Received: 27 Jan 2025 | **Accepted:** 22 July 2025 **Genel Tıp Derg. Volume** 35/**Issue** 4 (August), 597-604

DOI: 10.54005/geneltip.1627508

ORIGINAL ARTICLE

Cross-Linguistic Evaluation of Artificial Intelligence Chatbots: Performance of ChatGPT-3.5, Copilot, and Gemini in Neuro-ophthalmologic Evaluation in English and Turkish

Yapay Zeka Sohbet Robotlarının Diller Arası Değerlendirmesi: ChatGPT-3.5, Copilot ve Gemini'nin Nöro-oftalmolojik Değerlendirmede İngilizce ve Türkçe Performansı

[1]Eyüpcan ŞENSOY, [2]Mehmet CITIRIK

[1]MD, FEBO, Department of Ophthalmology, Etlik City Hospital, University of Health Sciences, Ankara, Turkiye
E-mail: dreyupcansensoy@yahoo.com
[2]Prof., MD., Department of Ophthalmology, Etlik City Hospital, University of Health Sciences, Ankara, Turkiye
E-mail: mcitirik@hotmail.com

Correspondence

Eyüpcan ŞENSOY, MD, FEBO
Varlık Mahallesi, Halil Sezai Erkut Caddesi Yenimahalle, Ankara, Turkiye

E-Mail: dreyupcansensoy@yahoo.com

How to cite ?

Şensoy E., Cıtırık M., Cross-Linguistic Evaluation of Artificial Intelligence Chatbots: Performance of ChatGPT-3.5, Copilot, and Gemini in Neuro-ophthalmologic Evaluation in English and Turkish, Genel Tıp Derg. 2025;35(4):597-604

ABSTRACT

Aim: To evaluate the performance of ChatGPT-3.5, Copilot, and Gemini artificial intelligence chatbots on the same questions in neuro-ophthalmologic evaluation in English and Turkish.
Methods: Forty questions related to neuro-ophthalmology were included in the study. After all English questions were translated into Turkish by a certified native speaker, both versions of the questions were asked to ChatGPT-3.5, Copilot, and Gemini chatbots. The answers were compared with the answer key and grouped as correct and incorrect. Their superiority over each other was compared statistically.
Results: ChatGPT-3,5 47.5%, Copilot 57.5%, and Gemini 32.5% answered the English questions correctly. ChatGPT-3.5 57.5%, Copilot 52.5%, and Gemini 32.5% answered the questions correctly in Turkish. No statistically significant difference was detected between chatbots in answering the same questions in English and Turkish, although there were different levels of success ($p>0.05$).
Conclusions: Although there is no statistically significant difference, chatbots can answer the same questions differently. In addition to improving the knowledge level of chatbots, their language skills also need to be improved.

Keywords: Artificial intelligence applications, ChatGPT-3.5, Copilot, English, Gemini, neuro-ophthalmology, Turkish

ÖZ

Amaç: ChatGPT-3,5, Copilot ve Gemini yapay zeka sohbet botlarının nöro-oftalmolojik değerlendirmede İngilizce ve Türkçe aynı sorulardaki performanslarını değerlendirmek.
Gereç ve Yöntemler: Nöro-oftalmoloji ile ilişkili 40 soru çalışmaya dahil edildi. Tüm İngilizce soruların sertifikasyonlu çevirmen (native speaker) tarafından Türkçeye çevirileri gerçekleştirildikten sonra soruların her iki versiyonu ChatGPT-3,5, Copilot ve Gemini sohbet botlarına soruldu. Verilen cevaplar cevap anahtarı ile karşılaştırılarak doğru ve yanlış olarak gruplandırıldı. Birbirlerine üstünlükleri istatistiksel olarak karşılaştırıldı.
Bulgular: Sorulan İngilizce sorulara ChatGPT-3,5 %47,5, Copilot %57,5 ve Gemini %32,5 oranında doğru cevap verdi. Sorulan Türkçe sorulara ChatGPT-3,5 %57,5, Copilot %52,5 ve Gemini %32,5 oranında doğru cevap verdi. Sohbet botları arasında, İngilizce ve Türkçe aynı soruları cevaplamada farklı başarı düzeyi olduğu halde, istatistiksel olarak anlamlı başarı farkı tespit edilmedi ($p>0,05$).
Sonuçlar: İstatistiksel olarak anlamlı bir fark izlenmemesine rağmen sohbet botları aynı sorulara farklı cevaplar verebilmektedir. Sohbet botlarının bilgi düzeylerinin geliştirilmesinin yanında dil becerilerinin de geliştirilmeye ihtiyacı vardır.

Anahtar Kelimeler: ChatGPT-3,5, Copilot, Gemini, İngilizce, nöro-oftalmoloji, Türkçe, yapay zeka uygulamaları.

Peer-Review: Double anonymized - Two External
Plagiarism Checks: Yes - intihal.net
Complaints: geneltip@selcuk.edu.tr

Selcuk University Press
Genel Tıp Dergisi | e-ISSN: 2602-3741
https://dergipark.org.tr/tr/pub/geneltip
https://yayinevi.selcuk.edu.tr/

INTRODUCTION

Neuro-ophthalmology is an important sub-branch of ophthalmology that connects ophthalmology and neurology, examining eye movements, visual pathways, and visual processing, and focusing on their pathologies. This science is a special field that requires intensive synthesis of information, such as history, neuroimaging, and laboratory tests, while examining the relationship between vision and life-threatening conditions (1). In a study, it was stated that approximately half of the patient referrals to neuro-ophthalmology were misdiagnosed and that patients were even harmed by the misdiagnosis during this process (2). Considering all these situations, it is clear that ophthalmologists need to receive intensive neuro-ophthalmology training in order to be able to diagnose and treat neuro-ophthalmological diseases effectively. In addition, the number of neuro-ophthalmologists who have received this training is quite low (3). One study stated that the average wait time for neuro-ophthalmology access in the United States was 6 weeks (4). These conditions pose significant challenges for both patients and general ophthalmologists in accurately diagnosing and treating neuro-ophthalmological disorders (1, 2). With the advancement of technology, artificial intelligence applications have become widely used in ophthalmology, especially since 2015, and have become an important resource for the diagnosis and treatment monitoring of diseases (5). One example of these artificial intelligence applications is the Large Language Model (LLM) based chatbots, an important branch of artificial intelligence that has recently been investigated for use in the field of ophthalmology. ChatGPT-3.5 (OpenAI), Copilot (Microsoft), and Gemini (Google AI) are important representatives of this group, and they are freely accessible (6).

Our study aims to investigate the effects of language differences (English and Turkish) on the success of ChatGPT-3.5, Copilot, and Gemini artificial intelligence chatbots in multiple-choice neuro-ophthalmology questions.

MATERIALS and METHODS

All 40 questions in the American Academy of Ophthalmology 2023-2024 Basic and Clinical Science Course (BCSC) Neuro-ophthalmology book study questions section were included in the study (7). The study questions were translated into Turkish by a certified translator (native speaker). The questions were applied in both English and Turkish versions in the same order on July 16, 2024, by opening a new account on the freely accessible ChatGPT-3.5 (OpenAI; San Francisco, CA), Copilot (Microsoft, Redmond, WA) and Gemini (Google, Mountain View, California, United States) artificial intelligence chat bots. Before each question was asked to the artificial intelligence chatbots, the programs were given the command, 'I will ask you multiple-choice questions. Please give me the correct answer option.' After each question, the session was closed and restarted. Additionally, a command was added to the given command to mitigate the memory effect: 'Do not memorize the questions and answers we ask you.' The answers given by the chatbots to the questions were compared with the answer key at the back of the book and grouped as correct or incorrect.

Since the data in our study is not from any

animal or human sources, ethics committee approval is not required.

Statistical Analysis

Statistical Package for the Social Sciences version 23 (SPSS Inc., Chicago, IL, USA) program was used for statistical analysis of the data, and percentage values were calculated. The Marascuilo test procedure was applied to compare the levels of answering questions of independent groups. The McNemar test is used to examine nominal parameters in dependent groups. A P value below 0.05 was accepted as the significance level.

RESULTS

Forty multiple-choice questions on neuro-ophthalmology were administered to artificial intelligence chatbots in English. ChatGPT-3.5 provided correct answers to 19 (47.5%) and incorrect answers to 21 (52.5%) of the questions. Copilot provided correct answers to 23 (57.5%) and incorrect answers to 17 (42.5%) of the questions. Gemini provided correct answers to 13 (32.5%) and incorrect answers to 27 (67.5%) of the questions (Table 1). No statistically significant difference was found between ChatGPT-3.5's level of correctly answering the English versions of the questions and Copilot and Gemini (ChatGPT-3.5 vs. Copilot, Value = 0.100, Critical Range = 0.218; ChatGPT-3.5 vs. Gemini: Value = 0.150, Critical Range = 0.212). Copilot's level of correctly answering the English versions of the questions was statistically significantly higher than Gemini's (Copilot vs. Gemini, Value = 0.250, Critical Range = 0.211).

Forty multiple-choice questions on neuro-ophthalmology were administered to artificial intelligence chatbots in Turkish. ChatGPT-3.5 gave correct answers to 23 (57.5%) and incorrect answers to 17 (42.5%) of the questions. Copilot gave correct answers to 21 (52.5%) and incorrect answers to 19 (47.5%) of the questions. Gemini gave correct answers to 13 (32.5%) and incorrect

Table 1: Responses of artificial intelligence chatbots to the same multiple-choice questions related to neuro-ophthalmology and their changes

Answers	ChatGPT-3.5 (English)	ChatGPT-3.5 (Turkish)	Copilot (English)	Copilot (Turkish)	Gemini (English)	Gemini (Turkish)
Correct	19 (47.5%)	23 (57.5%)	23 (57.5%)	21 (52.5%)	13 (32.5%)	13 (32.5%)
Incorrect	21 (52.5%)	17 (42.5%)	17 (42.5%)	19 (47.5%)	27 (67.5%)	27 (67.5%)
***p*-value**	0.454*		0.791*		>0.999*	
Producing the same answers	24 (60%)		27 (67.5%)		26 (65%)	
Producing different answers	16 (40%)		13 (32.5%)		14 (35%)	
Correct answers when asked in English	6 (37.5)		8 (61.6%)		8 (57.1%)	
Correct answers when asked in Turkish	10 (62.5%)		5 (38.4%)		6 (42.9%)	

*: McNemar test

answers to 27 (67.5%) of the questions (Table 1). No statistically significant difference was found between Copilot's level of correct answering the Turkish versions of the questions and ChatGPT-3.5 and Gemini's level of correct answering (ChatGPT-3.5 vs. Copilot, Value=0.05, Critical Range=0.218; Copilot vs. Gemini: Value=0.200, Critical Range=0.212). ChatGPT-3.5's ability to correctly answer the Turkish versions of the questions was statistically significantly higher than Gemini's (ChatGPT-3.5 vs. Gemini, Value=0.250, Critical Range=0.211).

ChatGPT-3.5 gave the same answer to 24 (60%) of the English and Turkish questions, while it gave different answers to 16 (40%) questions. Of the questions that produced different answers to the English and Turkish versions, 10 (62.5%) were answered correctly when asked in Turkish, while 6 (37.5%) were answered correctly when asked in English. No significant difference was found in terms of performance in answering the questions in English and Turkish (p=0.454, McNemar test) (Table 1). No statistically significant agreement was observed between ChatGPT-3.5's answers to the English and Turkish versions of the questions (Cohen's κ = 0.206, p=0.184).

Copilot gave the same answer to 27 (67.5%) of the English and Turkish questions while giving different answers to 13 (32.5%) questions. Of the questions that produced different answers to the English and Turkish versions, 5 (38.4%) were answered correctly when asked in Turkish, while 8 (61.6%) were answered correctly when asked in English. No significant difference was found in terms of performance in answering the questions in English and Turkish (p=0.791, McNemar test) (Table 1). A fair agreement was observed between Copilot's answers to the English and Turkish versions of the questions (Cohen's κ = 0.392, p=0.011).

Gemini gave the same answer to 26 (65%) of the English and Turkish questions while giving different answers to 14 (35%) questions. Of the questions that produced different answers to the English and Turkish versions, 6 (42.9%) were answered correctly when asked in Turkish, while 8 (57.1%) were answered correctly when asked in English. No significant difference was found in terms of performance in answering the questions in English and Turkish (p>0.999, McNemar test) (Table 1). No statistically significant agreement was observed between Gemini's answers to the English and Turkish versions of the questions (Cohen's κ = 0.202, p=0.201).

DISCUSSION

Advances in technology have also made their impact felt in artificial intelligence applications and have added a wide range of new features to these programs. This development has increased the usability of the programs in various fields, and their various advantages and disadvantages have become a frequently researched topic (8, 9). In this study, the success of these artificial intelligence programs in different languages in the field of neuro-ophthalmology has been examined, and the factors affecting this situation have been tried to be understood. Our study shows that artificial intelligence programs do not answer the same questions in different languages in the same way. This situation brings the trust in artificial intelligence back to the agenda and reveals the need for development.

In the past, artificial intelligence

applications were primarily deep learning-based artificial intelligence applications which focused on understanding and recognizing patterns. With the latest developments in artificial intelligence technologies, LLM-based applications have emerged. These applications are current artificial intelligence applications that can understand the words that come before them, evaluate this word structure, establish the connection between them, and, as a result, derive answers that are appropriate for the human mindset (10). These LLM-based applications have found a place in medical education through their various benefits and have become applications that can successfully perform many tasks on their own, such as rapid access to accurate information, literature review, summarization, and language translation (11). In addition, it has managed to attract the attention of health professionals at all levels thanks to its ability to explain the diagnosis, differential diagnosis, and treatment modalities of a wide variety of diseases (12). Recently, these beneficial features of chatbots in the field of medicine have also been popular in the field of ophthalmology, and neuro-ophthalmology has been an important subject where various benefits have been investigated. In a study where the diagnosis of 22 cases commonly seen in neuro-ophthalmology was asked ChatGPT-3.0 and ChatGPT-4.0 and their success levels were compared with ophthalmologists, it was stated that ChatGPT-3.0 and ChatGPT-4.0 gave correct answers to the cases at a rate of 59% and 82%, respectively, and researchers stated that with the development of these programs, ChatGPT-4.0, in particular, could help in the rapid and accurate diagnosis of neuro-ophthalmological diseases (1). In another study, the diagnosis of 10 neuro-ophthalmology cases was asked to ChatGPT-3.5, Bing, and Gemini programs, and the programs answered 4 out of 10 cases correctly. Looking at this result, researchers concluded that artificial intelligence programs could be potentially useful and stated that these programs could be used as consultants (13). The success of artificial intelligence chatbots on multiple-choice questions has also been studied. In a study that included three hundred questions, it was stated that ChatGPT-3.0 answered 55% of the questions correctly, while ChatGPT-4.0 answered 70% of the questions correctly. These programs answered neuro-ophthalmology questions correctly at 57% and 54%, respectively, and there was no statistical difference between their success on neuro-ophthalmology questions. Based on these results, researchers have stated that these programs can gain an important place in medical education with their development (14). In another study examining the success of ChatGPT-3.5 and Bing in ophthalmology questions, 913 multiple-choice questions were applied to these programs; ChatGPT-3.5 answered 59.7% of the questions correctly, while Bing answered 73.6%. Neuro-ophthalmology questions were answered 64% and 80% correctly, respectively. The researchers stated that Bing's internet access could provide additional contributions to these results and could guide researchers thanks to its citation feature, but that the information provided should be approached critically (15). In a study evaluating the success of artificial intelligence chatbots in multiple-choice questions related to neuro-ophthalmology and comparing them with

neuro-ophthalmologists, it was stated that the joint production of answers by experts and artificial intelligence increased the success rate and that this situation should be investigated in more detail (16). In a different study investigating the existence of changes in multiple-choice questions depending on the region where internet access is provided, Gemini showed success rates ranging from 40% to 80% in neuro-ophthalmology questions, and it was stated that the performance of chatbots may vary depending on the countries where access is provided. Based on these results, the researchers stated that this issue should be investigated to improve the performance of chatbots (17). In another study examining the success of artificial intelligence chatbots in Turkish ophthalmology questions, ChatGPT-3.5, Bing, and Bard answered questions correctly at 51%, 63%, and 45.5%, respectively; and they were successful at 44.4%, 66.7%, % and 33.3% in neuro-ophthalmology and strabismus questions, respectively. The researchers interpreted these results as indicating that LLM-based applications are potentially useful but need to be developed (18).

While we acknowledge that newer models such as Gemini Advanced and GPT-4 may provide more accurate results, our study focused on ChatGPT-3.5 due to its accessibility and widespread recognition. Future research could incorporate these newer models to compare their performance and evaluate whether they provide improved responses in neuro-ophthalmology tasks, particularly considering their advancements in both language processing and domain-specific knowledge.

In this study, we focused on a different topic. To our knowledge, our study is the first to investigate the effect of asking the same neuro-ophthalmology questions in English or Turkish on the performance of artificial intelligence chatbots. In our study, while artificial intelligence chatbots were more successful than Copilot Gemini in English questions, we did not find a significant difference in success between the other programs. While ChatGPT-3.5 was more successful than Gemini in answering Turkish questions, we did not find a significant difference between the other chatbots. These differences in performance may stem from variations in how the models interpret questions, as well as discrepancies in their knowledge base and language processing abilities. However, the fact that the success advantages of these programs differ between languages also creates an important dilemma. In addition, when chatbots were evaluated within themselves, each of them showed similar performance in both English and Turkish questions. However, when these performances were examined, we determined that the chatbots produced different answers to the same questions. Except for Copilot, we could not observe a statistically significant agreement in ChatGPT-3.5 and Gemini's answers to questions in different languages. In Copilot's responses, the agreement was statistically significant but remained at a fair level. This further increased the distrust in the responses of artificial intelligence programs to different languages. We believe that this situation may be caused by chatbots' deficiencies in their ability to make sense of things, associate words, and translate languages. We also believe that differences in the English and Turkish

literature that programs can access may have an impact on question answering levels in different languages. Although we could not detect a significant difference, chatbots giving different answers to the same questions may undermine users' trust in chatbots who are looking for accurate information. We believe that this is another important issue that needs to be investigated in the development of artificial intelligence chatbots.

The study questions of the American Academy of Ophthalmology 2023-2024 Basic and Clinical Science Course (BCSC) Neuro-ophthalmology book, which contains important information and is among the basic books, included 40 questions, and we asked these to chatbots. We foresee that the small number of questions may affect whether the statistical result is significant or not. However, we did not find it appropriate to add additional questions to the questions in this book, which measure basic information. We can foresee that it should be investigated whether different values will be obtained in tests with more questions.

Our study had several limitations. The inclusion of single-source questions in the study, the small number of questions, and the fact that the questions were asked to the chatbots only once constitute these limitations of the study.

CONCLUSION

As a result, although there is no difference in performance between chatbots in answering the same questions in different languages, chatbots can produce different answers to the same questions in different languages. This situation can be a significant obstacle to the preference of chatbots by users to access the correct information. While this study provides important insights into the performance of artificial intelligence chatbots in neuro-ophthalmology, further research is needed to address the discrepancies in performance across languages. Future studies should explore the potential of incorporating newer artificial intelligence models like GPT-4 or Gemini Advanced to evaluate whether they offer improvements in accuracy and consistency. Additionally, expanding the study to include a larger dataset and multiple question formats could provide more robust conclusions on the utility of artificial intelligence chatbots in medical education and clinical practice.

Conflict of interest

The author declares that there is no conflict of interest.

Financial support

There was no funding for this study.

Acknowledgement

None

REFERENCES

1. Madadi Y, Delsoz M, Lao PA, Fong JW, Hollingsworth T, Kahook MY, et al. ChatGPT Assisting Diagnosis of Neuro-ophthalmology Diseases Based on Case Reports. medRxiv 2023.

2. Stunkel L, Sharma RA, Mackay DD, Wilson B, Van Stavern GP, Newman NJ, et al. Patient Harm Due to Diagnostic Error of Neuro-Ophthalmologic Conditions. Ophthalmology 2021; 128:1356–1362.

3. Frohman LP. The human resource crisis in neuro-ophthalmology. J Neuroophthalmol 2008; 28:231–234.

4. Debusk A, Subramanian PS, Scannell Bryan M, Moster ML, Calvert PC, and Frohman LP. Mismatch in Supply and Demand for Neuro-Ophthalmic Care. J Neuroophthalmol 2022; 42:62–67.

5. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY,

Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019; 103:167.

6. Sensoy E and Citirik M. A comparative study on the knowledge levels of artificial intelligence programs in diagnosing ophthalmic pathologies and intraocular tumors evaluated their superiority and potential utility. Int Ophthalmol 2023; 43:4905–4909.

7. Bhatti TM, Chen JJ, Danesh-Meyer H V., Levin LA, Moss HE, Philips PH, et al., editors. Neuro-Ophthalmology. San Francisco: American Academy of Ophthalmology; 2023.

8. Şensoy E, Çıtırık M. ChatGPT-3.5, Copilot ve Gemini'nin oküler inflamasyon ve üveit konusundaki çoktan seçmeli sorularda performans analizi: Dil farklılıklarının etkisi: Kesitsel araştırma. Turkiye Klinikleri J Ophthalmol 2025;34:12-16

9. Şensoy E and Çıtırık M. Performance of ChatGPT-3.5, Copilot, and Gemini in answering English and Turkish questions related to ocular surface diseases and cornea: a comparison study. Turkish Journal of Clinical and Experimental Ophthalmology 2025; 20:37-41.

10. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS digital health 2023; 2:e0000198.

11. Khan RA, Jawaid M, Khan AR, and Sajjad M. ChatGPT - Reshaping medical education and clinical management. Pak J Med Sci 2023; 39:605.

12. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports. 2022.

13. Shukla R, Mishra AK, Banerjee N, and Verma A. The Comparison of ChatGPT 3.5, Microsoft Bing, and Google Gemini for Diagnosing Cases of Neuro-Ophthalmology. Cureus 2024; 16.

14. Haddad F and Saade JS. Performance of ChatGPT on Ophthalmology-Related Questions Across Various Examination Levels: Observational Study. JMIR Med Educ 2024; 10:e50842.

15. Tao BKL, Hua N, Milkovich J, and Micieli JA. ChatGPT-3.5 and Bing Chat in ophthalmology: an updated evaluation of performance, readability, and informative sources. Eye 2024 2024; 1–6.

16. Tailor PD, Dalvin LA, Starr MR, Tajfirouz DA, Chodnicki KD, Brodsky MC, et al. A Comparative Study of Large Language Models, Human Experts, and Expert-Edited Large Language Models to Neuro-Ophthalmology Questions. J Neuroophthalmol 2024.

17. Mihalache A, Grad J, Patil NS, Huang RS, Popovic MM, Mallipatna A, et al. Google Gemini and Bard artificial intelligence chatbot performance in ophthalmology knowledge assessment. Eye 2024 2024; 1–6.

18. Canleblebici M, Dal A, and 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. Turkiye Klinikleri J of Ophthalmol 2024;33:163-170.