Yalova Tıp Dergisi 1(1):8-12, 2026 | Yalova Medical Journal 1(1):8-12, 2026 https://yamj.yalova.edu.tr

YALOVA TIP DERGİSİ

YALOVA MEDICAL JOURNAL

ÖZGÜN MAKALE | ORIGINAL ARTICLE

Performance of Large Language Models in Ophthalmology Questions: Do Language Differences Matter?

Elvin Halili Celenk[1], Nursel Melda Yenerel[2], Peykan Turkcuoglu[3]

[1] Yalova University, Faculty of Medicine, Department of Ophthalmology, Yalova, Turkey.
[2] Medipol Acibadem Regional Hospital, Istanbul, Turkey.
[3] Veni Vidi Eye Hospital Atasehir, Istanbul, Turkey.

Abstract

Aim: To evaluate the effectiveness of ChatGPT-4.0, Copilot AI, Gemini AI, and Claude AI chatbots in answering the fundamental principles of ophthalmology.

Materials and Methods: All forty questions in the study questions section of the American Academy of Ophthalmology (AAO) 2024-2025 Basic and Clinical Science Course (BCSC) Fundamentals and Principles of Ophthalmology book were asked to ChatGPT-4.0, Copilot AI, Gemini AI, and Claude AI in both English and Turkish. The questions were asked only once and at different times of the day for each language. The responses provided by the chatbots were compared with the official answer key provided at the end of the book and classified as correct or incorrect.

Results: For English questions, Copilot AI (95%) and Claude AI (92.5%) showed higher accuracy than ChatGPT-4.0 (75%) and Gemini AI (87.5%). In Turkish questions, the accuracy rates of all models were found to be close to each other (85–92.5%), and no significant superiority was detected. Across languages, all AI applications showed medium-high consistency. The highest level of consistency was detected in Copilot AI and Claude AI.

Conclusion: Although Copilot AI and Claude AI outperformed the other two bots in English questions, they did not demonstrate a meaningful superiority over each other in Turkish questions.

Keywords: ChatGPT- 4.0, Copilot, Gemini, Claude, Ophthalmology Education

Corresponding Author: Assistant Professor, Elvin Halili Celenk

Address: *Yalova University, Faculty of Medicine, Department of Ophthalmology, Yalova, Turkey.*
E-mail: *elvin.celenk@yalova.edu.tr*

Introduction

Today, AI is used in many fields, and significant developments have been made in the field of medicine. Deep learning and broad-based language models can perform human-like language-based tasks such as text comprehension, question answering, and text generation [1,2]. These technologies also show high success in image analysis and clinical decision support processes. Ophthalmology is one of the disciplines that benefits most from AI applications, as it is a field that makes intensive use of visual data. Algorithms developed for retinal image analysis, early diagnosis of diseases, risk classification, and patient follow-up have begun to transform clinical applications.

The fundamental principles of ophthalmology are among the most important topics that must be learned during ophthalmology training. It is important for the quality of education that tools are available to help find answers to the questions we use during ophthalmology training and while preparing for exams, and that these tools ensure standardization in the answers.

The reliability of artificial intelligence chatbots in answering questions on different topics in ophthalmology is also a subject of research. Research has revealed both the benefits and shortcomings of artificial intelligence chatbots [3-7]. A study conducted to prepare for the "Fellowship of the European Board of Ophthalmology" and "Royal College of Ophthalmologists Fellowship" exams reported that chatbots were effective [3-7].

In November 2023, Microsoft renamed the new Bing to Copilot. Google renamed it Bard Gemini in 2024.
The aim of our study is to evaluate the effectiveness of ChatGPT-4.0 (OpenAI; San Francisco, USA), Copilot AI (Microsoft; Redmond, USA), Gemini AI (Google; Mountain View, USA), and Claude AI (Anthropic, USA), recently referred to as "ethical artificial intelligence," in terms of the basic principles of ophthalmology. To the best of our knowledge, there is no publication in the literature comparing the performance of these chatbots in answering basic ophthalmology questions in different languages.

Materials and Methods

All forty questions from the study questions section of the American Academy of Ophthalmology (AAO) 2024-2025 Basic and Clinical Science Course (BCSC) Fundamentals and Principles of Ophthalmology book were included in the study, and their Turkish translations were performed by a certified translator [8]. This book contains approximately the same number of questions as the other volumes in the AAO BCSC series. The accuracy and appropriateness of the translations were evaluated by an expert ophthalmologist. We did not deem it appropriate to add to these questions, adhering to the AAO basic and clinical sciences book.

Questions were asked in both English and Turkish to ChatGPT-4.0, Copilot AI, Gemini AI, and Claude AI on January 30, 2025.All artificial intelligence models included in this study were accessed via their official web-based interfaces using standard personal using accounts. Mobile applications were not utilized at any stage of the study. Regarding model settings, all AI platforms were used with their default configurations Before each question, the command "I will ask you a multiple-choice question; please provide the answer as an option" was given. After the bot responded, the session was closed, and the same command was reused for each question. Questions were asked only once and at different times of the day for each language. The responses provided by the chatbots were compared with the official answer key provided at the end of the book and classified as correct or incorrect. The evaluatıon of chatbot responces was conducted by a single experienced ophthalmologist. Since our study did not involve human or animal data, no ethics committee approval was obtained.

ChatGPT 4.0 is the latest version of GPT, released in March 2023. It is a Large Language Model (LLM)-based artificial intelligence chatbot that has access to a vast data network, is trainable, and can generate responses similar to human intelligence [8]. Copilot AI is an LMM-based AI bot used in many fields thanks to its natural language processing capability integrated with the GPT-4 AI system [9]. Gemini AI, on the other hand, uses the LaMDA language family to produce realistic language in natural language processing and strives to provide realistic responses. The 1.5 pro version was used in our study [10]. Claude AI is a generative artificial intelligence (AI) chatbot and large language model (LLM) family developed by the research company Anthropic. The Claude 3.5 sonnet version was used in our study.

Results

All 40 questions in the basic information and principles section of the ophthalmology textbook were asked in English to artificial intelligence chatbots. ChatGPT 4.0 answered 30 of these questions (75%) correctly and 10 (25%) incorrectly. Copilot AI answered 38 (95%) questions correctly and 2 (5%) incorrectly. Gemini AI answered 35 (87.5%) questions correctly and 5 (12.5%) questions incorrectly. Claude AI answered 37 (92.5%) questions correctly and 3 (7.5%) questions incorrectly (Figure 1). Copilot AI and Claude AI outperformed the other two chatbots (Chat GPT 4.0, Gemini AI) in correctly answering questions in English. A moderate level of agreement was observed between the responses of the two AI applications (κ=0.362). This agreement was found to be statistically significant (p=0.019) (Table 1).

The Turkish versions of the same questions were applied to the AI chatbots. Chat GPT 4.0 answered 35 of these questions correctly (87.5%) and 5 incorrectly (12.5%). Copilot AI answered 37 questions correctly (92.5%) and 3 incorrectly (7.5%). Gemini AI answered 34 questions correctly (85%) and 6 questions incorrectly (15%). Claude AI answered 34 questions correctly (85%) and 6 questions incorrectly (15%) (Figure 1), (Table 1). No superiority was detected among the four artificial intelligences in answering Turkish questions.

ChatGPT-4.0 provided identical answers to 33 questions of English and Turkish questions (82.5%) and different answers to 7 questions(17.5%). Of the questions with different answers, 6 were answered correctly (15%) when asked in Turkish, while 1 was answered incorrectly (2.5%) when asked in Turkish. There was no statistically significant difference between ChatGPT 4.0's success rates in answering English and Turkish questions (p=0.125). The Cohen's Kappa coefficient between ChatGPT- 4.0 (English) and ChatGPT-4.0 (Turkish) was calculated as κ=0.440, indicating that the artificial intelligence application provided consistent answers in different languages (Table 1).

Copilot AI answered 39 questions of the English and Turkish questions (97.5%) identically and 1 question(2.5%) differently. The differently answered question was incorrect when asked in Turkish. There was no statistically significant difference between Copilot AI's success rates in answering English and Turkish questions (p=0.997). The Cohen's Kappa coefficient between Copilot AI-English and Copilot AI-Turkish was calculated as κ=0.497, indicating that the AI application provided consistent answers in different languages (Table 1).

Gemini AI provided identical answers to 37 questions of the English and Turkish questions (92.5%) and different answers to 3 questions(7.5%). Of the questions with different answers, 2 were answered incorrectly (5%) when asked in Turkish, and 1 was answered correctly (2.5%) when asked in English.

Table 1. Artificial intelligence chatbots' responses to questions related to the American Academy of Ophthalmology and Ophthalmology Fundamentals and Principles of Ophthalmology and their changes.

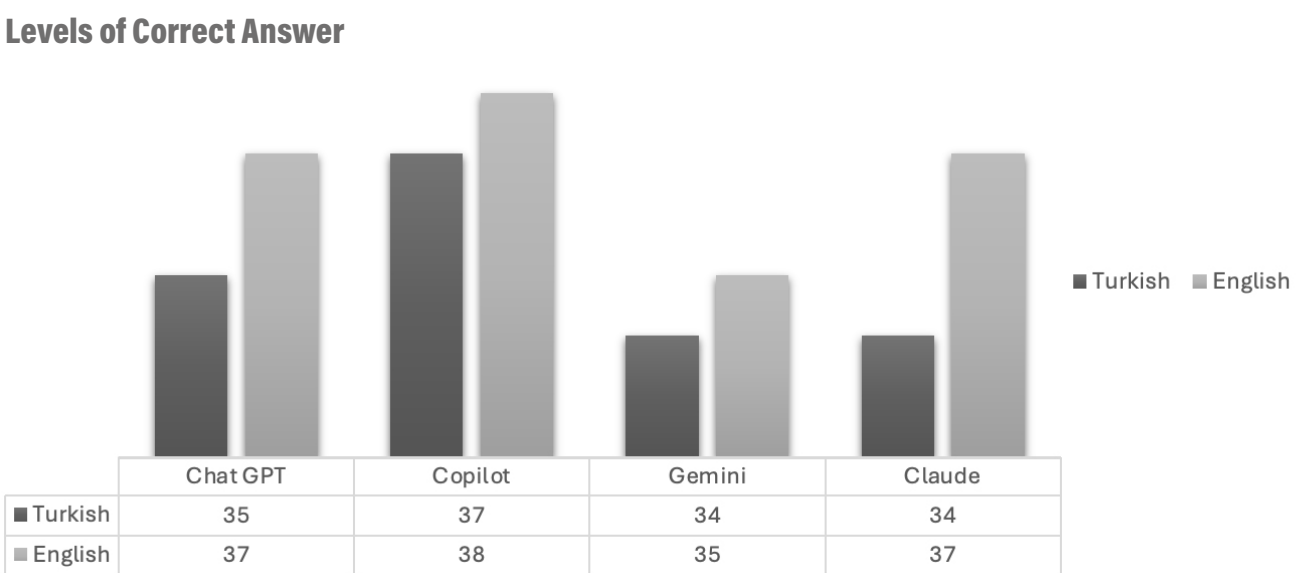
Answers	Chat GPT (English)	Chat GPT (Turkish)	Copilot (English)	Copilot (Turkish)	Gemini (English)	Gemini (Turkish)	Claude (English)	Claude (Turkish)
True	30	35	38	37	35	34	34	37
False	10	5	2	3	5	6	6	3
Giving the same answer	33 (82.5%)		39 (97.5%)		37 (92.5%)		37(92.5%)	
Giving a different answer	7 (17.5%)		1 (2.5%)		3 (7.5%)		3 (7.5%)	
True-false change	1 (2.5%)		1 (2.5%)		2 (5%)		3 (7.5%)	
False-true change	6 (15%)		0 (0%)		1 (2%)		0 (0%)	
[a]p	0.125		0.997		0.995		0.250	
κ	0.440 p=0.002**		0.497 p=0.001**		0.684 p=0.001**		0.630 p=0.001**	

*[a]Mc Nemar Testi, **p< 0,01. *p< 0.05, Cohen's Kappa katsayısı ; κ*

There was no statistically significant difference between the success rates for English and Turkish (p=0.995). The Cohen's Kappa coefficient between Gemini AI-English and Gemini AI-Turkish was calculated as κ=0.684, indicating that the AI application showed a high level of consistency in both languages (Table 1).
Claude AI answered 37 questions of the same questions in English and Turkish correctly(92.5%), and 3 questions incorrectly(7.5%). Of the questions answered differently, 3 (7.5%) were answered incorrectly when asked in Turkish.

There was no statistically significant difference between Claude AI's success rates in answering English and Turkish questions (p=0.250). The Cohen's Kappa coefficient between Claude AI-English and Claude AI-Turkish was calculated as κ=0.630, indicating that the AI application provided consistent answers in different languages (Table 1).

Figure 1. Levels of Correct Answer for Turkish and English



Discussion

In the field of ophthalmology, AI- based systems are increasingly being used to assess students' knowledge levels and improve educational processes [6,7,11,12]. Various studies in the literature show that these systems perform at different levels in exams and knowledge assessment processes. Raimondi et al. achieved the highest accuracy rate of 82.9% with Bing Chat in chatbots used in the Royal College of Ophthalmologists' specialty exams in the UK, without any guidance or instruction adjustments [7]. Kung H. T. et al., in their study evaluating ChatGPT's performance on the United States Medical Licensing Examination (USMLE), showed that it achieved success close to or at the 60% accuracy threshold and could therefore assist human learners as a preliminary step toward future integration into clinical decision-making processes in medical education settings [12]. Antaki F. et al. reported that ChatGPT demonstrated promising performance in a simulated ophthalmology knowledge assessment exam and that domain-specific pre-training may be necessary to enhance performance with customized large language models (LLMs) [13]. Panthier et al., in their study examining ChatGPT's successful completion of the French version of the European Board of Ophthalmology (EBO) exam, reported a 91% success rate, indicating a high level of proficiency in ophthalmology knowledge and practice [6]. Şensoy E. et al., in their study where questions on Ophthalmic Pathology and Intraocular Tumors were asked to AI chatbots, reported that ChatGPT, Bing, and Bard answered correctly at rates of 58.6%, 63.9%, and 69.4%, respectively; However, they reported that the differences in accuracy rates between the three programs were not statistically significant (p>0.05) [14].

In our study, the accuracy rates of correct answers to English questions were similar across the four chatbots, with Copilot AI providing more correct answers than the others. It was determined that there was a good level of agreement between the responses given by Claude AI and Gemini to English questions (Cohen's Kappa: 0.448) and that this was statistically significant (p=0.003). When comparing Claude AI and Copilot, a moderate level of agreement was found (Cohen's Kappa: 0.362), which was also statistically significant (p=0.019). When the four chatbots were evaluated together, it was observed that Gemini AI, Copilot AI, and Claude AI, in particular, provided similar responses to questions regarding the fundamentals and principles of eye diseases.

Mihalache A. et al. evaluated the performance of Gemini and Bard in different countries via the "EyeQuiz" platform and reported that these chatbots performed acceptably in answering exam questions, with no statistically significant differences between country versions [3]. In our study, the accuracy rates of the four chatbots were similar for Turkish questions, with Copilot providing more correct answers than the others. Claude AI and ChatGPT showed a high level of agreement for the Turkish questions (Cohen's $\kappa = 0.684$, $p = 0.001$). In addition, a good level of agreement was observed between Claude AI–Turkish and Gemini AI–Turkish (Cohen's $\kappa = 0.608$, $p = 0.001$), as well as between ChatGPT–Turkish and Copilot AI–Turkish (Cohen's $\kappa = 0.448$, $p = 0.003$).

Şensoy E. et al., in their study evaluating the effects of language differences on multiple-choice questions about eye inflammation and uveitis using ChatGPT-3.5, Copilot AI, and Gemini AI, reported accuracy rates of 63.9%, 63.9%, and 50% for English questions, respectively; and 52.8%, 52.8%, and 66.7% accuracy rates for Turkish questions, respectively. The researchers noted that the AI programs had different accuracy rates when answering English and Turkish questions, but there was no statistically significant difference between their performance (p>0.05) [15]. In our study, four chatbots also provided correct answers at different rates when answering English and Turkish questions, but no statistically significant difference was found since the p-value was equal to the traditional threshold of 0.05.

The limitations of this study include the use of a limited number of questions and the lack of comparative analysis on specific subtopics within the field. It would be beneficial to conduct assessments in different languages using a larger number of questions.

In conclusion, this study evaluated not only the performance of AI chatbots but also the existence of performance differences between the English and Turkish versions of the same questions. While Copilot AI and Claude AI outperformed the other two bots on English questions, they did not demonstrate a significant advantage over each other on Turkish questions. We believe that the small number of questions may have affected the statistical significance of the results. Our findings show that AI-based chatbots can be used as auxiliary tools in ophthalmology education, particularly facilitating students' quick access to information during the learning of basic principles and exam preparation processes. Clinically, it can be said that these systems are not yet mature enough to be used as decision support mechanisms, but they have the potential to serve as additional tools that will reduce the burden on clinicians in the future. Therefore, both the development of customized models that will provide greater accuracy and a critical approach by users are of great importance for safe integration in education and clinical practice. We believe that this topic requires further evaluation through comprehensive studies that address additional questions.

Acknowledgments and Funding Statement: Not applicable

Conflict of Interest: The authors have no conflict of interest to declare.

Author Contributions: Concept: E. H. C.; Design: E. H. C.; Supervision: N. M. Y., P. T.; Materials: E. H. C.; Data Collection and/or Processing: E. H. C.; Analysis and/or Interpretation: E. H. C., N. M. Y.; Literature Review: E. H. C., P. T; Writing: E. H. C.; Critical Review: E. H. C.

References

1. Flanagin A, Bibbins-Domingo K, Berkwits M, Christiansen SL. Nonhuman "authors" and implications for the integrity of scientific publication and medical knowledge. JAMA. 2023;329(8):637–9. doi:10.1001/jama.2023.1344.

2. Rahimy E. Deep learning applications in ophthalmology. Curr Opin Ophthalmol. 2018;29(3):254–60. doi:10.1097/ICU.0000000000000470.

3. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol. 2023;141(6):589–97. doi:10.1001/jamaophthalmol.2023.1144.

4. Canleblebici M, Dal A, Erdağ M. Evaluation of the performance of large language models (ChatGPT-3.5, ChatGPT-4, Bing and Bard) in Turkish ophthalmology chief-assistant exams: a comparative study. Turk Klin J Ophthalmol. 2024;33(3):163–70. doi:10.5336/ophthal.2024-102632.

5. Tan TF, Thirunavukarasu AJ, Campbell JP, Keane PA, Pasquale LR, Abramoff MD, et al. Generative artificial intelligence through ChatGPT and other large language models in ophthalmology: clinical applications and challenges. Ophthalmol Sci. 2023;3(4):100394. doi:10.1016/j.xops.2023.100394.

6. Panthier C, Gatinel D. Success of ChatGPT, an artificial intelligence language model, on the French version of the European Board of Ophthalmology examination: a new approach to medical knowledge assessment. J Fr Ophtalmol. 2023;46(7):706–11. doi:10.1016/j.jfo.2023.05.006.

7. Raimondi R, Tzoumas N, Salisbury T, Di Simplicio S, Romano MR; North East Trainee Research in Ophthalmology Network (NETRiON). Comparative analysis of major language models in Royal College of Ophthalmologists fellowship examinations. Eye (Lond). 2023;37(17):3530–3. doi:10.1038/s41433-023-02563-3.

8. Waisberg E, Ong J, Masalkhi M, Zaman N, Sarker P, Lee AG, et al. Google's AI chatbot "Bard": a side-by-side comparison with ChatGPT and its utilization in ophthalmology. Eye (Lond). 2024;38(4):642–5. doi:10.1038/s41433-023-02760-0.

9. Microsoft. Bing Chat | Microsoft Edge. Accessed July 4, 2024. Available from: https://www.microsoft.com/en-us/edge/features/bing-chat.

10. Google. Google AI updates: Bard and new AI features in Search. Accessed July 4, 2024. Available from: https://blog.google/technology/ai/bard-google-ai-search-updates.

11. Kleinig O, Gao C, Kovoor JG, Gupta AK, Bacchi S, Chan WO. How to use large language models in ophthalmology: from prompt engineering to privacy protection. Eye (Lond). 2024;38(4):649–53. doi:10.1038/s41433-023-02772-w.

12. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on the USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198.

13. Antaki F, Touma S, Milad D, El-Khoury R, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci. 2023;3(4):100324. doi:10.1016/j.xops.2023.100324.

14. Sensoy E, Citirik M. A comparative study on the knowledge levels of artificial intelligence programs in diagnosing ophthalmic pathologies and intraocular tumors. Int Ophthalmol. 2023;43(12):4905–9. doi:10.1007/s10792-023-02893-x.

15. Şensoy E, Çıtırık M. Performance analysis of ChatGPT-3.5, Copilot and Gemini in multiple choice questions on ocular inflammation and uveitis: the effect of language differences. Turk Klin J Ophthalmol. 2025;34(1):12–6.