AJO International 1 (2024) 100070

Contents lists available at ScienceDirect

AJO International

journal homepage: www.sciencedirect.com/journal/ajo-international

The performance of artificial intelligence-based large language models on ophthalmology-related questions in Swedish proficiency test for medicine: ChatGPT-4 omni vs Gemini 1.5 Pro

Mehmet Cem Sabaner [a,*], Arzu Seyhan Karatepe Hashas [b], Kemal Mert Mutibayraktaroglu [c], Zubeyir Yozgat [a], Oliver Niels Klefter [d,e], Yousif Subhi [d,f]

[a] Department of Ophthalmology, Kastamonu University, Kastamonu Training and Research Hospital, Kastamonu, Turkey
[b] Department of Ophthalmology, Region Västra Götaland, Sahlgrenska University Hospital, Gothenburg, Sweden
[c] Department of Ophthalmology, Region Västra Götaland, South Älvsborg Hospital, Borås, Sweden
[d] Department of Ophthalmology, Rigshospitalet, Glostrup, Denmark
[e] Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark
[f] Department of Clinical Research, University of Southern Denmark, Odense, Denmark

ARTICLE INFO

Keywords:
Artificial intelligence
ChatGPT-4 omni
E-learning
Gemini 1.5 Pro
Large language model
Medical education
Ophthalmology

ABSTRACT

Purpose: To compare the interpretation and response context of two commonly used artificial intelligence (AI)-based large language model (LLM) platforms to ophthalmology-related multiple choice questions (MCQs) in the Swedish proficiency test for medicine ("*kunskapsprov för läkare*") exams.
Design: Observational study.
Methods: The questions of a total of 29 exams held between 2016 and 2024 were reviewed. All ophthalmology-related questions were included in this study, and categorized into ophthalmology sections. Questions were asked to ChatGPT-4o and Gemini 1.5 Pro AI-based LLM chatbots in Swedish and English with specific commands. Secondly, all MCQs were asked again without feedback. As the final step, feedback was given for questions that were still answered incorrectly, and all questions were subsequently re-asked.
Results: A total of 134 ophthalmology-related questions out of 4876 MCQs were evaluated via both AI-based LLMs. The MCQ count in the 29 exams was 4.62 ± 2.21 (range: 0–8). After the final step, ChatGPT-4o achieved higher accuracy in Swedish (94 %) and English (95.5 %) compared to Gemini 1.5 Pro (both at 88.1 %) ($p = 0.13$, and $p = 0.04$, respectively). Moreover, ChatGPT-4o provided more correct answers in the neuro-ophthalmology section ($n = 47$) compared to Gemini 1.5 Pro across all three attempts in English ($p < 0.05$). There was no statistically significant difference either in the inter-AI comparison of other ophthalmology sections or in the inter-lingual comparison within AIs.
Conclusion: Both AI-based LLMs, and especially ChatGPT-4o, appear to perform well in ophthalmology-related MCQs. AI-based LLMs can contribute to ophthalmological medical education not only by selecting correct answers to MCQs but also by providing explanations.

1. Introduction

Artificial intelligence (AI) demonstrates an increasing potential across multiple domains, including education and medicine. It is a continuously evolving entity that progresses to become an indispensable part of medical practice and education of the future.[1–6] Although many varieties have emerged and become available for use in recent times, ChatGPT-4 Omni (ChatGPT-4o) and Gemini 1.5 Pro are state-of-the-art and the most up-to-date models available (ChatGPT-4o, release date May 13, 2024; Gemini 1.5 Pro, release date Feb 15, 2024, last updated May 14, 2024), designed to deliver high-accuracy and contextually relevant answers, and they can also perform tasks such as text generation, language understanding, translation, and summarization. Fundamentally, they are classified as large language models (LLMs), which are

* Corresponding author at: Kastamonu University, Department of Ophthalmology, Training and Research Hospital, 37150 Kastamonu, Turkey.
E-mail addresses: drmcemsabaner@yahoo.com (M.C. Sabaner), arzuskaratepe@hotmail.com (A.S.K. Hashas), kemalmuti@gmail.com (K.M. Mutibayraktaroglu), zubeyiryozgat@gmail.com (Z. Yozgat), oliver.niels.klefter.01@regionh.dk (O.N. Klefter), ysubhi@gmail.com (Y. Subhi).

https://doi.org/10.1016/j.ajoint.2024.100070
Received 28 July 2024; Received in revised form 18 September 2024; Accepted 25 September 2024
Available online 26 September 2024

deep learning models trained on extensive datasets to learn various linguistic features. LLMs utilize natural language processing techniques to generate highly accurate and contextually relevant responses across a broad range of languages.[7] AI chatbots utilize advanced natural language processing (NLP) and machine learning algorithms to understand and respond to user queries. Today, AI-based LLM chatbots are widely used in various fields, particularly in education, healthcare, customer service, and content production.[7–9]

E-learning is becoming increasingly popular, and AI is causing a paradigm shift in education models and habits worldwide.[8,9] For instance, homework or research can be completed, multiple choice questions (MCQs) can be answered, and even academic theses can be written through AI-based LLMs.[10] On the other hand, many academic and educational institutions have established regulations to govern the use of AI among students, and numerous scientific journals now mandate the disclosure of AI technology usage in the preparation of manuscripts.[8–10] However, questions may arise about the accuracy or usefulness of the AI approach.

One potential application could be as an adjunct in preparing medical students and doctors for MCQ exams. One such exam is the Swedish proficiency test for medicine "kunskapspröv för läkare", a certification exam for doctors qualified outside the European Union to work as physicians in Sweden, which also includes ophthalmology-related questions. To apply LLMs as an adjunct in exam preparation, at least two prerequisites should be fulfilled. First, the LLM should be capable of answering the questions correctly preferably being able to provide relevant explanation and references, and secondly, it should be usable in the same language as the exam. However, studies specifically assessing the performance of AI chatbots answering ophthalmological MCQs are limited.[11–16] To test if AI-based LLMs could hold the potential to be integrated into teaching and MCQ preparation for ophthalmology questions in a non-English language, the current study aimed to compare the performance of ChatGPT-4o and Gemini 1.5 Pro AI-based LLMs to ophthalmology-related MCQs in the Swedish proficiency test for medicine exams.

2. Methods

2.1. Study design and data collection

This cross-sectional study included evaluating the answers of two different AI-based LLMs to ophthalmology-related questions in Swedish proficiency test for medicine exams. This is the main exam to practice as a medical doctor in Sweden for doctors qualified outside EU/EEA, and it consists of two steps: Theoretical exam and practical exam. The theoretical exam is in Swedish and aims to measure the knowledge, understanding and assessment skills necessary to work as a physician. This test is conducted digitally in an exam room in Sweden by Umeå University, commissioned by the National Board of Health and Welfare. It is held four times a year, takes place on a single day, and consists of 7.5 h. An official open-access website (https://www.umu.se/utbildning/sok/kunskapsprov/kunskapsprov-for-lakare/teoretiskt-delprov/) provides the opportunity to practice and solve kunskapsprov exams without requiring registration ("The user agrees to use the Website and the content and services offered thereon in accordance with the law, these Terms and Conditions, good faith, good practices, and public order."). Transparency and accessibility, especially in educational materials, are key priorities in Scandinavian countries.

The questions of a total of 29 exams held between 2016 and 2024 are open access and have been made available along with the correct answers.[17] The theoretical exam consists of three parts, which are clinical cases, MCQ140, and scientific article sub-tests. The clinical cases sub-test is a section focused on clinical knowledge, typically comprising an average of five clinical cases, with each case consisting of approximately six MCQs, totaling around 30 MCQs. However, the exact number may vary with each exam. On the other hand, the MCQ140 sub-test is a mixed test with a total of 140 MCQs in each exam, primarily focused on basic medical sciences but also including clinical questions. The clinical cases and MCQ140 sub-tests were included in this study, and all questions from the above-mentioned exams were examined in detail one by one by two researchers (MCS and KMM). Ophthalmology-related questions that could be solved with ophthalmological medical knowledge were identified by same researchers through consensus and included in this study. For questions that remained unclear, assistance was sought from a third researcher (ASKH or ZY). All four researchers hold a C1-level Swedish language certificate and have previously successfully passed the "*kunskapsprov för läkare*" exam. Three of these researchers are senior ophthalmologists, while one is an ophthalmology resident (KMM). The questions were then categorized into ophthalmology sections as "neuro-ophthalmology", "retina and vitreous", "glaucoma", "cornea & anterior segment", "pediatric ophthalmology and strabismus", "oculoplastics and ocular oncology", and "uveitis". Only the shared answer keys were accepted as correct, without any comments, contributions or changes to the answers.

In the next step, the question-answer capabilities of AI-based LLMs were evaluated. The flowchart of the study evaluation process is shown in Fig. 1. Using ChatGPT Plus and Gemini Advanced, access was achieved with the latest up-to-date versions, ChatGPT-4o (OpenAI, San Francisco, CA) and Gemini 1.5 Pro (Google, Mountain View, CA) AI-based LLM chatbots, respectively. The chatbot question-answer process for this study started in the last week of May 2024 and was completed within one week, with no updates occurring during this period. Even if ChatGPT Plus and Gemini Advanced access is used, there is a message limit for both AI-based LLM chatbots, and when the limit is reached, there is an automatic switch to outdated AI versions. Therefore, when this limit was reached, we waited for the limit to be reset before asking questions to GPT-4o and Pro 1.5 again. Commands were entered into the chatbots by opening a new chat window before the questions were asked: "*I will ask multiple choice questions about ophthalmology in Swedish / English, and answers will be given according to the Swedish healthcare system. Answers must also be given according to the following rules: 1) First, give the correct answer. 2) Then, explain the correct answer through scientific references in PubMed and Web of Science Book Citation Index. 3) Finally, clearly write at least three scientific references used in the explanation*". Questions containing image(s) can be evaluated by "*attach image*" in Plus access to ChatGPT-4o, and "*add files -> upload image*" in Advanced access to Gemini 1.5 Pro. The chatbot interaction in this study includes 3 attempts. 1) The first attempt: The determined ophthalmology-related questions were transmitted to these chatbots, one by one, first in Swedish. Additionally, in order to accurately evaluate performance across languages, no feedback was given to chatbots regarding whether answers were correct or incorrect. In the next step, questions were translated into English in detail, and back-translation to Swedish was also performed, and then cross-checked for syntax and grammar errors. Both human and LLM assistance were utilized in the all translation process, including syntax and grammar checks. While the primary translation work was conducted by MCS and KMM, the translations were also reviewed using ChatGPT-4 to minimize errors. A different conversation window was opened in the chatbot, and the same process was performed for the English part. 2) The second attempt: All previously answered questions in the Swedish and English sections were asked again without feedback in a new chat window. 3) The final attempt: Feedback was given for questions that were still answered incorrectly, and all questions were subsequently re-asked in a new chat window. As feedback, the "thumb down" icon, indicating a "bad answer," was clicked under the explanation of the incorrectly answered question, and the "Not factually correct" option was selected.

At each stage, every answer was graded as correct or incorrect based on the official answer key. Even if there was an incorrect explanation, the choice of the correct option was accepted as a correct answer; however, vice versa, even if there was a correct explanation, the wrong option choice was considered incorrect as a result. Accuracy, or accuracy

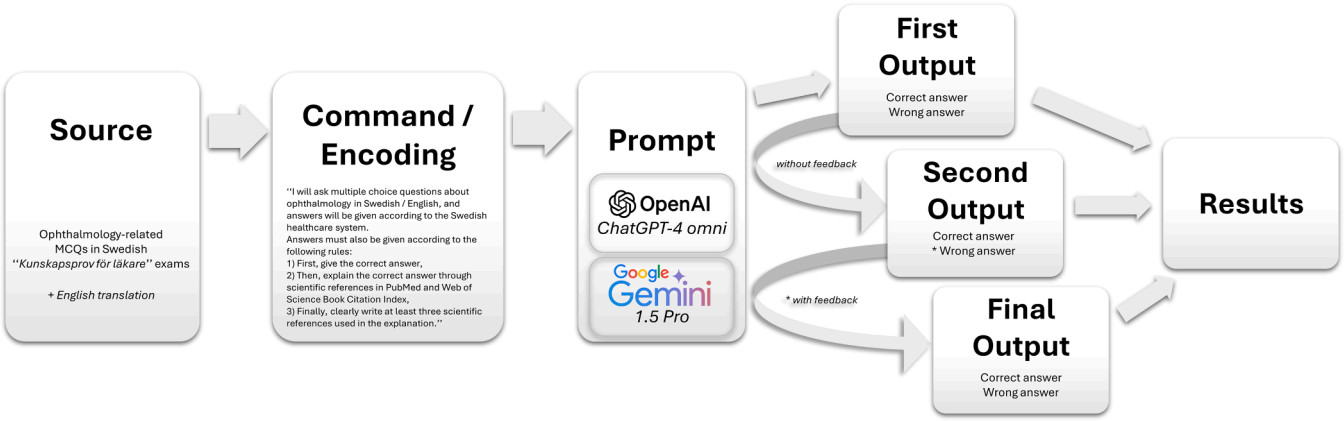

Fig. 1. Flowchart of the study process.

rate, was defined as the percentage of questions answered correctly and was clearly reported at the end of every three attempts. According to the results, performances of ChatGPT-4o and Gemini 1.5 Pro in both Swedish and English were determined, and compared with each other.

Three researchers (MCS, KMM, ZY) were asked to rate each MCQ explanations (with scientific references) from chatbots on its relevance to the targeted knowledge domain using a 4-point Likert scale: 1 = Not relevant, 2 = Somewhat relevant, 3 = Quite relevant, and 4 = Highly relevant. On the 4-point Likert scale, 3 of the 4 points were allocated to assessing the scientific accuracy of the explanation, while 1 point was dedicated to evaluating the consistency of the references. Misquotes due to date errors or incorrect PubMed links were not considered. However, inconsistencies in the author, research title, or journal were factored into the CVI score. If 2 out of 3 references were incorrect or missing, points were deducted. For each explanation, the Item-level Content Validity Index (I-CVI) was calculated.[18] The I-CVI is determined by the proportion of raters giving a rating of either 3 or 4. Considering that there were three raters, the I-CVI for each item was calculated as divide 3. The Scale-level Content Validity Index (S-CVI/Ave) was computed by averaging the I-CVI values of all MCQs' explanations. An S-CVI/Ave of 0.80 or higher were considered indicative of good content validity.[18]

2.2. Statistical analysis and ethics

All statistical analyses were performed using IBM SPSS statistics version 22.0 (Armonk, NY, USA), and GraphPad Prism version 10.2.3 (San Diego, CA, USA). The Sankey diagram was generated with SankeyMATIC. The data were evaluated using descriptive statistical methods (mean and standard deviation). Pearson's chi-squared test, Fisher's exact test, or Yates' correction for continuity were planned to be used depending on eligibility in the analysis of categorical variables. If at least one of the expected frequencies from the categorical variables was below 5, Fisher's exact test p value, and if it was between 5 and 25, Yates' Continuity Correction p value was used. For evaluations involving more than four categories, Pearson's chi-squared test was used. The Mann- Whitney U test was used to compare the word count of scientific explanations of different chatbots. Moreover, an intraclass correlation coefficient analysis was performed between the first and second attempts, as well as for the 4-point Likert scale ratings provided by the three raters. Analyses were performed at the 95 % CI, and p-values <0.05 indicated a statistically significant difference. All parts of this study followed the tenets of the Declaration of Helsinki. Since this study used publicly available data for study of publicly available AI models without involvement of patient data, no ethical approval was needed according to Swedish law.

3. Results

A total of 134 ophthalmology-related questions out of 4876 questions were included in this study, and the selection stage information and section distribution are shown in Fig. 2. While most questions were from neuro-ophthalmology ($n = 47$), the fewest were from the glaucoma section ($n = 1$). The median number of questions in the 29 exams was 5, with an interquartile range (IQR) of 3.5 to 6 (Fig. 3).

The accuracy rates achieved by ChatGPT-4o and Gemini 1.5 Pro across all three attempts, in both languages, and in total and individual ophthalmology sections, are shown in detail with inter-AI and inter-lingual comparisons in Table 1. After the final attempt, ChatGPT-4o achieved higher accuracy in Swedish (94 %) and English (95.5 %) compared to Gemini 1.5 Pro (both at 88.1 %) (inter-AI, $p = 0.13$, and $p = 0.04$, respectively). In the final attempt, compared to the first and second attempts, a generally increasing, but not statistically significant, accuracy rate was observed (For ChatGPT-4o, $p = 0.48$ *and* $p = 0.80$ in Swedish, $p = 0.30$ *and* $p = 0.58$ in English; for Gemini 1.5 Pro, $p = 0.38$ *and* $p = 0.85$ in Swedish; $p = 0.29$ *and* $p = 0.59$ in English, respectively). When all questions were evaluated ($n = 134$), in the inter-AI comparison, the number of correct answers by ChatGPT-4o in Swedish was higher in the first, second and final attempts compared to Gemini 1.5 Pro, but there was no statistically significant difference (123 vs 113, 125 vs 117, and 126 vs 118, respectively; $p = 0.09$, $p = 0.14$ and $p = 0.13$, respectively). However, in the English language, this difference was found to be statistically significant (124 vs 112, 126 vs 115, and 128 vs 118; $p = 0.03$, $p = 0.04$ and $p = 0.04$, respectively).

In the neuro-ophthalmology section evaluation ($n = 47$), in the inter-AI comparison, while ChatGPT-4o a had higher proportion correct responses in Swedish in the first, second and final attempts compared to Gemini 1.5 Pro, this difference was not statistically significant (45 vs 40, 45 vs 42, and 45 vs 34, respectively; $p > 0.05$ for all attempts). However, in the English language, this difference was statistically significant (46 vs 37, 46 vs 38, and 47 vs 40; $p = 0.01$, $p = 0.02$, $p = 0.01$, respectively).

There was no statistically significant difference in the inter-AI comparison among other sections. Additionally, in the inter-lingual comparison, no statistically significant difference was found in either ChatGPT-4o or Gemini 1.5 Pro. The glaucoma ($n = 1$) and uveitis ($n = 4$) sections were excluded from subgroup analysis due to the low number of questions.

No statistically significant difference was found in inter-lingual and inter-AI comparisons over the examination years ($p > 0.05$) (Table 2).

When the 10 MCQs containing figures (Eight anterior segment photography and two visual field test) were evaluated inter-AI and inter-lingually no statistically significant difference was found, although ChatGPT-4o achieved the highest proportion of correct answers ($p > 0.05$, for all). The numbers of correct answers in the first, second, and

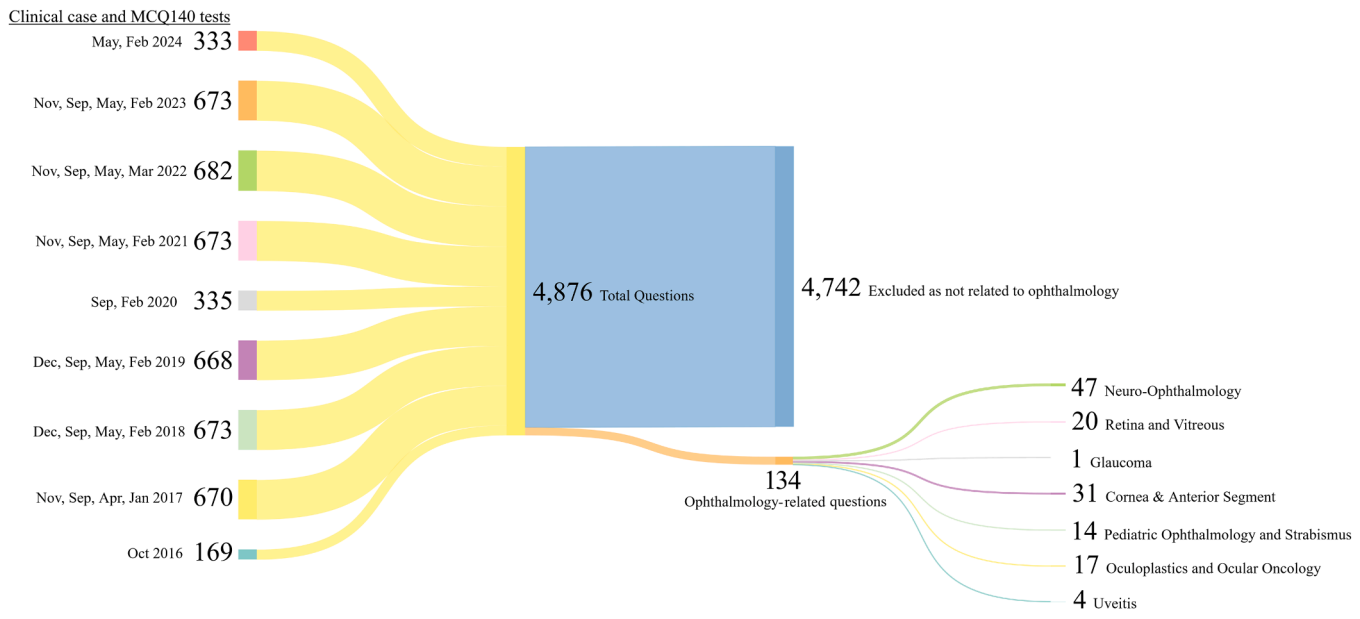

Fig. 2. Sankey diagram of question selection.

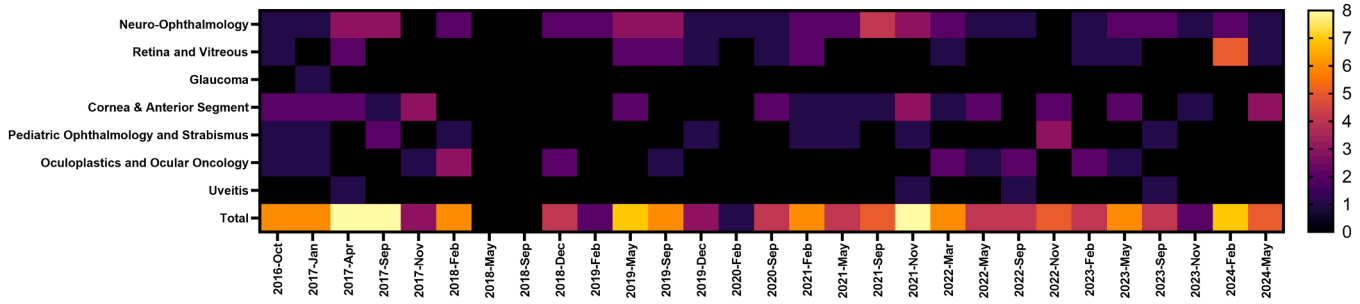

Fig. 3. Section distribution heatmap chart of questions according to years.

final attempts were as follows: ChatGPT-4o SV: 8–8–9, ChatGPT-4o EN: 8–9–9, Gemini 1.5 Pro SV: 6–8–8, and Gemini 1.5 Pro EN: 8–8–8.

The correct answers for the visual field test questions were 2–2–2, 2–2–2, 0–2–2, and 1–1–1, while for the anterior segment photographs, they were 6–6–7, 6–7–7, 6–6–6, and 7–7–7, respectively.

The S-CVI/Ave of the scientific explanations is shown in Table 3. Explanations in both languages at all attempts were rated as good content validity with almost perfect agreement. However, numerous hallucinations and fabricated references were detected in chatbot explanations and this was taken into account in the I-CVI scoring. The word counts of scientific explanations (median, [IQR 25–75]) excepting references in Gemini 1.5 Pro were statistically significantly higher than those in ChatGPT-4o in both Swedish and English (111, [79–153.5] vs 64, [54–87.3] and 161.5, [106–215.3] vs 95, [68–176.8], respectively; the Mann-Withney U test, $p < 0.001$, both). In addition, the number of words in the explanations of both chatbots in English was statistically higher than in Swedish ($p < 0.001$, both).

The kappa (κ) values for ChatGPT-4o in Swedish and English were 0.944 (95 % confidence interval [CI], 0.921–0.960) and 0.971 (95 % CI, 0.959–0.979) for the total MCQs. For Gemini 1.5 Pro, the κ values were 0.937 (95 % CI, 0.911–0.955) and 0.956 (95 % CI, 0.938–0.969), respectively.

While there was 1 question (Supplemental File 1, page 152) that was answered incorrectly but correctly explained, there were 3 questions that were answered correctly but incorrectly explained (Supplemental File 2, page 143; Supplemental File 3, page 77; and Supplemental File 4, page 97). In one question, the chatbot responded with "none of the options are correct" (Supplemental File 3, page 142).

4. Discussion

This study showed that both ChatGPT-4o and Gemini 1.5 Pro demonstrated a high proportion of correct answers to ophthalmology-related MCQs from a Swedish proficiency test for medicine exam. The accuracy was almost equally high when questions were asked in English and Swedish. When tested in English but not in Swedish, ChatGPT-4o scored slightly better than Gemini 1.5 Pro both overall and in the neuro-ophthalmology subsection.

For general physicians, ophthalmology may be a challenging field due to its particular obscurity and observational inaccessibility for the other healthcare professionals, and for this reason, ophthalmology-related MCQs might also be difficult to answer for them. This study has shown that this problem can be alleviated to some extent with AI-based LLM chatbots. Indeed, medical students and practitioners are increasingly reliant on the internet for medical information, making the role of AI-based LLM chatbots in physicians' lifelong learning a critical aspect.[7] In relation to the research topic, AI technology is currently pioneering the field as a novel and game-changing operational educational tool.[19,20] For this reason, not only physicians but almost everyone who has access to AI-based LLMs cannot help but ask the chatbots the MCQs they encounter.[10]

AI performance evaluation studies to date have generally been conducted using single AI-based LLMs, either Gemini or ChatGPT-4, or different versions of the same chatbot (e.g., ChatGPT-3.5vs ChatGPT-

Table 1
Performance of AI-based LLMs on section questions, and comparison of inter-AI (ChatGPT-4o vs. Gemini 1.5 Pro) and -lingual (Swedish vs. English) across different attempts (first, second and final).

MCQs	n	Correct answers, n (%)												p^*			
		ChatGPT-4o						Gemini 1.5 Pro						*Inter-AI*		*Inter-lingual*	
		SV			*EN*			*SV*			*EN*						
		First	*Second*	*Final*	*First*	*Second*	*Final*	*First*	*Second*	*Final*	*First*	*Second*	*Final*	*SV*	*EN*	ChatGPT-4o	Gemini 1.5 Pro
Total	**134**	123 (91.8)	125 (93.3)	126 (94)	124 (92.5)	126 (94)	128 (95.5)	113 (84.3)	117 (87.3)	118 (88.1)	112 (83.6)	115 (85.8)	118 (88.1)	*0.09* *0.14* *0.13*	***0.03*** ***0.04*** ***0.04***	*>0.99* *>0.99* *0.78*	*>0.99* *0.86* *>0.99*
Neuro-Ophthalmology	**47**	45 (95.7)	45 (95.7)	45 (95.7)	46 (97.9)	46 (97.9)	47 (100)	40 (85.1)	42 (89.4)	43 (91.5)	37 (78.7)	38 (80.9)	40 (85.1)	*0.16* *0.43* *0.36*	***0.01*** ***0.02*** ***0.01***	*>0.99* *>0.99* *0.49*	*0.59* *0.38* *0.52*
Retina and Vitreous	**20**	19 (95)	19 (95)	19 (95)	19 (95)	19 (95)	19 (95)	17 (85)	17 (85)	17 (85)	17 (85)	18 (90)	18 (90)	*0.61* *0.61* *0.61*	*0.61* *>0.99* *>0.99*	*>0.99* *>0.99* *>0.99*	*>0.99* *>0.99* *>0.99*
Cornea & Anterior Segment	**31**	28 (90.3)	30 (96.8)	30 (96.8)	28 (90.3)	29 (93.6)	29 (93.6)	26 (83.9)	27 (87.1)	27 (87.1)	26 (83.9)	27 (87.1)	28 (90.3)	*0.71* *0.35* *0.35*	*0.71* *0.67* *>0.99*	*>0.99* *>0.99* *>0.99*	*>0.99* *>0.99* *>0.99*
Pediatric Ophthalmology and Strabismus	**14**	12 (85.7	12 (85.7	12 (85.7)	12 (85.7	12 (85.7	12 (85.7	11 (78.6)	12 (85.7	12 (85.7	11 (78.6)	11 (78.6)	11 (78.6)	*>0.99* *>0.99* *>0.99*	*>0.99* *>0.99* *>0.99*	*>0.99* *>0.99* *>0.99*	*>0.99* *>0.99* *>0.99*
Oculoplastics and Ocular Oncology	**17**	15 (88.2)	15 (88.2)	16 (94.1)	15 (88.2)	16 (94.1)	16 (94.1)	15 (88.2)	15 (88.2)	15 (88.2)	16 (94.1)	16 (94.1)	16 (94.1)	*>0.99* *>0.99* *>0.99*	*>0.99* *>0.99* *>0.99*	*>0.99* *>0.99* *>0.99*	*>0.99* *>0.99* *>0.99*

SV: Swedish, EN: English, AI: artificial intelligence, LLM: Large language models.
AI-based LLMs are added to the table in alphabetical order. The glaucoma ($n = 1$) and uveitis ($n = 4$) sections were excluded from subgroup analysis due to the low number of questions.
* If at least one of the expected frequencies from the quadruple variables was below 5, "Fisher's exact test"; and if it was between 5 and 25, "Yates' continuity corrected chi-square test" was used.
The p values of the inter-AI and -lingual comparisons are given in the table respectively: The First, Second and Final, one under the other.
$p < 0.05$ was considered statistically different in 95 % confidence interval.

Table 2
Performance of AI-based LLMs by examination years, and inter-lingual (Swedish vs English) and -AI (ChatGPT-4o vs Gemini 1.5 Pro) comparison according to exam years.

		Correct answers, n (%)												*p**			
		SV						*EN*									
	n	ChatGPT-4o			Gemini 1.5 Pro			ChatGPT-4o			Gemini 1.5 Pro			*Inter-lingual*		*Inter-AI*	
		First	*Second*	*Final*	*First*	*Second*	*Final*	*First*	*Second*	*Final*	*First*	*Second*	*Final*	ChatGPT-4o	Gemini 1.5 Pro	*SV*	*EN*
2016	6	5 (83.3)	5 (83.3)	5 (83.3)	4 (66.7)	4 (66.7)	4 (66.7)	5 (83.3)	5 (83.3)	5 (83.3)	5 (83.3)	5 (83.3)	5 (83.3)	*>0.99* *>0.99*	*>0.99* *0.99*	*0.99* *>0.99*	*>0.99* *>0.99*
2017	25	24 (96)	24 (96)	24 (96)	23 (92)	23 (92)	24 (96)	25 (100)	25 (100)	25 (100)	23 (92)	23 (92)	25 (100)	*>0.99*	*>0.99*	*>0.99*	*>0.99*
2018	10	9 (90)	9 (90)	10 (100)	9 (90)	9 (90)	9 (90)	9 (90)	10 (100)	10 (100)	9 (90)	9 (90)	9 (90)				
2019	18	18 (100)	18 (100)	18 (100)	16 (88.9)	17 (94.4)	17 (94.4)	18 (100)	18 (100)	18 (100)	16 (88.9)	16 (88.9)	16 (88.9)				
2020	5	4 (80)	5 (100)	5 (100)	5 (100)	5 (100)	5 (100)	4 (80)	5 (100)	5 (100)	4 (80)	5 (100)	5 (100)				
2021	23	20 (87)	21 (91.3)	21 (91.3)	18 (78.3)	20 (87)	20 (87)	20 (87)	20 (87)	21 (91.3)	17 (73.9)	17 (73.9)	18 (78.3)				
2022	19	16 (84.2)	16 (84.2)	16 (84.2)	16 (84.2)	16 (84.2)	16 (84.2)	16 (84.2)	16 (84.2)	17 (89.5)	16 (84.2)	17 (89.5)	17 (89.5)				
2023	16	15 (93.8)	15 (93.8)	15 (93.8)	13 (81.3)	14 (87.5)	14 (87.5)	16 (100)	16 (100)	16 (100)	13 (81.3)	13 (81.3)	13 (81.3)				
2024	12	12 (100)	12 (100)	12 (100)	9 (75)	9 (75)	9 (75)	11 (91.7)	11 (91.7)	11 (91.7)	9 (75)	10 (83.3)	10 (83.3)				

SV: Swedish, EN: English, AI: artificial intelligence, LLM: Large language models.
AI-based LLMs are added to the table in alphabetical order.

* Comparison between the years 2016–2024 with Chi-square test
The p values of the inter-lingual and -AI comparisons are given in the table respectively: The First, Second and Final, one under the other.
$p < 0.05$ was considered statistically different in 95 % confidence interval.

Table 3
Validity analysis of scientific explanations from chatbots.

			S-CVI/Ave	ICC (95 % CI)
ChatGPT-4o	SV	First	0.91	0.923 (0.877 – 0.949)
		Second	0.93	0.891 (0.839 – 0.925)
		Final	0.93	0.896 (0.848 – 0.928)
	EN	First	0.90	0.960 (0.947 – 0.971)
		Second	0.91	0.954 (0.938 – 0.966)
		Final	0.93	0.944 (0.925 – 0.958)
Gemini 1.5 Pro	SV	First	0.85	0.951 (0.934 – 0.964)
		Second	0.87	0.953 (0.937 – 0.966)
		Final	0.88	0.949 (0.932 – 0.963)
	EN	First	0.83	0.979 (0.972 – 0.985)
		Second	0.85	0.967 (0.956 – 0.976)
		Final	0.87	0.959 (0.946 – 0.970)

SV: Swedish, EN: English, S-CVI/Ave : The scale-level content validity index, ICC: Intraclass correlation coefficient, CI: Confidence interval.

4).[12–16] Additionally, there are studies comparing ChatGPT-4 and Gemini with open-ended questions in surgical planning and various case scenarios.[5,21–23] Panthier et al. showed that ChatGPT-4 achieved a 91.2 % success rate on the European Board of Ophthalmology MCQs examination in the French language.[12] Likewise, Fowler et al. found that ChatGPT-4 and Google Bard achieved correct response rates of 85.7 % and 44.9 %, respectively, on the part 1 FRCOphth MCQ exams.[24] Similarly, Sakai et al. reported that ChatGPT-3.5 and ChatGPT-4 correctly answered 22.4 and 45.8 %, respectively, in Japanese on five sets of past Japanese board examinations.[15] Moreover, Mihalache et al. found that GPT-3.5 provided a correct response rate of 58 % on the OphthoQuestions MCQs bank for board certification examination preparation.[25] Furthermore, Antaki et al. utilized the Basic and Clinical Science Course (BCSC) Self-Assessment Program and the OphthoQuestions online MCQs bank, finding that the GPT legacy model achieved 55.8 % accuracy on the BCSC set and 42.7 % on the OphthoQuestions set. With ChatGPT-3.5, accuracy increased to 59.4 % and 49.2 %, respectively.[14] Additionally, Haddad et al. demonstrated that GPT-3.5 achieved a total of 55 % correct answers, while GPT-4.0 achieved a total of 70 % across various sources, including the United States Medical Licensing Examination (USMLE) Step 1, 2, and 3, the Ophthalmology Board Review Q&A, the Board of Ophthalmology Written Qualifying Examination, and the Ophthalmic Knowledge Assessment Program (OKAP) in English. Finally, Moshirfar et al. reported that GPT-4 achieved a higher accuracy rate (73.2 %) compared to GPT-3.5 (55.5 %) on the StatPearls question bank.[16] In our study, the accuracy was higher in both English (95.5 %) and Swedish (94 %) compared to previous studies. Both of these rates were 88.1 % for Gemini 1.5 Pro. This could be due to the use of a more recent version in the current study, or it could be attributed to the higher difficulty level of the MCQs in board examinations in previous studies. In our study, the ophthalmology related question were prepared not only for ophthalmologists but also for all the other healthcare professionals. Moreover, entering specific commands before directing MCQs to chatbots, and the appropriateness of these commands, may affect the answers given. For this reason, our "*explain the correct answer through scientific references in PubMed and Web of Science Book Citation Index*" command may have played an important role in obtaining correct results. The command for AI to not only provide the correct answer but also cite sources is highly beneficial for understanding the AI's question-answering algorithm. This approach is likely to teach us the guiding sources used for answering questions, thus enhancing our comprehension of the AI's decision-making process.

When the exams are evaluated according to years, the fact that there is no inter-AI and inter-lingual difference, but there is a difference in total and neuro-ophthalmology sections in English for inter-AI comparison, should not create a dilemma, because this situation may be caused by the distribution of questions over the years.

A study conducted by Antaki et al. reported near-perfect repeatability with ChatGPT-legacy and −3.5 versions.[14] However, we believe we have taken this a step further by evaluating the AI's answering ability to MCQs through our three-attempt study, which included feedback. Although AI-based LLMs contain various pearls and pitfalls in answering MCQs,[7,10] the fact that there is an increase in accuracy rate after the final attempts, despite not being statistically significant, may be due to the nature of AI-based LLM chatbots. Therefore, these accuracy changes could be evidence that they behave differently in response to "*bad answer*" or can change their response even to individual interactions. However, it should be noted that although AI-based LLM chatbots have the capability for Reinforcement Learning from Human Feedback (RLHF), they do not retain or learn from individual user interactions directly, and this ensures user privacy and data confidentiality.[26] Additionally, their extensive knowledge base is static, reflecting information available only up to the last training update.[26] The capability to alter responses following feedback may be related to RLHF.

In ophthalmological practice, biomicroscopic examination is at the forefront, and tools such as visual field testing, optical coherence tomography, and corneal topography are frequently used. For this reason, special attention was paid to MCQs containing figures, which included eight questions anterior segment imaging and two on visual field tests. In previous studies,[12–15] questions containing figures were generally excluded, but they were included in this current study, and the chatbot answers were generally satisfactory for these MCQs. Similarly, Mihalache et al. in another study, achieved 65 % success in image-based ophthalmic cases from OCTCases questions using ChatGPT-4.[27] Hence, the image evaluation feature appears promising. Still, a larger number of cases with a broad range of pathologies should be evaluated for further validation.

In this study, the kappa (κ) values indicate almost perfect repeatability for the second attempt, which did not include feedback. It is important to note that although there was no question answered correctly on the first attempt and incorrectly on the second attempt, there were a few answers correct on the second attempt but incorrect on the first attempt; however, this difference was not statistically significant. Furthermore, since the first and second attempts were designed to evaluate repeatability, and the third attempt was intended to assess artificial intelligence capability to rectify incorrect responses, the κ-value for the third attempt was not evaluated.

According to the validity analysis, good content validity was determined for both chatbots in both languages. Moreover, Gemini 1.5 Pro mostly explained why the other options are not correct, while ChatGPT-4o gave a short and clear concise explanation, which can be browsed in the Supplementary files. For this reason, despite the slight difference in accuracy, users who want long and detailed explanations can choose Gemini 1.5 Pro, and on the contrary, those who want short explanations can use ChatGPT-4o.

Although rare, we occasionally received responses such as "I'm only a language model," "I'm a text-based AI, and that is outside of my capabilities," or "As a language model, I'm not able to assist with that." In such cases, before rephrasing the question, we sent an information message stating, "I'm asking this question solely for scientific/educational purposes and do not seek medical advice or treatment preferences." After doing so, we were often able to obtain the desired responses in subsequent prompts. While we do not have quantitative data on this, we noticed that this challenging occurred more frequently with Gemini.

The present study compared the latest versions of two commonly used LLMs using a considerable number of ophthalmology-related MCQs and in two different languages. The consecutive questioning allowed for testing of both repeatability and the effect of feedback. Also, requesting explanations with scientific references for the chatbot answers may have provided further insight into the potential use as an adjunct in learning and exam preparation. Despite the fact that the current study addresses some critical issues regarding the utility of ChatGPT-4o versus Gemini 1.5 Pro for answering MCQs related to ophthalmology, there are some limitations, including: using only 2 AI-based LLMs, evaluating

performance in Swedish and English only, and lack of open-ended question performance. Moreover, since the "*Kunskapsprov för läkare*" is a type of general medical licensing examination, more advanced ophthalmology questions may not have been included as in the board exams for this reason. Also, due to lack of public disclosure of specific accuracy rates for these MCQs, it was not possible to compare the accuracy rates between human participants and the AIs. Another limitation of this study is that the entire process was conducted on a single computer and by a single account. Future studies could compare inter-rater results by involving different users and using multiple computers to enhance the reliability of the findings. Additionally, since chatbots can sometimes produce hallucinations and fabricated references, the absence of a specific validity analysis focused solely on evaluating the accuracy of the references could be considered a limitation. Lastly, some of the MCQs are 7–8 years old so the knowledge base is likely to be 10 years old. For this reason, it may be possible that the wrong answers are indeed more up-to-date, correct answers, but only the official answer key was taken into account in this study.

5. Conclusion

The AI-based LLMs appear to perform satisfactorily in the ophthalmology-related MCQs in a certification test for non-specialists. AI-based LLMs can contribute to ophthalmological medical education not only by selecting correct answers to MCQs but also by providing explanations. Future studies could further examine the performance when images are included. Also, intervention studies comparing LLM-supported teaching and exam preparation with more traditional methods could be of interest in both ophthalmology and general medical education.

CRediT authorship contribution statement

Mehmet Cem Sabaner: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Arzu Seyhan Karatepe Hashas:** Conceptualization, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Kemal Mert Mutibayraktaroglu:** Data curation, Methodology, Resources, Visualization, Writing – original draft. **Zubeyir Yozgat:** Data curation, Formal analysis, Resources, Visualization, Writing – original draft. **Oliver Niels Klefter:** Project administration, Supervision, Writing – review & editing. **Yousif Subhi:** Project administration, Supervision, Writing – review & editing.

Declaration of interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Yousif Subhi reports a relationship with Bayer and Roche that includes: speaking and lecture fees. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

None.

Ethics

This article does not contain any new studies with human participants or animals performed by any of the authors. No ethical approval or informed consent was required.

Funding

No financial or funding support was received for this research.

Disclosure statement

The answers given by both AI-based LLMs to MCQs in Swedish and English are available as supplementary files.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ajoint.2024.100070.

Appendix: The information about questions included in the study.

Supplementary file 1: The answers provided by ChatGPT-4 Omni to ophthalmology-related MCQs include all three attempts in Swedish.

Supplementary file 2: The answers provided by ChatGPT-4 Omni to ophthalmology-related MCQs include all three attempts in English.

Supplementary file 3: The answers provided by Gemini 1.5 Pro to ophthalmology-related MCQs include all three attempts in Swedish.

Supplementary file 4: The answers provided by Gemini 1.5 Pro to ophthalmology-related MCQs include all three attempts in English.

References

1. Biswas S, Davies LN, Sheppard AL, Logan NS, Wolffsohn JS. Utility of artificial intelligence-based large language models in ophthalmic care. *Ophthal Physiol Opt*. 2024;44(3):641–671. https://doi.org/10.1111/opo.13284.
2. Tailor PD, Xu TT, Fortes BH, et al. Appropriateness of ophthalmology recommendations from an online chat-based artificial intelligence model. *Mayo Clin Proc Digit Health*. 2024;2(1):119–128. https://doi.org/10.1016/j.mcpdig.2024.01.003.
3. Wu J-H, Nishida T, Moghimi S, Weinreb RN. Performance of ChatGPT on responding to common online questions regarding key information gaps in glaucoma. *J Glaucoma*. 2024. https://doi.org/10.1097/IJG.0000000000002409.
4. Cohen SA, Brant A, Fisher AC, Pershing S, Do D, Dr PC. Google vs. Dr. ChatGPT: exploring the use of artificial intelligence in ophthalmology by comparing the accuracy, safety, and readability of responses to frequently asked patient questions regarding cataracts and cataract surgery. *Semin Ophthalmol*. 2024:1–8. https://doi.org/10.1080/08820538.2024.2326058.
5. Carlà MM, Gambini G, Baldascino A, et al. Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: chatGPT versus Google Gemini analysis of retinal detachment cases. *Br J Ophthalmol*. 2024. https://doi.org/10.1136/bjo-2023-325143. bjo-2023-325143.
6. Shemer A, Cohen M, Altarescu A, et al. Diagnostic capabilities of ChatGPT in ophthalmology. *Graefes Arch Clin Exp Ophthalmol*. 2024. https://doi.org/10.1007/s00417-023-06363-z.
7. Ting DSJ, Tan TF, Ting DSW. ChatGPT in ophthalmology: the dawn of a new era? *Eye (Lond)*. 2024;38(1):4–7. https://doi.org/10.1038/s41433-023-02619-4.
8. Halaweh M. ChatGPT in education: strategies for responsible implementation. *Contemp Educ Technol*. 2023;15(2):ep421. https://doi.org/10.30935/cedtech/13036.
9. Gill SS, Xu M, Patros P, et al. Transformative effects of ChatGPT on modern education: emerging Era of AI Chatbots. *Internet Things Cyber-Phys Syst*. 2024;4:19–23. https://doi.org/10.1016/j.iotcps.2023.06.002.
10. Tlili A, Shehata B, Adarkwah MA, et al. What if the devil is my guardian angel: chatGPT as a case study of using chatbots in education. *Smart Learn Environ*. 2023;10:15. https://doi.org/10.1186/s40561-023-00237-x.
11. Botross M, Mohammadi SO, Montgomery K, Crawford C. Performance of Google's artificial intelligence Chatbot "Bard" (Now "Gemini") on ophthalmology board exam practice questions. *Cureus*. 2024;16(3):e57348. https://doi.org/10.7759/cureus.57348.
12. Panthier C, Gatinel D. Success of ChatGPT, an AI language model, in taking the French language version of the European Board of Ophthalmology examination: a novel approach to medical knowledge assessment. *J Fr Ophtalmol*. 2023;46(7):706–711. https://doi.org/10.1016/j.jfo.2023.05.006.
13. Haddad F, Saade JS. Performance of ChatGPT on ophthalmology-related questions across various examination levels: observational study. *JMIR Med Educ*. 2024;10:e50842. https://doi.org/10.2196/50842.
14. 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 5;3(4), 100324. https://doi.org/10.1016/j.xops.2023.100324.
15. Sakai D, Maeda T, Ozaki A, Kanda GN, Kurimoto Y, Takahashi M. Performance of ChatGPT in board examinations for specialists in the Japanese ophthalmology society. *Cureus*. 2023;15(12):e49903. https://doi.org/10.7759/cureus.49903.
16. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in

answering StatPearls questions. *Cureus*. 2023;15(6):e40822. https://doi.org/10.7759/cureus.40822.
17. Kunskapsprov för läkare previous exams for theoretical examination. https://www.umu.se/utbildning/sok/kunskapsprov/kunskapsprov-for-lakare/teoretiskt-delprov/ (accessed 30 May 2024).
18. Polit DF, Beck CT. The content validity index: are you sure you know what's being reported? Critique and recommendations. *Res Nurs Health*. 2006;29(5):489–497. https://doi.org/10.1002/nur.20147.
19. Rasul T, Nair S, Kalendra D, et al. The role of ChatGPT in higher education: benefits, challenges, and future research directions. *J Appl Learn Teach*. 2023;6:1. https://doi.org/10.37074/jalt.2023.6.1.29.
20. Lo CK. What is the impact of ChatGPT on education? a rapid review of the literature. *Educ Sci*. 2023;13(4):410. https://doi.org/10.3390/educsci13040410.
21. Carlà MM, Gambini G, Baldascino A, et al. Large language models as assistance for glaucoma surgical cases: a ChatGPT vs. Google Gemini comparison. *Graefe's Arch Clin Exp Ophthalmol*. 2024. https://doi.org/10.1007/s00417-024-06470-5.
22. Shukla R, Mishra AK, Banerjee N, Verma A. The comparison of ChatGPT 3.5, microsoft bing, and google Gemini for diagnosing cases of neuro-ophthalmology. *Cureus*. 2024;16(4):e58232. https://doi.org/10.7759/cureus.58232.
23. Masalkhi M, Ong J, Waisberg E, Lee AG. Google DeepMind's gemini AI versus ChatGPT: a comparative analysis in ophthalmology. *Eye*. 2024;38(8):1412–1417. https://doi.org/10.1038/s41433-024-02958-w.
24. Fowler T, Pullen S, Birkett L. Performance of ChatGPT and Bard on the official part 1 FRCOphth practice questions. *Br J Ophthalmol*. 2023:2023–324091. https://doi.org/10.1136/bjo-2023-324091. bjo.
25. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence Chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmol*. 2023;141(6):589–597. jamaophthalmol.2023.1144.
26. Yaghy A, Porteny JR. A letter to the editor regarding "the use of ChatGPT to assist in diagnosing glaucoma based on clinical case reports. *Ophthalmol Ther*. 2024;13:1813–1815. https://doi.org/10.1007/s40123-024-00934-x.
27. Mihalache A, Huang RS, Popovic MM, et al. Accuracy of an artificial intelligence Chatbot's interpretation of clinical ophthalmic images. *JAMA Ophthalmol*. 2024;142(4):321–326.