Performance of ChatGPT-4 Omni and Gemini 1.5 Pro on Ophthalmology-Related Questions in the Turkish Medical Specialty Exam

Mehmet Cem Sabaner, Zübeyir Yozgat

Kastamonu University Faculty of Medicine; Kastamonu Training and Research Hospital, Department of Ophthalmology, Kastamonu, Türkiye

Abstract

Objectives: To evaluate the response and interpretative capabilities of two pioneering artificial intelligence (AI)-based large language model (LLM) platforms in addressing ophthalmology-related multiple-choice questions (MCQs) from Turkish Medical Specialty Exams.

Materials and Methods: MCQs from a total of 37 exams held between 2006-2024 were reviewed. Ophthalmology-related questions were identified and categorized into sections. The selected questions were asked to the ChatGPT-4o and Gemini 1.5 Pro AI-based LLM chatbots in both Turkish and English with specific prompts, then re-asked without any interaction. In the final step, feedback for incorrect responses were generated and all questions were posed a third time.

Results: A total of 220 ophthalmology-related questions out of 7312 MCQs were evaluated using both AI-based LLMs. A mean of 6.47±2.91 (range: 2-13) MCQs was taken from each of the 33 parts (32 full exams and the pooled 10% of exams shared between 2022 and 2024). After the final step, ChatGPT-4o achieved higher accuracy in both Turkish (97.3%) and English (97.7%) compared to Gemini 1.5 Pro (94.1% and 93.2%, respectively), with a statistically significant difference in English (p=0.039) but not in Turkish (p=0.159). There was no statistically significant difference in either the inter-AI comparison of sections or interlingual comparison.

Conclusion: While both AI platforms demonstrated robust performance in addressing ophthalmology-related MCQs, ChatGPT-4o was slightly superior. These models have the potential to enhance ophthalmological medical education, not only by accurately selecting the answers to MCQs but also by providing detailed explanations.

Keywords: Artificial intelligence, large language model, ChatGPT-4 Omni, Gemini 1.5 Pro, medical education, ophthalmology, e-learning

Cite this article as: Sabaner MC, Yozgat Z. Performance of ChatGPT-4 Omni and Gemini 1.5 Pro on Ophthalmology-Related Questions in the Turkish Medical Specialty Exam.
Turk J Ophthalmol. 2025;55:177-185

Address for Correspondence: Mehmet Cem Sabaner, Kastamonu University Faculty of Medicine; Kastamonu Training and Research Hospital, Department of Ophthalmology, Kastamonu, Türkiye
E-mail: drmcemsabaner@yahoo.com **ORCID-ID:** orcid.org/0000-0002-0958-9961

Received: 23.08.2024 **Accepted:** 10.06.2025

DOI: 10.4274/tjo.galenos.2025.27895

Introduction

"Understanding these marvels of our era, the thinking machines, does not necessitate a diabolical intelligence; rather, simple common sense suffices."[1] Developing thinking machines and effectively integrating them into educational and business environments represents a significant breakthrough for humanity. Artificial intelligence (AI) is exhibiting significant potential across education platforms, notably in the medical sciences.[2,3] AI currently contributes to medicine not only by enhancing educational approaches but also by advancing diagnostic and treatment recommendations in contemporary medical practice.[2,3,4,5,6,7,8,9] Although it is widely assumed that AI can almost never fully replace humans, its potential contributions to medicine and medical education are subjects of considerable interest and curiosity. ChatGPT-4o (omni) and Gemini 1.5 Pro are cutting-edge models designed to provide highly reliable and context-sensitive outputs across a broad range of languages. They are mainly categorized as large language models (LLMs), which are advanced deep learning frameworks trained on extensive datasets to assimilate diverse language characteristics.[3,10] AI-based LLM chatbots are now increasingly utilized in numerous fields, notably in digital education, personalized healthcare, autonomous systems, client support, data science, and software engineering.[10]

AI-driven e-learning is swiftly gaining traction, transforming educational paradigms and practices on a global scale.[10] Tasks such as completing homework, conducting research, answering multiple-choice questions (MCQs), and even composing academic theses can now be efficiently managed using AI-based LLMs.[11] However, the reliability and efficacy of these AI-driven approaches remain under scrutiny. Previous research has illustrated the potential of AI in addressing MCQs, emphasizing its role in enhancing the acquisition of accurate information.[6] Nonetheless, there is a dearth of studies specifically evaluating the capabilities of AI chatbots in answering ophthalmology-related MCQs.[12,13,14,15,16,17] Consequently, this assessment aimed to reveal critical insights into how chatbots can be effectively utilized for ophthalmology-related MCQs in both English and Turkish. To this end, the study evaluated the responses of these two chatbots to ophthalmology-related MCQs in the Turkish Medical Specialty Exam (MSE).

Materials and Methods

Study Design and Data Collection

This cross-sectional study evaluated the performance of two AI-based LLM chatbot models in answering ophthalmology-related MCQs obtained from past Turkish MSEs. The MSE (known as TUS in Turkish) is a nationwide standardized exam held twice yearly by Türkiye's Student Selection and Placement Centre (ÖSYM in Turkish) for admission to medical specialty training. The MSE consists of two parts, the basic and clinical medical sciences tests.

All questions from a total of 32 exams held in 2006-2021[18] and a specified 10% of the questions from 5 exams held in 2022-2024[19,20,21] are considered works under the Law on Intellectual and Artistic Works that are copyrighted by ÖSYM and available to the public as open access with restrictions on reproduction, distribution, and re-publication. These questions were reviewed in detail by two senior ophthalmologists. Ophthalmology-related questions were identified by consensus and included in this study. These questions were also classified into ophthalmology subtopics.

The evaluation of the chatbots' capacity to answer ophthalmology-related MCQs was conducted using the most current premium versions of Gemini 1.5 Pro (Google, Mountain View, CA) and ChatGPT-4o (OpenAI, San Francisco, CA), accessed via Gemini Advanced and ChatGPT Plus platforms. The overall interaction process, including input prompts and response evaluation, is outlined in Figure 1. Each chatbot session began with a standardized prompt instructing the model to answer MCQs in either Turkish or English following a three-step format: (1) state the correct answer, (2) justify the answer using scientific sources indexed in the Web of Science (WoS) Citation Index and PubMed, and (3) list a minimum of three cited references. For questions containing visual data, the chatbots' image upload features were utilized. The evaluation was conducted in three distinct attempts.

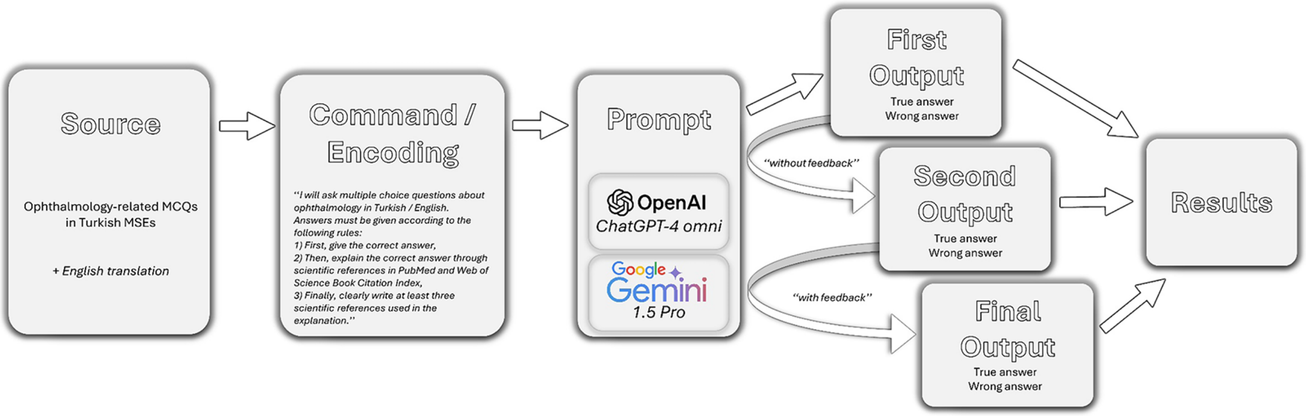
In the first attempt, all selected ophthalmology-related MCQs were presented to the chatbots individually, starting in Turkish. No feedback was given as to whether the responses were correct or incorrect. The same items were then professionally translated into English, followed by back-translation and cross-verification by two researchers. To minimize bias introduced by linguistic structure, answer options were reordered during translation. The English questions were presented to the chatbots individually with no feedback about correctness.

In the second attempt, all previously used questions in both languages were re-entered again without providing feedback.

In the final attempt, chatbot answers that remained incorrect were flagged via the "thumbs down" icon, with the "Not factually correct" reason selected. All questions were then submitted once more for reassessment.

Each attempt was conducted in a separate chatbot session. In each attempt, the correctness of responses was judged solely according to the official answer keys. Selection of the correct option was required for a response to be marked as accurate, regardless of the explanation's quality. Conversely, if an incorrect option was selected—even with a correct explanation—the answer was considered incorrect. For each attempt, the accuracy rate was computed as the percentage of correct answers.

Each explanation generated by the chatbots, including its cited references, was evaluated independently by two senior



"I will ask multiple choice questions about ophthalmology in Turkish / English. Answers must 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."

Figure 1. Flowchart of the study
MCQs: Multiple-choice questions, MSEs: Medical Specialty Exams

ophthalmologists. A 4-point Likert scale was employed to assess the relevance of each response to the intended ophthalmic knowledge: 1 = Not relevant, 2 = Somewhat relevant, 3 = Quite relevant, and 4 = Highly relevant. Of the four points, three were designated to assess the scientific soundness of the explanation, while the remaining point focused on the reliability of the cited references. Minor errors such as incorrect publication dates or faulty hyperlinks were not penalized; however, inconsistencies in author names, article titles, or journal sources were considered during scoring. If two or more references were missing or erroneous, point deductions were applied. The item-level content validity index (I-CVI) was determined by calculating the proportion of raters who assigned a score of 3 or 4 to each item.[22] To assess overall validity, average CVI values were calculated by averaging the I-CVI scores across all items for each attempt by both chatbots. An average CVI value ≥0.80 was interpreted as acceptable content validity, following the criteria established by Polit and Beck.[22]

Informed consent and institutional review board approval were not required for this AI-based LLM chatbot evaluation study.

Statistical Analysis

To analyze the data, statistical evaluations were conducted using GraphPad Prism (v10.2.3, San Diego, CA, USA) and IBM SPSS Statistics software (v22.0, Armonk, NY, USA). The Sankey diagram illustrating question flow and categorization was created using the online tool SankeyMATIC. Descriptive statistics were presented as mean ± standard deviation (SD) or median with interquartile ranges (25th–75th percentile), as appropriate. For categorical variables, Pearson's chi-square test was primarily employed. However, Fisher's exact test or Yates' continuity correction was applied when assumptions of expected frequency counts were not met (i.e., expected cell count <5 or 5-25, respectively). In comparisons involving more than four categorical groups, Pearson's chi-square remained the default method. Differences in the word count of explanations between the two chatbot systems were analyzed using the non-parametric Mann-Whitney U test, given the non-normal distribution of the data. To assess consistency across attempts and rater agreement in the Likert-scale evaluations, intraclass correlation coefficients were calculated. Statistical significance was defined as a p value below 0.05, and all analyses were conducted within a 95% confidence interval (CI).

Results

Of the 7312 MCQs reviewed from 37 past Turkish MSEs, a total of 220 questions were identified as ophthalmology-related and selected for further analysis. Detailed information regarding the question selection process and the subspecialty distribution is visualized in Figure 2. Due to ÖSYM's copyright restrictions, the full text of the questions and answers could not be published. However, details about the questions included are provided in the Appendix 1, and Turkish MSE-like questions and chatbot answer examples are presented in the Supplemental Material. Neuro-ophthalmology was the most frequently represented subspecialty (n=72), while glaucoma and uveitis were the least (n=13 each). Across the evaluated exams (32 full exams and the pooled 10% of exams shared between 2022-2024), the average number of ophthalmology questions was 6.47 (SD: 2.91), with a minimum of 2 and a maximum of 13, as illustrated in Figure 3.

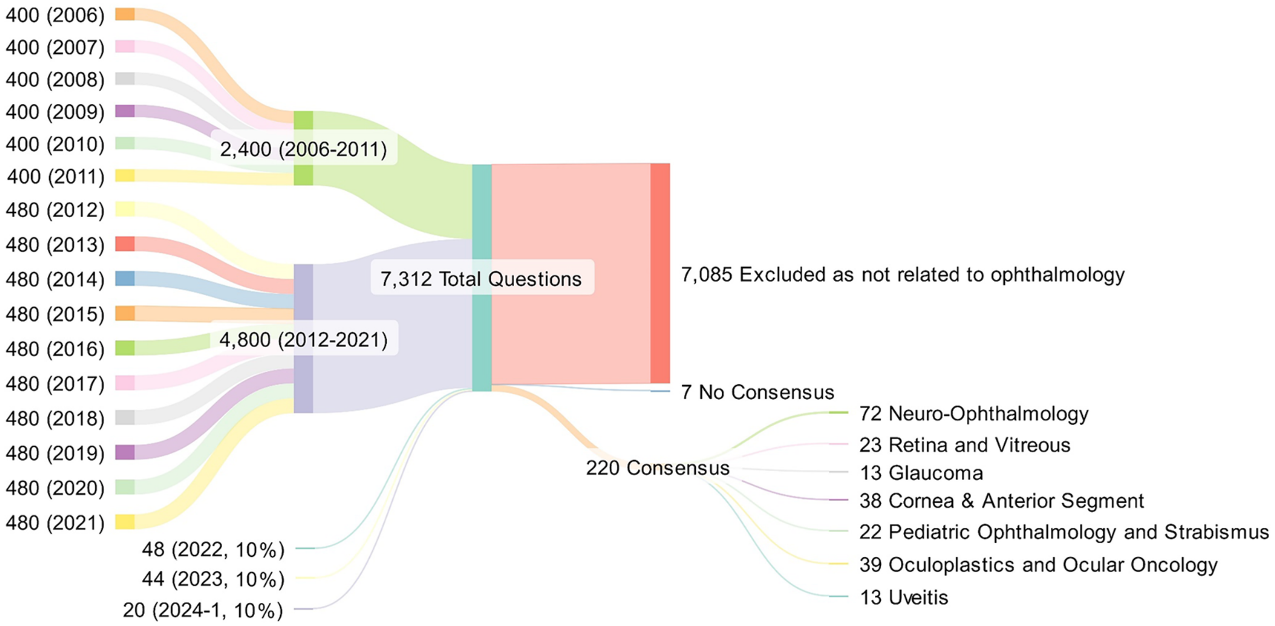


Figure 2. Sankey study diagram illustrating Medical Specialty Exam question selection and subspecialty distribution

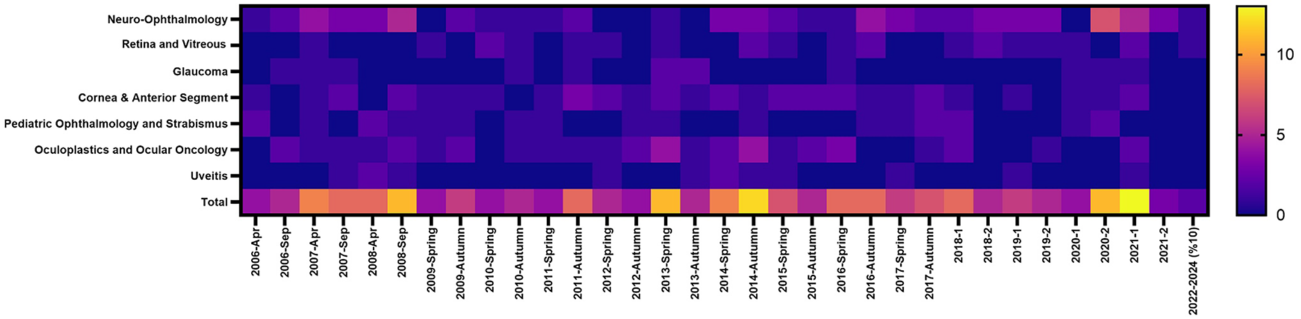


Figure 3. Subspecialty distribution heatmap chart of the Medical Specialty Exam questions according to year.

Detailed accuracy outcomes across all three attempts, stratified by language (Turkish and English) and by AI model, are presented in Table 1. In the final attempt, ChatGPT-4o demonstrated higher accuracy rates in both Turkish (97.3%) and English (97.7%) compared to Gemini 1.5 Pro (94.1% and 93.2%, respectively). This difference reached statistical significance in the English-language comparison (p=0.039), while it did not reach significance in Turkish (p=0.159). Although a progressive increase in accuracy was observed across successive attempts for both models, these changes were not statistically significant (p>0.05). In the overall analysis (n=220), ChatGPT-4o demonstrated superior performance across all attempts in terms of the number of correct responses. In Turkish, ChatGPT-4o achieved 209, 210, and 214 correct answers, while Gemini 1.5 Pro produced 202, 204, and 207, respectively (p>0.05 for all comparisons). In English, this difference reached statistical significance in the final attempt (215 vs. 205; p=0.039), although earlier attempts did not achieve significant differences (p=0.312).

When evaluated across individual ophthalmic subspecialties, no statistically significant differences were observed between the two AI platforms. Furthermore, no statistically significant interlingual variation was noted for either model when answering the same set of questions in Turkish versus English. The detailed distribution of performance according to exam years is provided in Table 2.

Among the evaluated items, only two MCQs included visual content. Notably, both chatbots correctly answered these questions in both languages, suggesting adequate visual interpretation capabilities under the tested conditions.

Content validity, assessed through average CVI values, demonstrated high agreement for both chatbots across all attempts and in both languages, as detailed in Table 3. Despite these high ratings, both models occasionally produced hallucinated or fabricated references. These instances—such as mismatched author names or journal titles—were systematically accounted for during I-CVI scoring.

In terms of explanation length, statistically significant differences were observed between Turkish and English responses for both models. Explanations generated in English were notably longer than their Turkish counterparts (ChatGPT-4o: median 178 vs. 88 words; Gemini 1.5 Pro: median 124 vs. 81.5 words; all comparisons, p<0.001). Furthermore, across both languages, ChatGPT-4o produced longer responses than Gemini 1.5 Pro (p<0.001 for both Turkish and English comparisons).

To assess response consistency across attempts, Cohen's kappa (κ) values were calculated for each AI model. In Turkish, κ values were 0.974 (95% CI, 0.967-0.980) for ChatGPT-4o and 0.967 (95% CI, 0.957-0.975) for Gemini 1.5 Pro. In English, both models achieved a κ value of 1.000, indicating perfect agreement. These results reflect near-perfect repeatability between the first and second attempts, during which no feedback was provided to the chatbots.

Discussion

The present study demonstrates that state-of-the-art LLM chatbots are capable of responding to ophthalmology-related MCQs in both Turkish and English with high levels of accuracy. Notably, ChatGPT-4o outperformed Gemini 1.5 Pro in the final evaluation attempt conducted in English, achieving statistically superior results. Despite this difference, both AI platforms exhibited robust performance across languages and attempts, supporting their potential as supplementary tools in ophthalmology education and assessment.

Due to the unique position and relative isolation of ophthalmology from other medical disciplines, ophthalmological questions can pose significant challenges for healthcare professionals. In parallel, the increasing reliance of healthcare professionals on online resources for up-to-date ophthalmological knowledge underscores the growing importance of AI-based LLMs in medical education. These models are rapidly gaining recognition as transformative tools in digital learning environments, capable of supplementing traditional instruction by providing immediate, structured, and reference-supported responses. Accordingly, AI-driven chatbots have emerged as accessible support mechanisms that can assist learners in interpreting complex MCQs across various languages and contexts.

Table 1. Interlingual and inter-model comparisons of chatbot performance on section questions across multiple attempts

	n	Correct answer count and accuracy (%)												p*			
		ChatGPT-4o						Gemini 1.5 Pro									
		TR			EN			TR			EN			Inter-AI		Interlingual	
		First	Second	Final	First	Second	Final	First	Second	Final	First	Second	Final	TR	EN	ChatGPT-4o	Gemini 1.5 Pro
Neuro-ophthalmology	**72**	67 (93.1)	68 (94.4)	70 (97.2)	66 (91.7)	66 (91.7)	70 (97.2)	65 (90.3)	65 (90.3)	67 (93.1)	64 (88.9)	64 (88.9)	64 (88.9)	0.763 0.745 0.441	0.779 0.779 0.097	>0.99 0.745 >0.99	>0.99 0.779 0.561
Retina and vitreous	**23**	22 (95.7)	22 (95.7)	23 (100)	23 (100)	23 (100)	23 (100)	22 (95.7)	22 (95.7)	22 (95.7)	22 (95.7)	22 (95.7)	22 (95.7)	>0.99 >0.99 >0.99	>0.99 >0.99 >0.99	>0.99 >0.99 >0.99	>0.99 >0.99 >0.99
Glaucoma	**13**	12 (92.3)	12 (92.3)	13 (100)	13 (100)	13 (100)	13 (100)	12 (92.3)	12 (92.3)	13 (100)	12 (92.3)	12 (92.3)	13 (100)	>0.99 >0.99 >0.99	>0.99 >0.99 >0.99	>0.99 >0.99 >0.99	>0.99 >0.99 >0.99
Cornea & anterior segment	**38**	38 (100)	38 (100)	38 (100)	38 (100)	38 (100)	38 (100)	36 (94.7)	36 (94.7)	36 (94.7)	38 (100)	38 (100)	38 (100)	0.493 0.493 0.493	>0.99 >0.99 >0.99	>0.99 >0.99 >0.99	0.493 0.493 0.493
Pediatric ophthalmology and strabismus	**22**	21 (95.5)	21 (95.5)	21 (95.5)	21 (95.5)	21 (95.5)	21 (95.5)	20 (90.9)	20 (90.9)	20 (90.9)	20 (90.9)	20 (90.9)	20 (90.9)	>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	**39**	37 (94.9)	37 (94.9)	37 (94.9)	38 (97.4)	38 (97.4)	39 (100)	36 (92.3)	36 (92.3)	37 (94.9)	37 (94.9)	37 (94.9)	37 (94.9)	>0.99 >0.99 >0.99	>0.99 >0.99 0.494	>0.99 >0.99 0.494	>0.99 >0.99 >0.99
Uveitis	**13**	12 (92.3)	12 (92.3)	12 (92.3)	11 (84.6)	11 (84.6)	11 (84.6)	11 (84.6)	12 (92.3)	12 (92.3)	11 (84.6)	11 (84.6)	11 (84.6)	>0.99 >0.99 >0.99	>0.99 >0.99 >0.99	>0.99 >0.99 >0.99	>0.99 >0.99 >0.99
Total	**220**	209 (95)	210 (95.5)	214 (97.3)	21 (95.5)	210 (95.5)	215 (97.7)	202 (91.8)	203 (92.3)	207 (94.1)	204 (92.7)	204 (92.7)	205 (93.2)	0.249 0.233 0.159	0.312 0.312 **0.039**	>0.99 >0.99 >0.99	0.858 >0.99 0.845

TR: Turkish, EN: English, AI: artificial intelligence, LLM: Large language models. The AI-based LLMs were listed in alphabetical order. The p values of the interlingual and inter-AI comparisons are listed in order for the first, second, and final attempts. *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. $p<0.05$ was considered statistically different in 95% confidence interval

Table 2. Evaluation of ChatGPT-4o and Gemini 1.5 Pro across examination years: language and model comparisons

Year	n	Correct answer count and accuracy (%)												p*			
		TR						EN									
		ChatGPT-4o			Gemini 1.5 Pro			ChatGPT-4o			Gemini 1.5 Pro			Interlingual		Inter-AI	
		First	Second	Final	First	Second	Final	First	Second	Final	First	Second	Final	ChatGPT-4o	Gemini 1.5 Pro	TR	EN
2006	**9**	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	>0.99 >0.99 >0.99	>0.99 >0.99 >0.99	>0.99 >0.99 >0.99	>0.99 >0.99 >0.99
2007	**17**	17 (100)	17 (100)	17 (100)	15 (88.2)	15 (88.2)	15 (88.2)	17 (100)	17 (100)	17 (100)	15 (88.2)	15 (88.2)	15 (88.2)				
2008	**19**	19 (100)	19 (100)	19 (100)	18 (94.7)	18 (94.7)	18 (94.7)	19 (100)	19 (100)	19 (100)	19 (100)	19 (100)	19 (100)				
2009	**10**	10 (100)	10 (100)	10 (100)	9 (90)	9 (90)	10 (100)	10 (100)	10 (100)	10 (100)	10 (100)	10 (100)	10 (100)				
2010	**9**	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)	9 (100)				
2011	**12**	12 (100)	12 (100)	12 (100)	12 (100)	12 (100)	12 (100)	12 (100)	12 (100)	12 (100)	12 (100)	12 (100)	12 (100)				
2012	**9**	7 (77.8)	7 (77.8)	7 (77.8)	8 (88.9)	8 (88.9)	8 (88.9)	8 (88.9)	8 (88.9)	8 (88.9)	8 (88.9)	8 (88.9)	8 (88.9)				
2013	**16**	16 (100)	16 (100)	16 (100)	15 (93.8)	15 (93.8)	15 (93.8)	16 (100)	16 (100)	16 (100)	15 (93.8)	15 (93.8)	15 (93.8)				
2014	**21**	19 (90.5)	19 (90.5)	21 (100)	21 (100)	21 (100)	21 (100)	20 (95.2)	20 (95.2)	21 (100)	19 (90.5)	19 (90.5)	19 (90.5)				
2015	**12**	11 (91.7)	11 (91.7)	12 (100)	12 (100)	12 (100)	12 (100)	10 (83.3)	10 (83.3)	12 (100)	10 (83.3)	10 (83.3)	10 (83.3)				
2016	**16**	15 (93.8)	15 (93.8)	15 (93.8)	13 (81.3)	13 (81.3)	13 (81.3)	15 (93.8)	15 (93.8)	15 (93.8)	14 (87.5)	14 (87.5)	14 (87.5)				
2017	**13**	13 (100)	13 (100)	13 (100)	12 (92.3)	12 (92.3)	13 (100)	12 (92.3)	12 (92.3)	13 (100)	11 (84.6)	11 (84.6)	11 (84.6)				
2018	**13**	13 (100)	13 (100)	13 (100)	11 (84.6)	11 (84.6)	11 (84.6)	13 (100)	13 (100)	13 (100)	12 (92.3)	12 (92.3)	12 (92.3)				
2019	**11**	10 (90.9)	11 (100)	11 (100)	8 (72.7)	9 (81.8)	9 (81.8)	10 (90.9)	10 (90.9)	10 (90.9)	10 (90.9)	10 (90.9)	10 (90.9)				
2020	**15**	13 (86.7)	13 (86.7)	13 (86.7)	14 (93.3)	15 (100)	15 (100)	13 (86.7)	13 (86.7)	13 (86.7)	15 (100)	15 (100)	15 (100)				
2021	**16**	14 (87.5)	14 (87.5)	15 (93.8)	14 (87.5)	14 (87.5)	15 (93.8)	15 (93.8)	15 (93.8)	16 (100)	14 (87.5)	14 (87.5)	15 (93.8)				
2022-2024 (10%)	**2**	2 (100)	2 (100)	2 (100)	2 (100)	2 (100)	2 (100)	2 (100)	2 (100)	2 (100)	2 (100)	2 (100)	2 (100)				

TR: Turkish, EN: English, AI: Artificial intelligence, LLM: Large language models. *Comparison between the years 2016-2024 via chi-square test. The AI-based LLMs were listed in alphabetical order. The p values of the interlingual and inter-AI comparisons are listed in order for the first, second, and final attempts. p<0.05 was considered statistically different in 95% confidence interval

Over the past few years, LLM chatbots like ChatGPT and Gemini have attracted growing interest as educational tools in medicine.[5,6,7,8,9] Early versions, such as ChatGPT-3.5 and Google Bard, delivered only moderate performance, with reported accuracies between 50% and 70% across different exam settings including the Turkish MSE, United States Medical Licensing Examination, and dedicated ophthalmology question banks.[15,23,24,25,26] These results, while promising, highlighted clear limitations in reasoning depth, domain-specific precision, and multilingual reliability.

As newer models like ChatGPT-4o and Gemini 1.5 Pro emerged, a marked improvement became evident. Several studies reported significantly higher success rates—often exceeding 70%, and in some cases over 90%—particularly in structured, multiple-choice exam formats and language-specific settings such as the medical proficiency tests for medicine or ophthalmological board assessments.[13,14,16,17,27]

Still, much of the available research has focused on open-ended questions or general medical content. Few have looked closely at ophthalmology—a highly specialized and visually driven field—and even fewer have explored how these models perform across different languages. This study was designed to address that gap by directly comparing ChatGPT-4o and Gemini 1.5 Pro on a bilingual (Turkish and English) set of ophthalmology-related MCQs, using standardized prompts that required scientific justification and citation. In doing so, our aim was not only to assess model accuracy, but also to explore the pedagogical and linguistic dimensions of AI-assisted learning in a focused clinical field. Interestingly, the comparatively high accuracy rates observed in our findings may

Table 3. Validity assessment of explanations generated by chatbots

			Average CVI	ICC (95% CI)
ChatGPT-4o	**TR**	First	0.95	0.849 (0.756-0.901)
		Second	0.96	0.850 (0.774-0.897)
		Final	0.97	0.834 (0.753-0.885)
	EN	First	0.96	0.951 (0.936-0.963)
		Second	0.96	0.942 (0.924-0.956)
		Final	0.98	0.885 (0.850-0.912)
Gemini 1.5 Pro	**TR**	First	0.93	0.862 (0.771-0.911)
		Second	0.94	0.878 (0.820-0.915)
		Final	0.95	0.850 (0.757-0.902)
	EN	First	0.92	0.927 (0.893-0.949)
		Second	0.93	0.925 (0.890-0.947)
		Final	0.93	0.918 (0.877-0.944)

TR: Turkish, EN: English, CVI: Content validity index, ICC: Intraclass correlation coefficient, CI: Confidence interval

be attributed to several methodological strengths. First, we employed the most recent versions of both AI platforms, each incorporating substantial architectural improvements over earlier iterations such as ChatGPT-3.5 or Google Bard. Second, the use of structured prompts that demanded not just correct answers, but also evidence-based reasoning, likely enhanced the quality of model outputs. Third, the bilingual design enabled controlled cross-linguistic comparison, offering valuable insight into model behavior in languages underrepresented during training. This combination of technological currency, prompt rigor, and linguistic breadth distinguishes the present study from prior work and reinforces the relevance of LLMs as adaptable tools in medical education.

Upon evaluating the exams over the years, we noted a lack of inter-AI and interlingual differences, but there was a significant difference in the inter-AI comparison for the total English MCQs in the final attempt. These results should not be viewed as contradictory, as it was likely influenced by the heterogeneous distribution of question types and difficulty levels across examination years.

One of the more intriguing findings in this study was the effect of user feedback on chatbot performance. While neither model truly “learns” in the traditional human sense during testing, both ChatGPT-4o and Gemini 1.5 Pro showed modest improvements in their final attempt after receiving a standardized negative feedback signal for incorrect answers. This raises an important question: to what extent do LLMs adapt their outputs in response to structured cues, even without persistent memory? These observations may reflect the underlying influence of reinforcement learning from human feedback, a core training mechanism that guides how these models prioritize factual consistency and contextual reasoning.[28,29] Although no real-time learning occurs during user interaction, feedback signals—such as rating a response as “factually incorrect”—can temporarily shift the model’s focus toward more cautious, evidence-based reasoning patterns.[28,29,30] In practical terms, this suggests that even a simple, well-designed correction can nudge a chatbot toward a more accurate and academically grounded answer, particularly in high-stakes domains like medicine. As previously emphasized by Antaki et al.[15], the educational value of LLMs lies not only in their ability to produce correct answers but also in their potential to facilitate reasoning and reflection. For medical educators and exam designers, this opens up new possibilities. If thoughtfully implemented, controlled feedback mechanisms could enhance the pedagogical role of chatbots—not just as static responders, but as adaptive tools that promote critical thinking and iterative learning.

The validity analysis indicated that both chatbots achieved satisfactory content validity across Turkish and English, as reflected by consistently high expert ratings. Notably, the explanations generated in English were more detailed than those in Turkish for both models, suggesting that users may access richer content when interacting in English. ChatGPT-4o, in particular, provided longer and more comprehensive responses in both languages, making it a potentially preferable tool for learners seeking in-depth justifications. Furthermore, both models frequently included brief comments on why alternative options were incorrect. This practice of addressing distractors may enhance the educational value of chatbot interactions by promoting a deeper understanding of the reasoning process underlying multiple-choice assessments.

AI-based LLM chatbot technology, readily accessible at people’s fingertips, continues to evolve rapidly, including in ophthalmology.[31,32,33] For instance, earlier versions of ChatGPT were limited by a knowledge cut-off (September 2021).[34,35,36] However, with the latest updates, ChatGPT has gained the ability to browse the internet and provide up-to-date content, demonstrating the potential for progressively improving accuracy rates. While this advancement is promising for research purposes, it also introduces the disadvantage of rapid publication obsolescence.[35] Additionally, it may lead to accuracy discrepancies between different versions of the same chatbot, posing challenges for consistent and reliable use in academic and professional settings. Also, despite these significant advancements, chatbots remain prone to generating hallucinations and fabricating references.[35] Therefore, maintaining a supervisory role while utilizing such tools is essential to ensure reliability.

Study Limitations

While this study elucidates critical aspects regarding the benefits of ChatGPT-4o over Gemini 1.5 Pro in addressing ophthalmology-related MCQs in Turkish MSEs, it is not devoid of limitations, including 1) evaluating performance only in Turkish and English languages, 2) the lack of assessment for open-ended question performance, and 3) the use of only two AI-based LLMs, despite the availability of many other models. Another significant limitation of this study is the focus on evaluating the effectiveness of LLMs using MSE questions specifically designed to assess the fundamental ophthalmology knowledge of general practitioners. Consequently, the findings presented here are not comprehensive enough to fully elucidate the potential role of LLMs in ophthalmology education. Further detailed studies focusing on various aspects of ophthalmology are required to better understand and define the utility of LLMs in this field. Also, in our study, only the officially published answer keys and question cancellations were taken into consideration. While rare, there have been instances in such examinations where questions were later contested, with appeals or legal proceedings initiated for their cancellation. However, these questions are typically not reflected in the officially released answer keys and therefore could not be accounted for in analysis. This represents a limitation of the study, as the inclusion of such contested questions might have provided a more comprehensive assessment of the data.

Although extremely rare, it was observed in both AI-based models that the logical explanation was provided but the wrong choice was chosen, or the wrong explanation was given but the correct choice was selected. One should always remember that everyone, including AI, can make errors, so it is always wise to check the results. Furthermore, as chatbots are prone to generating fabricated references and hallucinations, the lack of a dedicated validity analysis specifically aimed at assessing reference accuracy may be regarded as a limitation. Lastly, since accuracy rates of participants for these exams were not known and not publicly available, a comparison between the human accuracy rates and those of the AIs could not be performed.

Even with these flaws, to the best of our knowledge, this is the first AI comparative study to reveal that ChatGPT-4o exhibits a modest performance advantage over Gemini 1.5 Pro in addressing ophthalmology-related MCQs in Turkish MSEs. Additionally, the evaluation of a substantial number of MCQs (n=220) and the inclusion of three consecutive attempts with and without feedback enhance this work. Furthermore, the requirement for scientific explanations from PubMed and the WoS Citation Index may have influenced these results. The use of the most up-to-date AI versions also strengthens the study. Finally, unlike most other studies, questions containing figures were evaluated in this study.

Conclusion

Both AI-based LLMs demonstrated robust performance in answering ophthalmology-related MCQs. They hold promise for improving ophthalmology education by not only accurately identifying the correct answers to ophthalmology-related MCQs but also offering explanations. While both AI platforms prove to be useful, ChatGPT-4o is significantly ahead. Further research on the contributions of AI-driven e-learning, particularly for med students and ophthalmology residents, is essential in this relatively nascent technological field.

Ethics

Ethics Committee Approval: Since this study did not involve human participants or animal experiments, ethics approval and informed consent were not required.

Informed Consent: Patient consent is not required for this article.

Acknowledgment

We would like to thank Hasan Nadir Rana, MD, native English speaker and UK resident, for linguistic review of this manuscript.

Declarations

Authorship Contributions

Surgical and Medical Practices: M.C.S., Z.Y., Concept: M.C.S., Z.Y., Design: M.C.S., Z.Y., Data Collection or Processing: M.C.S., Z.Y., Analysis or Interpretation: M.C.S., Z.Y., Literature Search: M.C.S., Z.Y., Writing: M.C.S., Z.Y.

Conflict of Interest: All authors certify that they have no affiliations with or involvement in any organization with any financial or non-financial interest in the subject matter or materials discussed in this article. The use of multiple-choice questions was intended for scientific purposes only.

Financial Disclosure: All expenses related to premium subscriptions of the chatbot platforms were personally borne by the authors. No external funding or financial support was obtained for the conduct of this study.

References

1. Cahit A. Can machines think and how can they think? Atatürk Üniversitesi 1958-1959 Öğretim Yılı Halk Konf. 1959:91-103.
2. Keskinbora K, Güven F. Artificial intelligence and ophthalmology. Turk J Ophthalmol. 2020;50:37-43.
3. Shemer A, Cohen M, Altarescu A, Atar-Vardi M, Hecht I, Dubinsky-Pertzov B, Shoshany N, Zmujack S, Or L, Einan-Lifshitz A, Pras E. Diagnostic capabilities of ChatGPT in ophthalmology. Graefes Arch Clin Exp Ophthalmol. 2024;262:2345-2352.
4. Tsui JC, Wong MB, Kim BJ, Maguire AM, Scoles D, VanderBeek BL, Brucker AJ. Appropriateness of ophthalmic symptoms triage by a popular online artificial intelligence chatbot. Eye (Lond). 2023;37:3692-3693.
5. Güler MS, Baydemir EE. Evaluation of ChatGPT-4 responses to glaucoma patients' questions: can artificial intelligence become a trusted advisor between doctor and patient? Clin Exp Ophthalmol. 2024;52:1016-1019.

6. Chen JS, Reddy AJ, Al-Sharif E, Shoji MK, Kalaw FGP, Eslani M, Lang PZ, Arya M, Koretz ZA, Bolo KA, Arnett JJ, Roginiel AC, Do JL, Robbins SL, Camp AS, Scott NL, Rudell JC, Weinreb RN, Baxter SL, Granet DB. Analysis of ChatGPT responses to ophthalmic cases: can ChatGPT think like an ophthalmologist? Ophthalmol Sci. 2024;5:100600.
7. Carlà MM, Gambini G, Baldascino A, Giannuzzi F, Boselli F, Crincoli E, D'Onofrio NC, Rizzo S. Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. Br J Ophthalmol. 2024;108:1457.
8. Ming S, Yao X, Guo X, Guo Q, Xie K, Chen D, Lei B. Performance of ChatGPT in ophthalmic registration and clinical diagnosis: cross-sectional study. J Med Internet Res. 2024;26:e60226.
9. David D, Zloto O, Katz G, Huna-Baron R, Vishnevskia-Dai V, Armarnik S, Zauberman NA, Barnir EM, Singer R, Hostovsky A, Klang E. The use of artificial intelligence based chat bots in ophthalmology triage. Eye (Lond). 2025;39:785-789.
10. Halaweh M. ChatGPT in education: strategies for responsible implementation. Contemp Educ Technol. 2023;15:e421.
11. Tlili A, Shehata B, Adarkwah MA, Bozkurt A, Hickey DT, Huang R, Agyemang B. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learn Environ. 2023;10:1-24.
12. 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:e57348.
13. 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;46:706-711.
14. Wu JH, Nishida T, Liu TYA. Accuracy of large language models in answering ophthalmology board-style questions: a meta-analysis. Asia Pac J Ophthalmol (Phila). 2024;13:100106.
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.
16. 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:e49903.
17. 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:e40822.
18. Previous MSEs in 2006-2021. https://www.osym.gov.tr/TR,15072/tus-cikmis-sorular.html.
19. 10% of MSE questions in 2022. https://www.osym.gov.tr/TR,22532/2022.html.
20. 10% of MSE questions in 2023. https://www.osym.gov.tr/TR,25279/2023.html.
21. 10% of MSE questions in 2024. https://www.osym.gov.tr/TR,29136/2024.html.
22. Polit DF, Beck CT. The content validity index: are you sure you know what's being reported? Critique and recommendations. Res Nurs Health. 2006;29:489-497.
23. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. Medicine (Baltimore). 2023;102:e34673.
24. Ilgaz HB, Çelik Z. The significance of artificial intelligence platforms in anatomy education: an experience with ChatGPT and Google Bard. Cureus. 2023;15:e45301.
25. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198.
26. Mihalache A, Huang RS, Popovic MM, Muni RH. Performance of an upgraded artificial intelligence chatbot for ophthalmic knowledge assessment. JAMA Ophthalmol. 2023;141:798-800.
27. Sabaner MC, Hashas ASK, Mutibayraktaroglu KM, Yozgat Z, Klefter ON, Subhi Y. 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. AJO Int. 2024:100070.
28. Yang Z, Wang D, Zhou F, Song D, Zhang Y, Jiang J, Kong K, Liu X, Qiao Y, Chang RT, Han Y, Li F, Tham CC, Zhang X. Understanding natural language: potential application of large language models to ophthalmology. Asia Pac J Ophthalmol (Phila). 2024;13:100085.
29. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29:1930-1940.
30. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 2022;35:27730-27744.
31. 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.
32. Sabaner MC, Anguita R, Antaki F, Balas M, Boberg-Ans LC, Ferro Desideri L, Grauslund J, Hansen MS, Klefter ON, Potapenko I, Rasmussen MLR, Subhi Y. Opportunities and challenges of chatbots in ophthalmology: a narrative review. J Pers Med. 2024;14:1165.
33. Postacı SA, Dal A. The ability of large language models to generate patient information materials for retinopathy of prematurity: evaluation of readability, accuracy, and comprehensiveness. Turk J Ophthalmol. 2024;54:330-336.
34. Vaishya R, Misra A, Vaish A. ChatGPT: is this version good for healthcare and research? Diabetes Metab Syndr. 2023;17:102744.
35. Gurnani B, Kaur K. Leveraging ChatGPT for ophthalmic education: a critical appraisal. Eur J Ophthalmol. 2024;34:323-327.
36. Pushpanathan K, Zou M, Srinivasan S, Wong WM, Mangunkusumo EA, Thomas GN, Lai Y, Sun CH, Lam JSH, Tan MCJ, Lin HAH, Ma W, Koh VTC, Chen DZ, Tham YC. Can OpenAI's new o1 model outperform its predecessors in common eye care queries? Ophthalmol Sci. 2025;5:100745.