Performance of ChatGPT-3.5, Copilot and Gemini in Answering English and Turkish Questions Related to Ocular Surface Diseases and Cornea: A Comparison Study

Eyüpcan Şensoy[1], *Mehmet Çıtırık*[1]

ABSTRACT

Purpose: To evaluate the performance of ChatGPT-3.5, Copilot, and Gemini artificial intelligence chatbots in answering the same questions in English and Turkish related to ocular external diseases and cornea.

Materials and Methods: Both English and Turkish versions of 41 multiple choice questions related to ocular external diseases and cornea were applied to ChatGPT-3.5, Copilot, and Gemini artificial intelligence chatbots. These questions were translated into Turkish by a certified native speaker. The answers given by the chatbots were compared with the answer key and grouped as correct and incorrect. The success rates of chatbots were compared statistically.

Results: In the English version of the questions, ChatGPT-3.5 provided correct answers at 53.7%, Copilot at 43.9%, and Gemini at 51.2% accuracy (p=0.655). In the Turkish version of the questions, ChatGPT-3.5 provided correct answers at 48.8%, Copilot at 41.5%, and Gemini at 43.9% accuracy (p=0.794). There was no statistically significant difference between chatbots in answering the Turkish versions of the questions, although there were fewer correct answers in all three applications (p>0.05).

Conclusion: Although artificial intelligence chatbots are a promising tool for obtaining information, they need to be developed and their performance improved both in terms of their knowledge level and ability to interpret and translate the meaning in different languages.

Keywords: ChatGPT-3.5, Copilot, English and Turkish, Gemini, Ocular surface diseases and Cornea.

INTRADUCTION

With the rapid developments in technology in recent times, there has been great progress in artificial intelligence technologies, and with these developments, it has become a frequently mentioned topic in all fields of medicine.[1] Deep learning-based artificial intelligence programs, an example of this group, have attracted intense interest in the field of ophthalmology, especially after 2015, and have begun to play an active role in the diagnosis and follow-up of a wide variety of diseases.[2,3] Another example of this group is artificial intelligence programs based on Large Language Models (LLMs). LLMs are a branch of artificial intelligence that can perceive data, summarize it, contextually evaluate these inputs, offer various solution possibilities, and mimic human thinking while doing so.[4] These programs have provided medical researchers with a wide range of benefits, including the ability to search literature and summarize and analyze data.[5] The wide range of benefits of artificial intelligence programs has led to their more widespread use in the field of ophthalmology and their performance being examined more frequently.[6–10]

The aim of our study is to investigate the effects of language differences on the success levels of ChatGPT-3.5 (OpenAI), Copilot (Microsoft), and Gemini (Google) artificial intelligence chatbots in multiple choice questions about ocular surface diseases and the cornea.

[1] Ankara Etlik Şehir Hastanesi, Göz Hastalıkları, Ankara, Türkiye

Received: 21.07.2024
Accepted: 31.10.2024
TJ-CEO 2025; 20: 37-41
DOI: 10.37844/TJ-CEO.2025.20.6

Corresponding author:
Eyüpcan Şensoy
Email: dreyupcansensoy@yahoo.com

MATERIALS AND METHODS

All 41 questions in the study questions section of the American Academy of Ophthalmology 2023-2024 Basic and Clinical Sciences External Diseases and Cornea book were included in the study.[11] The Turkish translations of the same questions were made by a certified translator (native speaker). The questions were applied to the artificial intelligence chatbots ChatGPT-3.5 (OpenAI; San Francisco, CA), Copilot (Microsoft, Redmond, WA), and Gemini (Google, Mountain View, California, United States), which have been made available for free by three major manufacturers, on July 11, 2024. Before the questions were applied to the artificial intelligence chatbots, the command 'I will ask you multiple choice questions. Please give me the correct answer option.' was given, and the chat session was ended after each question was applied. The answers given by the artificial intelligence programs to the questions were compared with the answer key at the back of the book and then grouped as correct or incorrect. Since our study does not contain data on human or animal subjects, ethics committee approval is not required.

ChatGPT-3.5

This LLM-based program, which has the ability to mimic human intelligence, has been trained with a very large data network of approximately 175 billion, thus finding a strong place for itself in its group.[12] While the disadvantages of the ChatGPT-3.5 artificial intelligence chatbot are that it was last updated in September 2021 and does not have internet access, its advantages include being accessible for free and having a very wide information system.[13]

Copilot

A LLM-based program integrated with GPT-4 as of February 2023, Copilot is an artificial intelligence chatbot with up-to-date internet access.[12] It provides guidance to the researcher in specifying the sources of the information they provide and thus providing more detailed information on the relevant subject.[6,14]

Gemini

A LLM-based chatbot trained with a vast knowledge network, Gemini can analyze a wide range of contextual data and produce precise answers to various problems it encounters. It has active internet access.[15,16]

Statistical Analysis

Statistical Package for the Social Sciences version 23 (SPSS Inc., Chicago, IL, USA) program was used for statistical analysis. Percentage values were calculated. McNemar test was used for comparing two dependent nominal groups, and the Pearson chi-square test was used for comparing independent nominal data. P<0.05 was accepted as the statistical significance level.

RESULTS

Forty-one English questions related to ocular surface diseases and the cornea were applied to AI chatbots. The ChatGPT-3.5 gave correct answers to 22 (53.7%) and incorrect answers to 19 (46.3%) of the questions. The copilot gave correct answers to 18 (43.9%) and incorrect answers to 23 (56.1%) of the questions. The Gemini gave correct answers to 21 (51.2%) and incorrect answers to 20 (48.8%) of the questions (Table 1). There was no statistically significant difference in the success of the three artificial intelligence chatbots in answering English questions (p=0.655 Pearson chi-square test).

The Turkish versions of the questions were applied to the chatbots. The ChatGPT-3.5 gave correct answers to 20 (48.8%) and incorrect answers to 21 (51.2%) of the questions. The Copilot gave correct answers to 17 (41.5%) and incorrect answers to 24 (58.5%) of the questions. The Gemini gave correct answers to 18 (43.9%) and incorrect answers to 23 (56.1%) of the questions (Table 1). There was no statistically significant difference in the levels of success in answering the Turkish questions among the three artificial intelligence chatbots (p=0.794 Pearson chi-square test).

ChatGPT-3.5 gave the same answer to 31 (75.6%) of the English and Turkish questions, while it gave different answers to 10 (24.4%). Of the questions to which it gave different answers, 6 (60%) were answered incorrectly when asked in Turkish, while 4 (40%) were answered correctly when asked in Turkish. No statistically significant difference was observed in ChatGPT-3.5's ability to answer the same questions correctly in English and Turkish (p=0.754 McNemar test) (Table 2).

Copilot gave the same answer to 30 (73.2%) of the English and Turkish questions, while giving different answers to 11 (26.8%). Of the questions to which it gave different answers, 6 (54.5%) were answered incorrectly when asked in Turkish, while 5 (45.5%) were answered correctly when asked in Turkish. No statistically significant difference was observed in the Copilot's correct answering of the same questions in English and Turkish (p=1.0 McNemar test) (Table 2).

Gemini gave the same answer to 26 (63.4%) of the English and Turkish questions while giving different answers to 15 (36.6%). Of the questions to which it gave different answers, 10 (66.7%) were answered incorrectly when asked in Turkish, while 5 (33.3%) were answered correctly when asked in Turkish. No statistically significant difference was observed in Gemini's correct answers to the same questions in English and Turkish (p=0.607 McNemar test) (Table 2).

DISCUSSION

In recent years, with the developing technology, the suggestion of increasing the integration of artificial intelligence with medical education and clinical applications

Table 1: *Accuracy Levels of ChatGPT-3.5, Copilot, and Gemini in Answering Multiple-Choice Questions Related to Ocular Surface Diseases and the Cornea in English and Turkish*

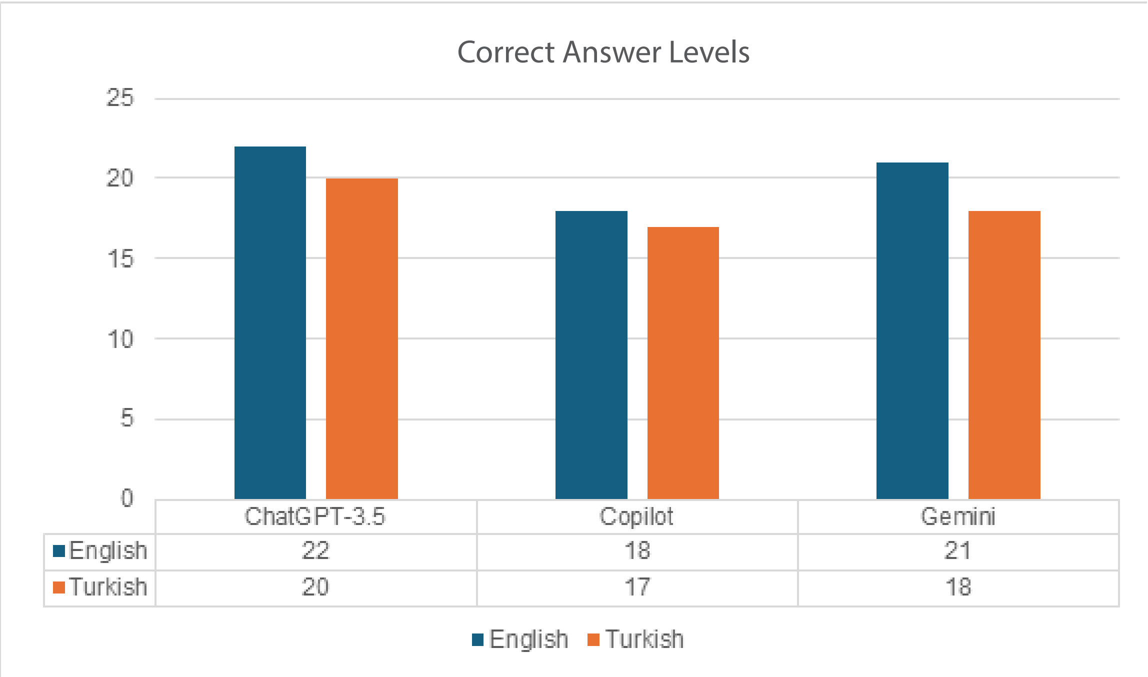

Table 2: *Responses and Changes Provided by Artificial Intelligence Chatbots to the Same Questions Related to Ocular Surface Diseases and the Cornea*

Answers	**ChatGPT-3.5 (English)**	**ChatGPT-3.5 (Turkish)**	**Copilot (English)**	**Copilot (Turkish)**	**Gemini (English)**	**Gemini (Turkish)**
Correct	22 (53.7%)	20 (48.8%)	18 (43.9%)	17 (41.5%)	21 (51.2%)	18 (43.9%)
Incorrect	13 (36.1%)	21 (51.2%)	23 (56.1%)	24 (58.5%)	20 (48.8%)	23 (56.1%)
P value	0.754*		1.0*		0.607*	
Giving the same answer	31 (75.6%)		30 (73.2%)		26 (63.4%)	
Giving a different answer	10 (24.4%)		11 (26.8%)		15 (36.6%)	
Correct-incorrect change	6 (60%)		6 (54.5%)		10 (66.7%)	
Incorrect-incorrect change	4 (40%)		5 (45.5%)		5(33.3%)	

*: McNemar test

has become widespread in every department of medicine.[13] One of these integration issues is that chatbots can be used as a consultant in accessing the right information.[5] Interest in this topic has increased significantly in recent times, and their success in medical questions has become a frequently tested topic. For example, in a study examining ChatGPT's success in answering questions on the USMLE, it was stated that the program answered 60% of the questions correctly, and it was suggested that ChatGPT would play an active role in the clinical decision-making process in the future.[17] These programs have also been frequently researched in the field of ophthalmology, and their performances have been examined in different subjects.[6–10] Haddad et al. tested the success of ChatGPT-3.5 and ChatGPT-4.0 programs on 380 ophthalmology questions and reported that their success rates were 55% and 70%, respectively. They also evaluated their success on cornea-related questions and reported that ChatGPT-3.5 was 66% successful and ChatGPT-4.0 was 74% successful and that these programs were not superior to each other in answering cornea questions.[9] In a study testing the success of ChatGPT-3.5 and Bing in ophthalmology questions, 913 questions were applied to artificial intelligence chatbots, and it was stated that ChatGPT-3.5 answered 59.69% of the questions correctly, while Bing answered 73.6% of the questions correctly. As a result of this data, the authors stated that Bing's information accessibility is more advanced and can be useful for ophthalmology students.[13] Canlebleci et al. evaluated the success of ChatGPT-3.5, ChatGPT-4.0, Bing, and Bard in Turkish ophthalmology questions and asked 200 multiple-choice questions to these programs. It was stated that the success of ChatGPT-3.5, ChatGPT-4.0, Bing, and Bard in answering these questions was 51%, 77.5%, 63%, and 45.5%, respectively, and the researchers emphasized that although LLMs have promising successes, their continuous development is still necessary.[10] Another study examined the success of artificial intelligence chatbots in ophthalmology questions according to the region where internet access is provided, and it was stated that their success in answering questions about ocular surface diseases and cornea was between 50% and 90%, depending on the country where internet access is available.[7]

In this study, we examined the success of artificial intelligence chatbots in Turkish and English versions of the same questions. Although it was determined that the effect of applying the questions in different languages on the success of the chatbots was not statistically significant, the success of artificial intelligence chatbots in Turkish versions decreased, and most of the questions that were answered differently consisted of questions that were answered correctly when asked in English but incorrectly when asked in Turkish. This situation may have arisen because the literature is often composed of English sources and the chatbots' weakness in understanding, interpreting, and language translation abilities of this information. The study questions of the American Academy of Ophthalmology 2023-2024 Basic and Clinical Science Course (BCSC) External Diseases and Cornea book, which contains important information and is among the basic books, included 41 questions, and we asked these to the 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 of this book, which measures basic knowledge. We can foresee that it should be investigated whether different values will be obtained in tests that include more questions.

The most important limitations of our study are the small number of questions, the inability to separate the questions into sub-topic branches, and the inability to investigate the existence of their superiority over each other.

As a result, our study is the first to examine the performance of three free artificial intelligence chatbots, ChatGPT-3.5, Copilot, and Gemini, on the same questions related to ocular surface diseases and the cornea in English and Turkish. For artificial intelligence chatbots to be used as accurate information tools in answering questions in different languages, their knowledge levels should be increased, and their ability to understand, interpret, and translate different languages should be developed.

REFERENCES

1. Evans RS. Electronic Health Records: Then, Now, and in the Future. Yearb Med Inform 2016;25:S48-S61. doi:10.15265/IYS-2016-S006/ID/JRS006-48/BIB
2. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019;103:167-175. doi:10.1136/BJOPHTHALMOL-2018-313173
3. Antaki F, Coussa RG, Kahwati G, et al. Accuracy

of automated machine learning in classifying retinal pathologies from ultra-widefield pseudocolour fundus images. Br J Ophthalmol 2023;107:90-95. doi:10.1136/BJOPHTHALMOL-2021-319030

4. Radford A, Wu J, Child R, et al. Language Models are Unsupervised Multitask Learners. Accessed June 26, 2023. https://github.com/codelucas/newspaper
5. Gao CA, Howard FM, Markov NS, et al. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. bioRxiv. Published online December 27, 2022:2022.12.23.521610. doi:10.1101/2022.12.23.521610
6. Tao BKL, Hua N, Milkovich J, et al. ChatGPT-3.5 and Bing Chat in ophthalmology: an updated evaluation of performance, readability, and informative sources. Eye 2024. Published online March 20, 2024:1-6. doi:10.1038/s41433-024-03037-w
7. Mihalache A, Popovic MM, Muni RH. Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. JAMA Ophthalmol 2023;141:589-597. doi:10.1001/JAMAOPHTHALMOL.2023.1144
8. Mihalache A, Grad J, Patil NS, et al. Google Gemini and Bard artificial intelligence chatbot performance in ophthalmology knowledge assessment. Eye 2024. Published online April 13, 2024:1-6. doi:10.1038/s41433-024-03067-4
9. Haddad F, Saade JS. Performance of ChatGPT on Ophthalmology-Related Questions Across Various Examination Levels: Observational Study. JMIR Med Educ 2024;10:e50842. doi:10.2196/50842
10. 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. Turkiye Klinikleri Journal of Ophthalmology. Published online June 11, 2024.
11. Feder RS, Berdy GJ, Luorno JD, et al., eds. External Disease and Cornea. American Academy of Ophthalmology; 2023.
12. Wen J, Wang W. The future of ChatGPT in academic research and publishing: A commentary for clinical and translational medicine. Clin Transl Med 2023;13. doi:10.1002/CTM2.1207
13. Tao BKL, Hua N, Milkovich J, et al. ChatGPT-3.5 and Bing Chat in ophthalmology: an updated evaluation of performance, readability, and informative sources. Eye 2024. Published online March 20, 2024:1-6. doi:10.1038/s41433-024-03037-w
14. Bing Chat | Microsoft Edge. Accessed July 4, 2024. https://www.microsoft.com/en-us/edge/features/bing-chat?form=MT00D8
15. Waisberg E, Ong J, Masalkhi M, et al. Google's AI chatbot "Bard": a side-by-side comparison with ChatGPT and its utilization in ophthalmology. Eye 2023 38:4. 2023;38:642-645. doi:10.1038/s41433-023-02760-0
16. Google AI updates: Bard and new AI features in Search. Accessed July 4, 2024. https://blog.google/technology/ai/bard-google-ai-search-updates/
17. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS digital health. 2023;2:e0000198. doi:10.1371/journal.pdig.0000198