Ankyra Medical Journal
Original Article
DOI: 10.51271/ANKMJ-0014

Exploring the competence of artificial intelligence programs in the field of oculofacial plastic and orbital surgery

Eyüpcan Şensoy, Mehmet Çıtırık
Department of Ophthalmology, Ankara Etlik City Hospital, Ankara, Turkiye

Cite this article: Şensoy E, Çıtırık M. Exploring the competence of artificial intelligence programs in the field of oculofacial plastic and orbital surgery. *Ank Med J.* 2024;3(3):63-65.

Received: 15.05.2024 ◆ **Accepted**: 20.05.2024 ◆ **Published**: 28.05.2024

ABSTRACT

Aims: It aims to evaluate the knowledge level of ChatGPT, Bing, and Bard artificial intelligence chatbots developed based on large language models (LLM) about oculofacial plastic surgery and to investigate the presence of superiority over each other.

Methods: Twenty-nine questions that tested knowledge about oculofacial plastic and orbital surgery were taken from the study questions section of the American Academy and Ophthalmology 2022-2023 Basic and Clinical Science Course Oculofacial Plastic and Orbital Surgery. The questions were asked to ChatGPT, Bing, and Bard programs, which are current artificial intelligence chatbots. The questions were classified as either correct or incorrect.

Results: ChatGPT gave 44.8% correct answers, Bing 48.3% correct answers, and Bard 58.6% correct answers to 29 questions about artificial intelligence chatbots. No statistical difference was observed between the rates of correct and incorrect answers given by 3 the intelligence programs (p=0.609, Pearson's chi-squared test).

Conclusion: The use of artificial intelligence to access information regarding oculofacial plastic and orbital surgery may provide limited benefits. Care should be taken in terms of accuracy and timeliness when evaluating the results of artificial intelligence programs.

Keywords: Bard, Bing, ChatGPT, oculofacial plastic surgery, orbital surgery

INTRODUCTION

Artificial intelligence is a sub-branch of computer science that aims to provide answers similarly by imitating the human mind.[1] The implementation of the first studies dates back to the early 1970s.[2] Although it has a wide variety of features, examples can be given to recognizing images, generating ideas to solve problems, and comprehending what is spoken.[3] Large language models (LLM), a sub-branch of artificial intelligence applications that can perceive inputted information, evaluate and summarize them, predict a wide variety of meaning possibilities, evaluate all these in a broad context, and draw conclusions as a result.[4] The development of LLMs has revolutionized the development of artificial intelligence chatbots.[5] Examples of these developed chatbots as Chat generative pre-trained transformer (ChatGPT) produced by OpenAI, Bing produced by Microsoft, and Bard artificial intelligence chatbots produced by Google AI.

With these recent developments, artificial intelligence programs have led to the formation of a new academic environment that has gained features such as synthesizing, processing, and analyzing data. Considering these advantages, interest in artificial intelligence programs in medical sciences has also increased.[6] They were even used in the writing of some articles and they were mentioned as the author.[7,8]

Our aim in this study is to evaluate the knowledge level of ChatGPT, Bing, and Bard artificial intelligence programs, which are offered free of charge by 3 different manufacturers with the increasing developments in artificial intelligence, in the field of oculofacial plastic and orbital surgery and to investigate the presence of superiority to each other.

METHODS

All 29 questions testing knowledge about oculofacial plastic and orbital surgery from the study questions section of the American Academy of Ophthalmology 2022-2023 Basic and Clinical Science Course Oculofacial Plastic and Orbital Surgery book were included in the study.[9] The questions were asked separately on 15 July 2023 to ChatGPT GPT-3.5 (OpenAI; San Francisco, CA), Bing (Microsoft, Redmond, Washington), and Bard (by Google) artificial intelligence chatbots, which can be used for free. First, we ask multiple-choice questions. Tell me the correct answer option."

Corresponding Author: Eyüpcan Şensoy, dreyupcansensoy@yahoo.com

command was given. After each question, the chat session was restarted to exclude memory retention features of artificial intelligence programs. The answers to the questions were compared with the answer keys and categorized as correct or incorrect. Additionally, common answers to the same question were determined and grouped as correct or incorrect.

Statistical Analysis

Statistical Package for the Social Sciences version 23 (SPSS Inc., Chicago, IL, USA) was used for statistical analysis of the data. The percentages were calculated using descriptive statistics. Pearson chi-square and Yates chi-square tests were used for the statistical comparison of independent nominal values. Differences were considered statistically significant at $p<0.05$.

RESULTS

Twenty-nine questions about oculofacial plastic and orbital surgery were asked in all three artificial intelligence programs. The ChatGPT artificial intelligence program gave correct answers to 13 (44.8%) questions and incorrect answers to 15 (51.7%) questions. ChatGPT replied to one of the questions asked, "As of the last update in September 2021, I could not find a definitive answer to this question." The Bing artificial intelligence program gave correct answers to 14 (48.3%) questions and incorrect answers to 15 (51.7%) questions. On the other hand, the Bard artificial intelligence program gave correct answers to 17 (58.6%) questions and incorrect answers to 12 (41.4%). The number of questions for which all 3 programs gave the same answers was 16 (55.2%). These programs gave correct answers to 9 (56.3%) of the questions and incorrect answers to 7 (43.8%) (Table).

Table. The success of artificial intelligence chatbots on questions related to oculofacial plastic and orbital surgery

Answers (n)	ChatGPT	Bing	Bard
Correct	13 (44.8%)	14 (48.3%)	17 (58.6%)
Incorrect	15 (51.7%)	15 (51.7%)	12 (41.4%)
Same answers (n)	16 (55.2%)		
Correct	9 (56.3%)		
Incorrect	7 (43.8%)		

There was no statistically significant difference between the correct and incorrect answers given by ChatGPT, Bing, and Bard artificial intelligence chatbots ($p=0.609$, Pearson chi-square test). There was no statistical difference between the correct and incorrect response rates of the ChatGPT and Bing chatbots ($p=1.0$, Yates chi-squared test). There was no statistically significant difference between the correct and incorrect response rates of the ChatGPT and Bard chatbots ($p=0.512$, Yates chi-squared test). There was no statistically significant difference between the correct and incorrect response rates of the Bing chatbot and the Bard chatbot ($p=0.599$, Yates chi-square test).

DISCUSSION

ChatGPT, developed based on LLM, is a program that has been processed with 175 billion parameters and aims to produce answers that are similar to the human mindset. Thanks to this complex structure, it has taken a step forward among similar programs.[10] ChatGPT; thanks to its various features such as making a personalized learning plan, performing translations, and helping research, has also found many uses in the field of medicine.[11] In general, ChatGPT is a useful artificial intelligence program for accessing information quickly and reliably. These benefits continue in the medical field. Its use in obtaining information about various diseases, making differential diagnoses, and obtaining information about a wide variety of treatment modalities is an example of its use in the medical field. In addition, being able to examine and analyze medical literature and find summaries of texts are examples of its benefits to medical researcher.[12] Considering all these advantages, it appeals to a wide audience from medical students to a wide range of health professionals, and has various benefits.[13] In addition to ChatGPT, LLM-based Bing and Bard artificial intelligence programs, which were introduced in 2023, also contain similar features. We believe that the advantage of these LLM-based artificial intelligence chatbots in accessing information quickly and reliably can contribute to the learning of diseases and treatment methods related to oculofacial plastic and orbital surgery. In addition, we believe that these programs can help people who are trained and specialized in oculofacial plastic and orbital surgery to save and use time more efficiently. In addition to the various advantages mentioned above, these programs have some disadvantages. Examples of these are the sources referenced for medical information, free online websites on the Internet, limited access to paid articles, and ChatGPT's last update in September 2021.[10,14] Considering all these, it comes to mind that artificial intelligence programs may have problems accessing up-to-date information in constantly renewing medical fields and may raise doubts about the accuracy of the information. Although its wide range of benefits is reassuring for use in the medical field, it is important to test its clinical usefulness and examine its performance.[15,16] We also conducted this study; we designed it to test the performance of artificial intelligence programs on whether they can be used as a resource to access accurate information about oculofacial plastic and orbital surgery under a wide variety of advantages and disadvantages.

In recent years, various studies have answered medical questions and examined their correct and incorrect response rates. Examples include the study in PubMed, which states that the accuracy rate of the models that test the answers to yes or no questions is 68.1%, and a study that examines a dataset of 12,723 questions and states that the accuracy rate is 36.7%.[17,18] With the recent developments in artificial intelligence, the use of artificial intelligence programs such as ChatGPT, Bing, and Bard, which can detect and respond to more complex questions, in answering medical questions has come to the fore. For this purpose, ChatGPT's success in answering the questions in the USMLE has been tested and it has been shown that it can provide more than 50% correct answers.[19] In a study examining the reliability of these artificial intelligence programs in answering ophthalmology-related questions, questions were asked to ChatGPT and Bing artificial intelligence chatbots. It was stated that artificial intelligence programs gave correct answers to the questions at a rate of 58.8% and 71.2%, respectively.[14] We asked 29 questions to the ChatGPT, Bing, and Bard artificial intelligence programs that test the knowledge about oculofacial plastic and orbital surgery, and we found the correct answer rates to be 44.8%, 48.3%, and 58.6%, respectively. ChatGPT's answer

to one question, "As of my last update in September 2021, I could not find a definitive answer to this question." supported our suspicion that this artificial intelligence program may be far from up-to-date information and will have limited effects on accessing the right answer. This rate of all 3 artificial intelligence programs answering the questions correctly may be related to the limited access to current information. Differences in the information tested by the questions and the structure of the questions may have reduced the success of artificial intelligence chatbots. The questions we tested were only related to a specialized field and were taken from a current book. Both the questioning of current information and the existence of a single specialized field may have caused the difference between success rates and other studies. When we evaluated the artificial intelligence chatbots in our study among themselves, although there was no statistically significant difference between the three programs, it was observed that Bing and Bard artificial intelligence programs, which came into use in 2023, had a higher correct answer rate. We think that this situation may be related to the fact that Bing and Bard chatbots can provide access to more up-to-date information.

Limitations

The limitations of the study are that artificial intelligence programs are not specific to medicine, there is no intervention in the operational steps, there is limited access to paid sites, and there is a limitation of ChatGPT in accessing data from 2022 and beyond.

CONCLUSION

As a result, to the best of our knowledge, our study is the first to test the knowledge levels of free artificial intelligence programs released by three different manufacturers based on LLM about oculofacial plastic and orbital surgery and investigate the existence of superiority to each other. Although more recently released artificial intelligence programs, such as Bing and Bard, have higher accuracy in answering questions, the information provided may make a limited contribution to providing users with access to correct information. Care should always be taken regarding the reliability and accuracy of the information provided by artificial intelligence programs.

ETHICAL DECLARATIONS

Ethics Committee Approval

Since our article does not contain human or animal subjects, it does not require an ethics committee approval.

Informed Consent

Since our article does not contain human or animal subjects, it does not require an informed consent.

Referee Evaluation Process

Externally peer-reviewed.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Financial Disclosure

The authors declared that this study has received no financial support.

Author Contributions

All of the authors declare that have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

Data availability statement

All data generated or analyzed during the present study are included in this published article.

REFERENCES

1. Rahimy E. Deep learning applications in ophthalmology. *Curr Opin Ophthalmol*. 2018;29(3):254-260.
2. Patel VL, Shortliffe EH, Stefanelli M, et al. The coming of age of artificial intelligence in medicine. *Artif Intell Med*. 2009;46(1):5-17.
3. Mikolov T, Deoras A, Povey D, Burget L, Černocký J. Strategies for training large scale neural network language models. *2011 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011, Proceedings*. Published online 2011:196-201.
4. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI* Ali MJ. ChatGPT and lacrimal drainage disorders: performance and scope of improvement. *Ophthalmic Plast Reconstr Surg*. 2023;39(3):221-225.
5. Alqahtani H, Kavakli-Thorne M, Kumar G. Applications of generative adversarial networks (GANs): an updated review. *Arch Comput Meth Engineering*. 2021;28(2):525-552.
6. ChatGPT Generative Pre-trained Transformer, Zhavoronkov A. Rapamycin in the context of Pascal's Wager: generative pre-trained transformer perspective. *Oncoscience*. 2022;9:82-84.
7. O'Connor S. Open artificial intelligence platforms in nursing education: tools for academic progress or abuse? *Nurse Educ Pract*. 2023;66:103537.
8. Korn BS, Burkat CN, Carter KD, et al, eds. Oculofacial Plastic and Orbital Surgery. Vol 7. American Academy of Ophthalmology: 2022.
9. 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(3):e1207.
10. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - reshaping medical education and clinical management. *Pak J Med Sci*. 2023;39(2):605.
11. 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*. 2022:2022.12.23.521610.
12. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol*. Published online December 30, 2022. doi: 10.1007/s00330-023-10213-1
13. Cai LZ, Shaheen A, Jin A, et al. Performance of generative large language models on ophthalmology board style questions. *Am J Ophthalmol*. 2023;254:141-149.
14. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of generative pre-trained transformer 3 (GPT-3) in healthcare delivery. *NPJ Digit Med*. 2021;4(1):93.
15. Nath S, Marie A, Ellershaw S, Korot E, Keane PA. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. *Br J Ophthalmol*. 2022;106(7):889-892.
16. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. Pubmedqa: a dataset for biomedical research question answering. *arXiv preprint arXiv: 1909.06146*.
17. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Appl Sci*. 2021;11(14):6421.
18. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health*. 2023;2(2):e0000198.