and learning. ChatGPT-4.0 was reported to show improved accuracy when asked the same questions about intraocular lenses six months apart.[13] In another study, when medical questions initially answered incorrectly by ChatGPT were re-asked a short time later (8-17 days), the model answered most of the questions correctly.[14]

While human learning is a gradual process shaped by experience, cognition, and context, large language models such as ChatGPT acquire knowledge through periodic large-scale retraining cycles.[15] Each new release, such as ChatGPT-4o, reflects a gradual progression, enhanced by insights from increasingly diverse, current, and domain-specific datasets. This process enables rapid

Table 3. Mean change in the percentage of correct answers between the two exams for residents and ChatGPT

Subspecialty	Percentage change, all residents (%) (n=501)	Percentage change, ChatGPT (%)	p*
Lens and cataract	3.48±21.38 (-66.67 to 66.67)	11.11	**<0.001**
Cornea/ocular surface/anterior segment	-4.06±20.78 (-57.58 to 72.73)	37.37	**<0.001**
Glaucoma	5.85±23.36 (-50.0 to 73.21)	21.43	**<0.001**
Neuro-ophthalmology	1.82±29.73 (-80.0 to 80.0)	40.0	0.655
Oculoplasty	33.46±26.34 (-50.0 to 100.0)	35.71	**<0.001**
Pediatric ophthalmology and strabismus	4.74±22.63 (-62.50 to 75.0)	75.0	**<0.001**
Optics	2.67±27.91 (-80.0 to 80.0)	40.0	0.029
Retina	1.95±17.17 (-50.0 to 62.50)	18.75	0.011
Uveitis	3.91±28.18 (-80.0 to 80.0)	20.0	**0.002**
Total	4.31±9.97 (-39.81 to 56.75)	30.38	**<0.001**

*Change in the percentage of correct answers between the two exams for residents who took both; Wilcoxon signed rank test

and effective improvements in information accuracy and functional performance. However, this development lacks the continuity, ethical reasoning, and experiential depth involved in human learning.[16] In contrast, humans experience a slower but more holistic learning process. Knowledge is not only acquired through formal education, but is also shaped through trial and error, emotional context, and social interaction.[17] Especially in medical education, this type of learning process enhances qualities such as clinical judgment, empathy, and adaptability, which current artificial intelligence systems have yet to attain.[18]

The performance of large language models and humans on ophthalmology-related questions has also been compared in previous studies.[19,20] In another study using ophthalmology residency exam questions from 2020 to 2023, large language models did not show a significant change in accuracy over the four years.[21] However, it was not specified exactly when the questions were posed to the large language models; if all the questions were asked at approximately the same time, accuracy rates would be expected to remain similar even if the test years were different.[21] In a study by Taloni et al.[22] using 1,023 questions from the BCSC (Basic and Clinical Science Course) question set of the American Academy of Ophthalmology, ChatGPT-4.0 outperformed its predecessor ChatGPT-3.5. Human participants ranked second in overall performance. Similarly, Maino et al.[23] evaluated 440 previously administered multiple-choice questions on the European Board of Ophthalmology Diploma Examination and reported that ophthalmology residents performed better than ChatGPT-3.5 but were less accurate than ChatGPT-4o.

Although our findings are generally consistent with these studies, there is an important difference in study design. While previous studies adopted a cross-sectional approach, our study involved two similar national exams administered to the same group of residents one year apart, thereby enabling the observation of longitudinal changes. Moreover, we assessed not only human learning, but also the change in performance between successive versions of the same large language model. To the best of our knowledge, our study is the first to provide a parallel view of the progress of both human and machine learning over time.

The overall increase in resident scores is a positive indicator of the effectiveness of training over time, suggesting that structured training programs together with clinical experience contribute to knowledge retention. Interestingly, the only subspecialty with no statistically significant improvement was neuro-ophthalmology. This area is known for its multidisciplinary nature and limited clinical exposure in many training centers.[24] The only area in which resident performance declined significantly was corneal and ocular surface diseases. This may point to factors such as insufficient emphasis on this subspecialty in the training curriculum or a scarcity of clinical cases. These findings may guide future modifications to residency training programs, especially in terms of identifying areas that need strengthening. In contrast, ChatGPT-4o performed strongly in all subspecialties and showed significant improvement over ChatGPT-3.5. ChatGPT-4o had an overall accuracy rate of 84.7%, exhibiting greater accuracy and consistency than the resident group, although it ranked 8th in terms of year-over-year performance improvement. This reinforces the increasing potential of large language models as educational tools in medical education, especially in terms of exam preparation and theoretical knowledge support. However, it should be noted that these models do not include elements important to medical practice such as contextual nuance, clinical judgment, and practical skills. Therefore, such AI tools should be considered a supportive and complementary component of traditional medical education rather than a replacement.

Study Limitations

Our study has some limitations that should be noted. First, although a longitudinal comparison was made, the effect of variables such as individual learning environments, level of clinical experience, and work-related habits is unknown. Second, although both exams were similar in content and structure, their psychometric equivalence has not been assessed at the item level. Therefore, the study evaluates year-over-year differences not as absolute values, but as relative change in performance under similar conditions. Regarding the AI methodology, the web-based interface offers limited control over response length and context memory compared to the application versions of ChatGPT. This may lead to minor differences in responses, which we consider a methodological limitation. Furthermore, the selection of residents who participated in both exams may have introduced selection bias, as this approach could select for individuals who are more motivated or academically inclined. Finally, the limited number of questions and the fact that the study is based on the national exam of a single country may limit the generalizability of the findings to different education systems.

Conclusion

ChatGPT-4o demonstrated improved accuracy over the previous version (ChatGPT-3.5) and outperformed the resident group in the 2024 national ophthalmology

resident training exam. While residents showed more modest improvement, the dramatic progress made by ChatGPT-4o underscores the evolving capabilities of large language models. However, it is important to note that despite their high accuracy, these models can occasionally generate erroneous or misleading responses. Therefore, their role in medical education should be complementary, regarded as a supportive tool rather than a substitute for the critical thinking and experience-based knowledge that develops in humans through training.

Ethics

Ethics Committee Approval: Not required.

Informed Consent: Not required.

Declarations

Authorship Contributions

Concept: A.S.B., Z.Y., B.T.Ö., Ç.A., Design: A.S.B., Z.Y., B.T.Ö., Ç.A., Data Collection or Processing: A.S.B., Z.Y., Analysis or Interpretation: A.S.B., Literature Search: A.S.B., Writing: A.S.B., Z.Y., B.T.Ö., Ç.A.

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

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

References

1. OpenAI. ChatGPT-4o: Advancements and new peatures. OpenAI Blog. Accessed June 11, 2025. Available from: https://openai.com/blog
2. Gumus Akgun G, Altan C, Balci AS, Alagoz N, Çakır I, Yaşar T. Using ChatGPT-4 in visual field test assessment. Clin Exp Optom. 2025;108:1031-1036.
3. Aydın FO, Aksoy BK, Ceylan A, Akbaş YB, Ermiş S, Kepez Yıldız B, Yıldırım Y. 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;54:313-317.
4. Durmaz Engin C, Karatas E, Ozturk T. Exploring the role of ChatGPT-4, BingAI, and Gemini as virtual consultants to educate families about retinopathy of prematurity. Children (Basel). 2024;11:750.
5. Bahar TS, Öcal O, Çetinkaya Yaprak A. Comparison of ChatGPT-4, Microsoft Copilot, and Google Gemini for pediatric ophthalmology questions. J Pediatr Ophthalmol Strabismus. 2025;62:428-434.
6. Gurnani B, Kaur K, Gireesh P, Balakrishnan L, Mishra C. Evaluating the novel role of ChatGPT-4 in addressing corneal ulcer queries: an AI-powered insight. Eur J Ophthalmol. 2025;35:1531-1541.
7. Sundaramoorthy S, Ratra V, Shankar V, Dorairajan R, Maskati Q, Fredrick TN, Ratra A, Ratra D. Conversational guide for cataract surgery complications: a comparative study of surgeons versus large language model-based chatbot generated instructions for patient interaction. Ophthalmic Epidemiol. 2025:1-8.
8. 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:e117-e123.
9. Moulaei K, Yadegari A, Baharestani M, Farzanbakhsh S, Sabet B, Reza Afrash M. Generative artificial intelligence in healthcare: A scoping review on benefits, challenges and applications. Int J Med Inform. 2024 Aug;188:105474.
10. Balci AS, Yazar Z, Ozturk BT, Altan C. Performance of ChatGPT in ophthalmology exam; human versus AI. Int Ophthalmol. 2024;44:413.
11. Sevgi M, Antaki F, Keane PA. Medical education with large language models in ophthalmology: custom instructions and enhanced retrieval capabilities. Br J Ophthalmol. 2024;108:1354-1361.
12. Balci AS, Çakmak S. Evaluating the Accuracy and Readability of ChatGPT-4o's responses to patient-based questions about keratoconus. Ophthalmic Epidemiol. 2025:1-6.
13. Aydin FO, Ermis S. Responses of ChatGPT-4 on intraocular lenses: an evolving artificial intelligence assessment. Clin Exp Optom. 2025:1-5.
14. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, Chang S, Berkowitz S, Finn A, Jahangir E, Scoville E, Reese T, Friedman D, Bastarache J, van der Heijden Y, Wright J, Carter N, Alexander M, Choe J, Chastain C, Zic J, Horst S, Turker I, Agarwal R, Osmundson E, Idrees K, Kieman C, Padmanabhan C, Bailey C, Schlegel C, Chambless L, Gibson M, Osterman T, Wheless L. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Sq [Preprint]. 2023:rs.3.rs-2566942.
15. Shi H, Xu Z, Wang H, Qin W, Wang W, Wang Y, Wang Z, Ebrahimi S, Wang H. Continual learning of large language models: a comprehensive survey. arXiv preprint arXiv: 2404.16789. 2024. https://doi.org/10.48550/arXiv.2404.16789
16. Komasawa N, Yokohira M. Learner-centered experience-based medical education in an AI-driven society: a literature review. Cureus. 2023;15:e46883.
17. De Felice S, Hamilton AFC, Ponari M, Vigliocco G. Learning from others is good, with others is better: the role of social interaction in human acquisition of new knowledge. Philos Trans R Soc Lond B Biol Sci. 2023;378:20210357.
18. Ten Cate TJ, Kusurkar RA, Williams GC. How self-determination theory can assist our understanding of the teaching and learning processes in medical education. AMEE guide No. 59. Med Teach. 2011;33:961-973.
19. Fowler T, Pullen S, Birkett L. Performance of ChatGPT and Bard on the official part 1 FRCOphth practice questions. Br J Ophthalmol. 2024;108:1379-1383.
20. Tao BK, Hua N, Milkovich J, Micieli JA. ChatGPT-3.5 and Bing Chat in ophthalmology: an updated evaluation of performance, readability, and informative sources. Eye (Lond). 2024;38:1897-1902.
21. Bahir D, Zur O, Attal L, Nujeidat Z, Knaanie A, Pikkel J, Mimouni M, Plopsky G. Gemini AI vs. ChatGPT: a comprehensive examination alongside ophthalmology residents in medical knowledge. Graefes Arch Clin Exp Ophthalmol. 2025;263:527-536.
22. Taloni A, Borselli M, Scarsi V, Rossi C, Coco G, Scorcia V, Giannaccare G. 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.

23. Maino AP, Klikowski J, Strong B, Ghaffari W, Woźniak M, Bourcier T, Grzybowski A. Artificial Intelligence vs. Human Cognition: A Comparative Analysis of ChatGPT and candidates sitting the european board of ophthalmology diploma examination. Vision. 2025;9:31.

24. Mollan SP, Menon V, Cunningham A, Plant GT, Bennetto L, Wong SH, Dayan M. Neuro-ophthalmology in the United Kingdom: providing a sustainable, safe and high-quality service for the future. Eye (Lond). 2024;38:2235-2237.