Original Article

HEALTH SCIENCES
MEDICINE

DOI: 10.32322/jhsm.1419341
J Health Sci Med. 2024;7(2):224-229

ChatGPT in dentomaxillofacial radiology education

Hilal Peker Öztürk[1], Hakan Aysever[1,2], Buğra Şenel[1,2], Şükran Ayran[1], Mustafa Çağrı Peker[3], Hatice Seda Özgedik[1], Nurten Baysal[4]

[1]Department of Dentomaxillofacial Radiology, Gülhane Dentistry Faculty, University of Health Sciences, Ankara, Turkiye
[2]Department of Dentomaxillofacial Radiology, Faculty of Dentistry, University of East Mediterranean, Gazimağusa, North Cyprus
[3]Electrical and Electronics Engineer Business School, University of Sussex, Brighton, England
[4]Department of Prosthodontics, Gülhane Dentistry Faculty, University of Health Sciences, Ankara, Turkiye

Cite this article as: Öztürk Peker H, Aysever H, Şenel B, et al. ChatGPT in dentomaxillofacial radiology education. *J Health Sci Med.* 2024;7(2):224-229.

Received: 14.01.2024 ◆ **Accepted**: 18.03.2024 ◆ **Published**: 25.03.2024

ABSTRACT

Aims: Artificial intelligence refers to the ability of computer systems or machines to perform cognitive functions and tasks that are similar to humans'. The aim of this study is to assess the knowledge and interpretative abilities of ChatGPT-versions by administering a dentomaxillofacial-radiology exam, comparing its performance with that of dentistry-students in Türkiye, and questioning the effectiveness of different languages.

Methods: It is a descriptive research comparing the data of ChatGPT versions 3.5 and 4 in both Turkish and English.

Results: Firstly 20 test-questions were evaluated. There is a significant difference($p<0.05$) between the ChatGPT answer-sheets. ChatGPT-4 in English demonstrated the highest performance. Answer-sheets of chatGPT-4 in Turkish and English demonstrated the best performance with 5 correct answers in open-ended-questions. Based on the answers of 89 students to 20 test-questions, a class-profile was created. ChatGPT answer-sheets and class-profile were analyzed($p<0.05$). Class-profile ranked first as ChatGPT-4 in English. A significant difference was found between the answer-sheets of ChatGPT and the class-profile for open-ended-questions($p<0.10$). The most successful results were obtained from ChatGPT-4 in Turkish and English, as well as the class-profile.

Conclusion: It is important to mention that ChatGPT 3.5's knowledge and perception in the field of dentomaxillofacial radiology are not sufficient, particularly for use in examinations.

Keywords: Artificial intelligence, dentistry, education, oral radiology

INTRODUCTION

Artificial intelligence (AI) refers to the ability of computer systems or machines to perform cognitive functions and tasks that are similar to those of humans. These functions include image or language recognition, learning, decision-making, problem-solving, and other processes performed by humans. AI systems utilize algorithms to execute tasks that were previously exclusive to human intelligence. As AI technology continues to evolve, its applications are becoming increasingly prevalent across various fields.[1-4]

A Large Language Model (LLM) is a type of machine learning model that employs deep-learning algorithms. They are also referred to as auto-regressive language models, and ChatGPT (Chat Generative Pre-trained Transformer) is the latest iteration within this category. Developed by OpenAI in San Francisco, California, ChatGPT utilizes the GPT architecture. It is engineered to generate responses resembling those of humans when prompted with questions or prompts, employing sophisticated algorithms to comprehend and process natural language. The conversational abilities of ChatGPT stem directly from its AI capabilities, positioning it as a chatbot facilitating advanced and natural interactions between humans and machines. ChatGPT finds application in various domains, including customer service, chatbots, text-based dialogue systems, automatic text completion tools, among others.[5-8]

Shortly after the announcement of the first version of the model, GPT-1 in 2018, GPT-2 was introduced in 2019 with a larger dataset and increased analytical power. With the release of GPT-3 in 2020, a series of models capable of understanding and generating natural language was established, making it the largest language model to date. The current default version, GPT-3.5, has incorporated code understanding and generation capabilities in addition to its superior language comprehension abilities. As of March 14, 2023, the limited beta version of GPT-4 has been released. Currently, GPT-3.5 represents an advanced version capable of handling both text and code, with potential future enhancements in image analysis capabilities.[6,8]

Over the past decade, the clinical application of AI programs in the medical profession, including dentistry, has gained significant traction. AI finds various applications in dentistry, not only in clinical practice but also in dental education and patient management. Dental education is evolving beyond traditional clinical simulator studies to include opportunities for students to reinforce or challenge theoretical knowledge. ChatGPT, as an AI tool, offers numerous benefits for dental students in their education and learning processes. Demonstrating ChatGPT's capabilities can provide crucial information for dental students regarding their education and establish the reliability of ChatGPT for learning purposes. It is essential for dental students to be able to evaluate the accuracy of medical and dental information generated by AI and to generate reliable, verified information for patients. Therefore, it is imperative to evaluate ChatGPT's ability to provide accurate answers to dental examination questions.[1,9-12]

The objective of this study is to evaluate the knowledge and interpretative capabilities of various versions of ChatGPT by administering a dentomaxillofacial radiology examination. This assessment will involve comparing the performance of ChatGPT with that of dentistry students in Turkiye and exploring the effectiveness of different language options.

The null hypothesis of this study posits that there will be no significant difference in the exam results between ChatGPT and students who have taken the dentomaxillofacial radiology course.

METHODS

This study is a descriptive research endeavor that involves comparing data obtained from dentomaxillofacial radiology exam questions routinely administered to 4th-year dental students with responses generated by ChatGPT versions 3.5 and 4, in both Turkish and English languages. Since there were no human or animal experiments conducted, no informed consent or ethical approval was required. In April 2023, a dentomaxillofacial radiology exam was conducted for 4th-year students at the University of Health Sciences, Gülhane Faculty of Dentistry. The exam questions were presented to two distinct versions of ChatGPT, with each version supporting both Turkish and English languages. The questions were administered twice in each version and language, resulting in a total of four ChatGPT answer sheets being generated.

All exam papers were evaluated by the same assessors, Associate Professor BŞ and Assistant Professor HPÖ. Answers obtained in Turkish were manually transcribed onto copies of the official exam paper and shuffled among all exam papers to ensure objectivity in the assessment conducted by the evaluators. The Turkish and English versions of the same questions were administered twice to both ChatGPT versions 3.5 and 4, and the average of the obtained results was used for the final assessment. The study involved comparing answer sheets generated by ChatGPT with the performance of dentistry students. This comparative analysis aimed to evaluate ChatGPT's performance relative to that of dentistry students and to assess variations and consistencies within the ChatGPT-generated answers. The examination comprised 28 questions, including 20 multiple-choice questions and 8 essay and short-answer questions. The 8 essay or short-answer questions were evaluated on a scale of 5 points. According to the evaluator's assessment, a score of 2.5 or higher for each question was considered correct, while a score below 2.5 was deemed incorrect in both the evaluation of students' and ChatGPT's answer sheets.

A total of 8 answer sheets generated by ChatGPT were evaluated in terms of both multiple-choice and open-ended question types. The objective was to compare the performance success in both types of questions and examine how well ChatGPT performed in each relative to the other. The topics covered in this exam were taught to students over a period of 12 weeks, comprising 24 lecture hours between November 2022 and April 2023. The total number of participating students was 89, and there were no exclusion criteria.

The overall performance of the 89 students and ChatGPT were evaluated. All statistical analyses were conducted using IBM SPSS Statistics 21.0 (IBM Corp., Armonk, NY, USA) and MS Excel 2007. Descriptive statistics, including the number, mean, and standard deviation, were calculated for the variables.

RESULTS

The chi-square test was used to evaluate the 20 test questions. With a p-value of 0.019 ($p<0.05$), there is a significant difference between the ChatGPT answer sheets. The exam performances were further analyzed to determine the nature of this difference (Table 1).

ChatGPT-4 in English demonstrated the highest performance by correctly answering 15 out of 20 multiple-choice questions. Similarly, the answer sheet of ChatGPT-4 in Turkish and ChatGPT-3.5 in English

Table 1. ChatGPT answer sheets' correct and incorrect answers to the 20 multiple-choice questions

Group	Incorrect answers	Correct answers	Total questions	Success ranking
ChatGPT 3.5 Turkish	10	10	20	3
ChatGPT 3.5 English	**6**	**14**	**20**	**2**
ChatGPT 4 Turkish	**6**	**14**	**20**	**2**
ChatGPT 4 English	**5**	**15**	**20**	**1**

each had 14 questions correctly answered. However, ChatGPT-3.5 in Turkish showed poorer performance, with only 10 correct answers.

There is a weak significant difference between the ChatGPT answer sheets in terms of open-ended questions, with a p-value of 0.073 ($p<0.10$). The answer sheets of ChatGPT-4 in both Turkish and English demonstrated the best performance with 5 correct answers each. Conversely, the poorest performance in open-ended questions was observed in ChatGPT-3.5 in Turkish, with only 2 correct answers (Table 2).

Table 2. ChatGPT answer sheets' correct and incorrect answers to the 8 open-ended questions

Group	Incorrect answers	Correct answers	Total questions	Success ranking
ChatGPT 3.5 Turkish	6	2	8	3
ChatGPT 3.5 English	4	4	8	2
ChatGPT 4 Turkish	3	5	8	1
ChatGPT 4 English	3	5	8	1

Additionally, question-based success charts for ChatGPT versions' answer sheets have been generated for both multiple-choice and open-ended questions (Figure 1 and Figure 2).

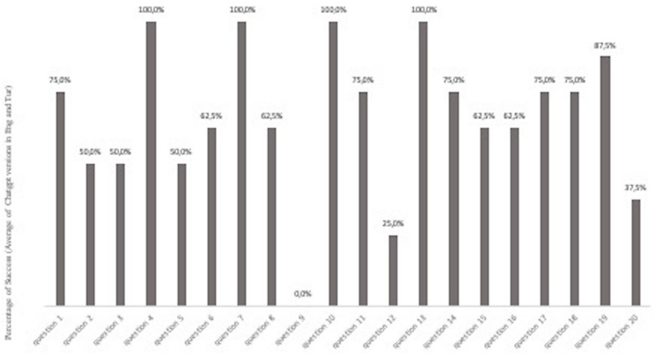

Figure 1. The correct response rates of ChatGPT versions' answer-sheets for 20 multiple-choice test questions

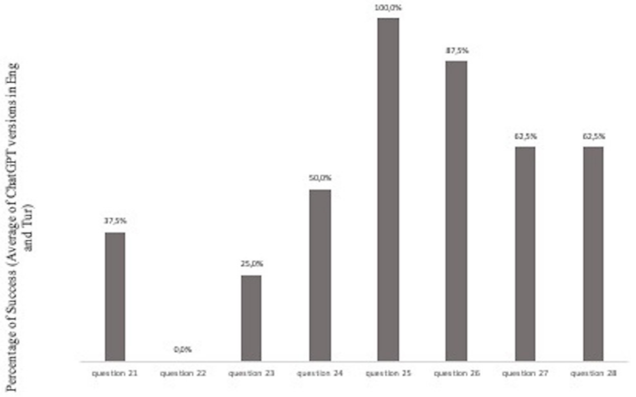

Figure 2. The correct response rates of ChatGPT versions' answer-sheets for 8 open-ended questions

Based on the answers obtained from the 89 students to the 20 test questions, a class profile was created (Table 3). The ChatGPT answer sheets and the answers of the students (class profile) for the 20 test questions were analyzed using the Kruskal-Wallis test. The p-value of 0.018 ($p<0.05$) indicates a significant difference. The class profile ranked first with 15 correct answers, matching the performance of ChatGPT-4 in English.

Table 3. The total scores of class-profile, ChatGPT 3.5 and 4

Answer-Sheet	Correct answers to the test-questions N/S	Correct answers to open-ended questions N/S	Totally Score S
Class-Profile(Students)	15/45	5/25	70
Chatgpt 3.5/ Turkish	10/30	2/10	40
Chatgpt 3.5/ English	14/42	4/20	62
Chatgpt 4/ Turkish	14/42	5/25	67
Chatgpt 4/ English	15/45	5/25	70

Another class profile was created for the open-ended questions of the 89 students. The class profile's question-based success chart has been developed for both multiple-choice and open-ended questions (Figure 3 and Figure 4). The Kruskal-Wallis test found a significant difference between the answer sheets of ChatGPT models and the class profile's answers for open-ended questions ($p=0.076$, $p<0.10$), indicating a weak significant difference.

Students achieved similar performance with some answer sheets of ChatGPT. The most successful results in terms of open-ended questions were obtained from ChatGPT-4 in both Turkish and English, as well as the class profile, with each correctly answering 5 out of the 8 open-ended questions.

When the answers were evaluated in total in the present study, the class profile and the English version of ChatGPT-4 achieved the highest performance, obtaining

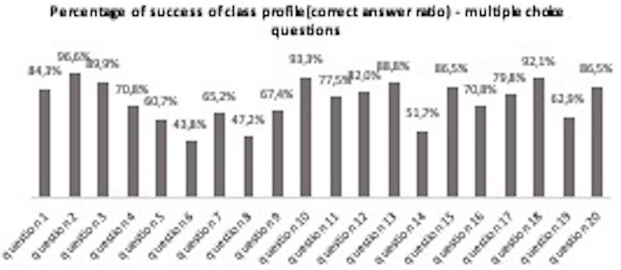

Figure 3. The correct response rates of students(class profile) for 20 multiple-choice test questions

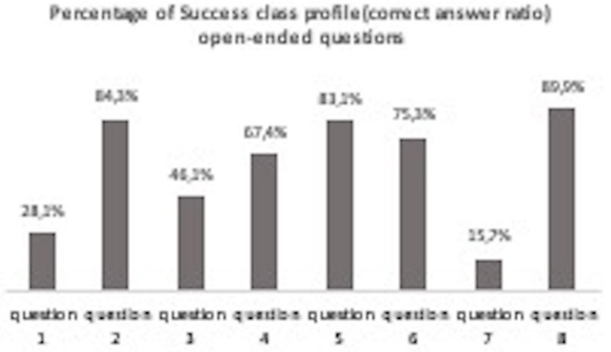

Figure 4. The correct response rates of students(class profile) for 8 open-ended questions

70 out of 100. Following closely, the Turkish version of ChatGPT-4 obtained 67 out of 100, demonstrating the second-highest performance. On the other hand, the lowest performance was observed in ChatGPT-3.5 in Turkish, obtaining 40 out of 100. The correct and incorrect answers given to both multiple-choice and open-ended questions, along with the total scores obtained, are provided in Table 3.

The performance of ChatGPT-3.5 and ChatGPT-4 in the English language remained consistent, exhibiting unchanged performance across both initial and subsequent evaluations. In open-ended questions, ChatGPT-4 demonstrated consistent results between its initial and subsequent evaluations in English, while ChatGPT-3.5 exhibited improved performance in the second evaluation compared to the first. In multiple-choice questions, the initial evaluation of ChatGPT-3.5 in Turkish displayed superior performance compared to the subsequent evaluation. However, both evaluations yielded similar results in open-ended questions. Conversely, in version-4, the second evaluation conducted in Turkish yielded improved results in both multiple-choice and open-ended questions compared to the initial evaluation. Consequently, these findings suggest that ChatGPT exhibited greater consistency in providing results in English during the Dentomaxillofacial Radiology exam.

DISCUSSION

The null hypothesis of this study, which posited that ChatGPT's exam results would outperform those of students, was partially rejected. In multiple-choice questions, the English version of ChatGPT-4 exhibited the highest performance, yet it was comparable to the performance of the student cohort. Conversely, in open-ended questions, both the English and Turkish versions of ChatGPT-4, as well as the student cohort, produced similar results.

As of January 2023, ChatGPT has amassed a user base of 100 million, solidifying its position as the AI program with the highest user ratio.[13] For researchers and practitioners to effectively utilize ChatGPT and mitigate potential issues, it is imperative to gain a comprehensive understanding of its capabilities and constraints.[7]

However, there exists a scarcity of studies that have explored ChatGPT within the realms of medical and dental fields thus far.[14]

In a similar study, medical school students' responses were compared to those generated by ChatGPT in an examination centered on medical parasitology. The performance of medical school students was found to surpass that of ChatGPT. In the current study, success was evaluated based on two question types: open-ended questions and test questions. Furthermore, two distinct versions of ChatGPT, in English and Turkish, underwent the examination twice each. Unlike Huh's study, ChatGPT-4 in English achieved the highest academic performance alongside the class-profile, both in terms of the overall exam score and the accuracy of responses to the test questions.[1]

In open-ended questions, both students and ChatGPT-4 in Turkish and English achieved comparable levels of success. However, ChatGPT-3.5 yielded varying results for Turkish and English. This discrepancy may stem from the complexity of these models, the training data utilized, and the inherent characteristics of languages. Consequently, it is believed that certain language models may generate different responses across different languages, and achieving perfect consistency in responses may not always be feasible.

The study has unveiled another noteworthy finding: the consistency between the results of the first and second evaluations of exams conducted in English by ChatGPT surpasses that of those conducted in the Turkish language. In fact, this consistency has reached 100% in version-4.

According to the evaluated versions of ChatGPT, the highest success rate in multiple-choice questions is 75%, whereas in open-ended questions, it decreases to 62.5%.

In this study, within the field of dentomaxillofacial radiology, ChatGPT has exhibited relatively superior performance in multiple-choice questions compared to open-ended questions. Nevertheless, the results from both sets of questions suggest that there are certain areas where the knowledge of ChatGPT appears to be deficient. These findings may not be universally applicable to all fields within dentistry but are specific to the domain of dentomaxillofacial radiology. However, it can be concluded that the knowledge base of ChatGPT may not be sufficiently robust for advanced levels within this particular field.

In their study, Khurana and Vaddi[15] have identified the potential applications of ChatGPT within the realm of dentomaxillofacial radiology. They highlighted several areas where ChatGPT can be effectively utilized, including; generating oral radiology reports, responding to multiple-choice questions, aiding in scientific writing, contributing to dental education by assisting in presentation creation, providing feedback on student assignments, aiding in the preparation of academic content outlines. However, they noted that ChatGPT exhibits limitations in addressing queries involving image-based questions.

The performance of ChatGPT has been assessed in a study concerning the United States Medical Licensing Examination(USMLE) Step 1, 2, and 3. The examination comprised a combination of open-ended questions with variable inputs and multiple-choice questions, specifically single-answer questions. As observed in the present study, ChatGPT exhibited its lowest performance in the Step 1 assessment, which primarily involved open-ended questions. This was followed by its performance in Step 2 and 3. Remarkably, the outcomes of the study align with those achieved by human subjects.[16] According to the results obtained from this study, it has been confirmed that ChatGPT-4 has been enhanced and improved compared to version-3.5.

In another study, Ali and his colleagues[17] utilized ChatGPT to generate clinical letters for patients. The results revealed that the correctness rates of the letters created using ChatGPT were significantly higher compared to those generated by physicians. One of the crucial aspects of studies lies in the act of writing. This is because a study can only be effectively introduced to the literature and the realm of academia through well-crafted composition. However, it is imperative to recognize that ChatGPT should not serve as the sole source. It is advisable to seek support from ChatGPT for language enhancement and rectification of certain errors.

The most significant advantage of ChatGPT can be evaluated as its ability to rapidly comprehend given information and arrive at evidence-based conclusions more swiftly than humans. Essentially, individuals can pose questions on any topic and promptly receive satisfactory answers through ChatGPT.[14,18]

CONCLUSION

Overall, in multiple-choice questions, ChatGPT-4 in English has exhibited the highest performance, mirroring that of the class profile. However, in open-ended questions, ChatGPT-4 in both English and Turkish, as well as the class profile, have demonstrated equivalent results as the highest performance. The divergent outcomes observed in ChatGPT-3.5 evaluations in Turkish and English underscore the necessity for further research into evaluation across different languages. Therefore, new studies need to be conducted in various languages and with diverse content.

As a precautionary note to dental students, it is important to mention that ChatGPT's knowledge and understanding in the field of Dentomaxillofacial Radiology are not yet sufficient, particularly for use in examinations. Nonetheless, it is rapidly evolving and improving.

ETHICAL DECLARATIONS

Ethics Committee Approval

Since there were no human or animal experiments conducted, no ethical approval required.

Informed Consent

Since there were no human or animal experiments conducted, no informed consent required.

Referee Evaluation Process

Externally peer-reviewed.

Conflict of Interest Statement

The authors have no conflict of interests to declare.

Financial Disclosure

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

Author Contributions

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

REFERENCES

1. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J *Educ Eval Health Prof.* 2023;20:1.
2. Keser G, Pekiner FN. Attitudes, perceptions and knowledge regarding the future of artificial intelligence in oral radiology among a group of dental students in Turkey: a survey. C*lin Exp Health Sci.* 2021;11(4):637-641.

3. Yüzbaşıoğlu E. Attitudes and perceptions of dental students towards artificial intelligence. *J Dent Educ*. 2021;85(1):60-68.
4. Sur J, Bose S, Khan F, Dewangan D, Sawriya E, Roul A. Knowledge, attitudes, and perceptions regarding the future of artificial intelligence in oral radiology in India: a survey. *Imaging Sci Dent*. 2020;50(3):193-198.
5. Thurzo A, Strunga M, Urban R, Surovková J, Afrashtehfar KI. Impact of artificial intelligence on dental education: a review and guide for curriculum update. *Educ Sci*. 2023;13(2):150.
6. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the united states medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9(1):e45312.
7. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst*. 2023;47(1):33.
8. Ollivier M, Pareek A, Dahmen J, et al. A deeper dive into ChatGPT: history, use and future perspectives for orthopaedic research. *Knee Surg Sports Traumatol Arthrosc*. 2023;31(4):1190-1192.
9. Ding H, Wu J, Zhao W, Matinlinna JP, Burrow MF, Tsoi JK. Artificial intelligence in dentistry-a review. *Front Dent Med*. 2023;4:1085251. doi: 10.3389/fdmed.2023.1085251
10. Agrawal P, Nikhade P. Artificial intelligence in dentistry: past, present, and future. *Cureus*. 2022;14(7):e27405.
11. Nguyen TT, Larrivee N, Lee A, Bilaniuk O, Durand R. Use of artificial intelligence in dentistry: current clinical trends and research advances. *J Can Dent Assoc.* 2021;87(l7):1488-2159.
12. Park SH, Do KH, Kim S, Park JH, Lim YS. What should medical students know about artificial intelligence in medicine? *J Educ Eval Health Prof*. 2019;16:18.
13. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Med Educ*. 2023;9(1):e46885. doi: 10.2196/46885
14. Fatani B. ChatGPT for future medical and dental research. *Cureus*. 2023;15(4):e37285. doi: 10.7759/cureus.37285
15. Khurana S, Vaddi A. ChatGPT from the perspective of an academic oral and maxillofacial radiologist. *Cureus*. 2023;15(6):e40053. doi: 10.7759/cureus.40053
16. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*. 2023;15(2):e35179. doi: 10.7759/cureus.35179
17. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. *Lancet Digit Health.* 2023;5(4):e179-e181. doi: 10.1016/S2589-7500(23)00048-1
18. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? *Crit Care*. 2023;27(1):75. doi: 10.1186/s13054-023-04380-2