DOI: 10.37990/medr.1567242 Med Records 2025;7(1):162-6

MEDICAL RECORDS-International Medical Journal

Research Article

Assessing the Performance of ChatGPT on Dentistry Specialization Exam Questions: A Comparative Study with DUS Examinees

Mustafa Temiz, Ceylan Guzel

İstanbul Medipol University, Faculty of Dentistry, Department of Oral and Maxillofacial Surgery, İstanbul, Türkiye

Abstract

Aim: This study aims to evaluate the performance of the ChatGPT-4.0 model in answering questions from the Turkish Dentistry Specialization Exam (DUS), comparing it with the performance of DUS examinees and exploring the model's clinical reasoning capabilities and its potential educational value in dental training. The objective is to identify the strengths and limitations of ChatGPT when tasked with responding to questions typically presented in this critical examination for dental professionals.
Material and Method: The study analyzed DUS questions from the years 2012 to 2017, focusing on the basic medical sciences and clinical sciences sections. ChatGPT's responses to these questions were compared with the average scores of DUS examinees, who had previously taken the exam. A statistical analysis was performed to assess the significance of the differences in performance between ChatGPT and the human examinees.
Results: ChatGPT significantly outperformed DUS examinees in both the basic medical sciences and clinical sciences sections across all years analyzed. The statistical analysis revealed that the differences in performance between ChatGPT and DUS examinees were statistically significant, with ChatGPT demonstrating superior accuracy in all years.
Conclusion: ChatGPT's performance on the DUS demonstrates its potential as a supplementary tool for dental education and exam preparation. However, future research should focus on integrating AI into practical dental training, particularly in assessing its real-world applicability. The limitations of AI in replicating hands-on clinical decision-making in unpredictable environments must also be considered.

Keywords: Artificial intelligence in dentistry, clinical decision support, chatgpt in medical education, dental exam performance

INTRODUCTION

Specialization in dentistry requires extensive and deep knowledge in both basic medical sciences and advanced clinical practices. In Türkiye, the Dentistry Specialization Examination (DUS) is one of the most critical exam for dental professionals who wish to pursue specialist training. Administered by the Evaluation Selection and Placement Center (ÖSYM), the DUS is a rigorous examination that consists of 120 questions divided into basic medical sciences and clinical sciences sections. These sections test the knowledge and clinical competence of examinees (1). Success in this examination is crucial for dental professionals looking to advance in their careers, as it determines eligibility for specialist education in Türkiye's competitive healthcare environment.

In recent years, advancements in artificial intelligence (AI) have sparked considerable interest in its potential applications in medical and dental education (2). One of the most notable innovations in AI is the development of large language models (LLMs) like ChatGPT by OpenAI. These models have been shown to perform comparably to human examinees in various standardized tests, including medical licensure exams such as the United States Medical Licensing Examination (USMLE) (3-5).

LLMs like ChatGPT have also been used for various educational purposes, such as assisting with scientific writing, conducting literature reviews, and formulating research questions (6). The potential of these models to complement traditional educational methods by providing personalized learning experiences is becoming increasingly

CITATION
Temiz M, Guzel C. Assessing the Performance of ChatGPT on Dentistry Specialization Exam Questions: A Comparative Study with DUS Examinees. Med Records. 2025;7(1):162-6. DOI:1037990/medr.1567242

Received: 15.10.2024 **Accepted:** 19.12.2024 **Published:** 14.01.2025
Corresponding Author: Mustafa Temiz, İstanbul Medipol University, Faculty of Dentistry, Department of Oral and Maxillofacial Surgery, İstanbul, Türkiye
E-mail: drmustafatemiz@gmail.com

apparent (7). In particular, studies have suggested that AI models can offer a significant advantage in preparing for critical exams by providing real-time feedback and simulating exam conditions (6,8).

However, while AI has proven to be effective in standardized testing environments, there is still much to learn about its potential to support clinical decision-making and practical skill development in fields such as dentistry (5,6). This study systematically evaluates the performance of ChatGPT in answering questions from the Turkish DUS, focusing on its clinical reasoning capabilities and reliability as an educational tool. The study compares ChatGPT's performance to that of actual DUS examinees who took the exam between 2012 and 2017, with the goal of assessing the strengths and limitations of AI in this specific context. Additionally, the research explores the potential role of AI in enhancing dental education, especially in preparing for specialized exams like the DUS.

MATERIAL AND METHOD

Study Design

This study was designed as a retrospective analysis aimed at evaluating the performance of the ChatGPT-4.0 model in answering multiple-choice questions from the Turkish DUS. The study compares ChatGPT's performance to that of DUS examinees who took the exam between 2012 and 2017. The analysis focused on net correct answers and considered the basic medical sciences and clinical sciences sections separately, as these sections encompass different areas of knowledge and testing formats.

Data Collection

Data was gathered from six DUS exams administered between 2012 and 2017. The exams were made publicly available by ÖSYM, and only those that provided full performance data and allowed open access were included in the study (https://www.osym.gov.tr/TR,25704/2023.html). Any exams with missing numerical data or restricted access were excluded from the analysis. The study utilized 120 multiple-choice questions from each DUS exam, covering both basic medical sciences and clinical sciences. Each exam consisted of 40 questions from the basic medical sciences and 80 questions from the clinical sciences, providing a comprehensive evaluation of the dental knowledge required for specialization.

Performance data for the DUS examinees were obtained from official ÖSYM reports. The net scores for each examinee were calculated using the standard scoring method, where one point was subtracted for every four incorrect answers from the total number of correct answers. This calculation method was applied consistently across all years to ensure comparability of the data.

AI Model

The ChatGPT-4.0 model, developed by OpenAI, was utilized to answer the DUS questions. Each question was presented to the model via screen recordings, with the model selecting one of the multiple-choice answers for each question. The model's performance was evaluated based on the accuracy of its selected answers compared to the correct responses. The analysis aimed to identify how well ChatGPT could perform on specialized dental exams and to compare its accuracy with that of human examinees.

RESULTS

The performance data comparing ChatGPT and DUS examinees across the years 2012 to 2017 is summarized. (Table 1).

Table 1. Descriptive statistics of performance (net correct answers)

Year	DUS examinees medical basic science	ChatGPT medical basic science	DUS examinees clinical science	ChatGPT clinical science
2012	15.42	40.00	41.08	67.50
2013	16.23	35.50	48.90	60.25
2014	19.88	40.00	43.52	62.50
2015	16.86	40.00	49.56	65.50
2016	14.04	36.25	50.46	66.25
2017	19.55	40.00	46.72	65.00

Figure 1 highlights that ChatGPT consistently outperformed the DUS examinees in both basic medical sciences and clinical sciences sections across all years. The most significant differences were observed in the basic medical sciences, where ChatGPT achieved near-perfect scores in several years, while the DUS examinees' scores were substantially lower. Even in the clinical sciences, which tend to require more complex reasoning and application of knowledge, ChatGPT consistently outscored the human examinees.

Figure 1 presents the success rates of ChatGPT and DUS examinees as a percentage of the total number of questions answered correctly in both the basic medical sciences and clinical sciences sections. As shown in the figure, ChatGPT's success rates were consistently high across all years and far exceeded those of the DUS examinees in both sections. ChatGPT's success rates in the basic medical sciences were particularly impressive, often exceeding 90%, while the DUS examinees' success rates were generally below 50%.

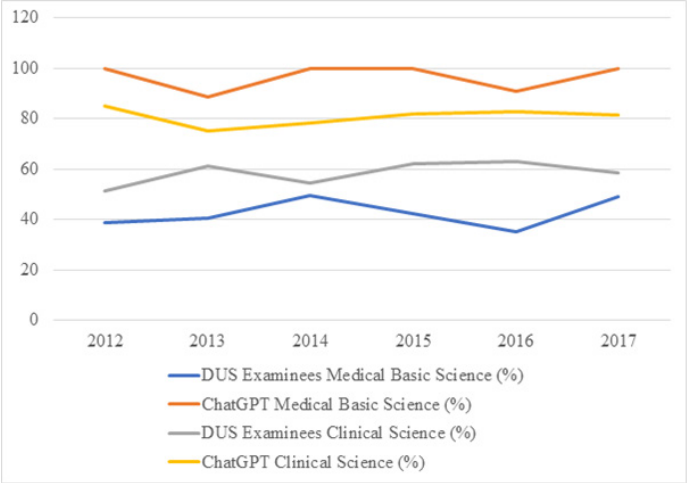

Figure 1. Success rates: DUS examinees vs. ChatGPT (medical basic science and clinical science)

A paired t-test was conducted to compare the average performance of ChatGPT and DUS examinees in both the basic medical sciences and clinical sciences sections. The results of the t-test indicated a statistically significant difference ($p<0.05$) in both sections. ChatGPT's mean score in the basic medical sciences was 38.63, compared to the DUS examinees' mean score of 16.66. In the clinical sciences, ChatGPT's mean score was 64.83, compared to the DUS examinees' mean score of 46.87 (Figure 2). These results demonstrate a substantial performance gap in favor of ChatGPT.

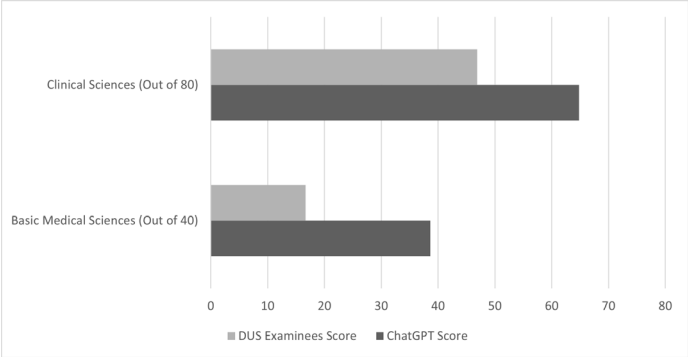

Figure 2. Comparison of mean net scores (absolute out of 40 for basic, 80 for clinical): ChatGPT vs DUS examinees

DISCUSSION

The results of this study indicate that ChatGPT consistently outperformed DUS examinees in both the basic medical sciences and clinical sciences sections of the exam. These findings highlight the growing potential of AI in medical and dental education, specifically in specialized exams like the DUS. Large language models such as ChatGPT are highly capable of processing vast amounts of information quickly and accurately, making them particularly suitable for answering questions that rely on recall. This is consistent with the findings of Brown et al., who stated that language models are effective in tasks that require memory-based knowledge (2).

One explanation for ChatGPT's superior performance may be that the exams are structured as multiple choice; these exams rely heavily on recall rather than critical thinking or clinical judgment. ChatGPT excels at providing accurate answers because it can access a vast repository of information and process it with incredible speed. This is consistent with research by Cascella et al., which shows that AI models such as ChatGPT are particularly suitable for tasks that require retrieving specified facts from a knowledge database (5).

In contrast, although test examinees may be knowledgeable, they may be affected by factors such as cognitive limitations, test anxiety, or variability in test preparation methods, which may lead to lower performance according to the AI model (6). Additionally, humans face the natural limitations of information processing speed and working memory, which can lead to errors, especially under time constraints. The variability in performance among DUS examinees over the years may also be attributed to changes in the difficulty of the exam questions, external pressures, and differing levels of preparedness among candidates (9).

ChatGPT's strong performance in the clinical sciences division was particularly notable. Clinical exams that require decision-making by synthesizing knowledge are often considered a challenging aspect of medical and dental examinations (10,11). Although ChatGPT's success in such exams is impressive, the fact that the exam consists of multiple-choice questions plays an important role in this success. However, as Davenport et all point out, AI systems may struggle in real-world clinical scenarios where decisions rely on the incorporation of uncertain or incomplete information (12). This suggests that ChatGPT's success in exams does not necessarily translate into practical clinical decision-making.

In the basic medical sciences section, ChatGPT achieved a success rate of 97.39%, while the average success rate for DUS examinees was only 42.49%. This significant difference in performance may be attributed to the nature of the questions in the basic sciences. Subjects like anatomy, histology, and pharmacology rely on well-documented and relatively stable knowledge bases, which are readily available in public databases. ChatGPT, being trained on vast amounts of such data, can quickly retrieve and accurately process this information. Moreover, the questions in these areas often have definitive answers (13-15). This is likely why ChatGPT achieved near-perfect scores in these sections, compared to the success rates of the DUS examinees.

However, it is important to acknowledge that clinical knowledge differs significantly from basic science knowledge, as it requires the practical application of skills acquired through experience (16). Dentists and other healthcare professionals, when encountering clinical cases, must go beyond abstract information and make patient-specific decisions. The complexity of managing diverse patient scenarios, especially in high-pressure clinical environments, requires a type of reasoning that goes beyond what AI models can currently replicate. Clinical

decisions often involve interpreting subtle patient cues, integrating hands-on experience, and considering patient preferences, all of which are difficult for AI to simulate accurately (16-18). In this study, ChatGPT demonstrated a 80.21% success rate in clinical sciences, compared to 58.36% for DUS examinees. Although ChatGPT's performance was still superior, the smaller margin of success suggests that clinical questions particularly those involving diagnostic reasoning or patient management may pose greater challenges for AI models.

Moreover, while ChatGPT demonstrated exceptional proficiency in the theoretical aspects of clinical sciences, it is essential to recognize that true clinical competence involves more than just answering questions correctly (19). Effective clinical decision-making requires the ability to weigh multiple factors simultaneously, to exercise judgment in the face of uncertainty, and to engage in hands-on procedures that require fine motor skills and the ability to adapt to real-time feedback (19-21). As such, while ChatGPT has proven to be a valuable tool for knowledge acquisition and standardized testing, its utility in real-world clinical practice remains limited by its inability to replicate these higher-order cognitive processes.

Despite these limitations, AI has significant potential as an educational tool. By integrating artificial intelligence into dental education, students can benefit from personalized learning experiences tailored to their individual strengths and weaknesses. AI-powered platforms can provide targeted learning materials and practice questions, helping students prepare more effectively for exams such as the DUS (22,23). In particular, AI systems can be used to identify areas where students struggle the most, enabling educators to offer more focused instruction in those areas. Furthermore, AI-driven assessments can give students an opportunity to test their knowledge in a simulated environment, providing feedback that can help them build confidence and improve their performance on future exams.

Additionally, AI can assist in enhancing the learning experience by providing detailed explanations for incorrect answers. This type of real-time feedback helps students understand their mistakes and develop better clinical reasoning skills over time. However, educators must be mindful of the risks associated with over-reliance on AI-generated answers. While AI can provide support in factual recall, it is crucial for students to cultivate their own problem-solving abilities and develop critical thinking skills, particularly in the context of clinical decision-making. As Davenport and Kalakota caution, AI should be viewed as a tool to complement, rather than replace, traditional learning methods (12).

Limitations of the Study

Despite the promising findings of this study, several limitations must be considered. First, the data used for this study were based on publicly available DUS exam results, and the performance of the DUS examinees may not be fully representative of the general population of dental students. Additionally, the ChatGPT model was developed primarily in English, which could have affected its performance when handling Turkish-language exam questions. This language discrepancy is a factor that must be considered when evaluating the model's accuracy and reliability in non-English exams. Future studies should explore the impact of language differences on AI performance, particularly in multilingual or non-English contexts.

Furthermore, while ChatGPT performed well in this study, it is important to remember that the model was evaluated in a controlled, multiple-choice exam environment. Real-world clinical practice is far more complex and dynamic, involving patient interactions, physical examinations, and hands-on procedures that cannot be easily replicated by AI (18,21,24). Therefore, future research should explore the use of AI in clinical practice settings to determine whether it can assist dental professionals in making accurate decisions when treating patients.

CONCLUSION

This study highlights the potential of AI, particularly large language models like ChatGPT, to support dental education by providing accurate knowledge recall and assisting in exam preparation. ChatGPT's superior performance on the DUS, particularly in the basic medical sciences and clinical sciences sections, demonstrates that AI can be a valuable tool for dental professionals preparing for critical exams. However, while AI shows promise in structured, fact-based testing environments, its limitations in real-world clinical practice, where situational judgment and hands-on skills are critical, must be acknowledged.

The future of dental education will likely involve integrating AI as a supplementary tool, enhancing students' ability to retain and recall knowledge while emphasizing the irreplaceable value of human clinical expertise. AI should be used to complement traditional learning methods, helping students build a solid foundation of knowledge that can be applied in practical, real-world scenarios. By embracing the potential of AI while recognizing its limitations, educators can help prepare the next generation of dental professionals for success in both academic and clinical settings.

Financial disclosures: *The authors declared that this study has received no financial support.*

Conflict of interest: *The authors have no conflicts of interest to declare.*

Ethical approval: *Ethical approval has been obtained from the Ethics Committee of İstanbul Medipol University (ethics committee approval number: 1135, issue number: E-10840098-202.3.02-7444).*

REFERENCES

1. T.C. Cumhurbaşkanlığı Mevzuat Bilgi Sistemi. Tıpta ve diş hekimliğinde uzmanlık eğitimi yönetmeliği. www.mevzuat.gov.tr/mevzuat?MevzuatNo=39700&MevzuatTur=7&MevzuatTertip=5 acces date 10.10.2024.

2. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. ArXiv. 2020 doi: 10.48550/arXiv.2005.14165

3. Patino GA, Amiel JM, Brown M, et al. The promise and perils of artificial intelligence in health professions education practice and scholarship. Acad Med. 2024;99:477-81.

4. 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:e0000198.

5. 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;47:33.

6. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. ArXiv. May 2017. doi: 10.48550/arXiv.1706.03762

7. Nagi F, Salih R, Alzubaidi M. et al. Applications of artificial intelligence (AI) in medical education: a scoping review. Stud Health Technol Inform. 2023;305:648-51.

8. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med. 2018;15:e1002686.

9. Cowan N. Working memory underpins cognitive development, learning, and education. Educ Psychol Rev. 2014;26:197-223.

10. Sumbal A, Sumbal R, Amir A. Can ChatGPT-3.5 pass a medical exam? a systematic review of ChatGPT's performance in academic testing. J Med Educ Curric Dev. 2024;11:23821205241238641.

11. Yu P, Fang C, Liu X, et al. Performance of ChatGPT on the Chinese postgraduate examination for clinical medicine: survey study. JMIR Med Educ. 2024;10:e48514.

12. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6:94-8.

13. Choi W. Assessment of the capacity of ChatGPT as a self-learning tool in medical pharmacology: a study using MCQs. BMC Med Educ. 2023;23:864.

14. Totlis T, Natsis K, Filos D, et al. The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. Surg Radiol Anat. 2023;45:1321-9.

15. Meo SA, Al-Masri AA, Alotaibi M, et al. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. Healthcare (Basel). 2023;11:2046.

16. Clement J, Maldonado AQ. Augmenting the transplant team with artificial intelligence: toward meaningful AI use in solid organ transplant. Front Immunol. 2021;12:694222.

17. Ouanes K, Farhah N. Effectiveness of artificial intelligence (AI) in clinical decision support systems and care delivery. J Med Syst. 2024;48:74.

18. Pashkov VM, Harkusha AO, Harkusha YO. Artificial intelligence in medical practice: regulative issues and perspectıives. Wiad Lek. 2020;73:2722-7.

19. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17:195.

20. Benzinger L, Ursin F, Balke WT, et al. Should artificial intelligence be used to support clinical ethical decision-making? A systematic review of reasons. BMC Med Ethics. 2023;24:48.

21. Mörch CM, Atsu S, Cai W. et al. Artificial intelligence and ethics in dentistry: a scoping review. J Dent Res. 2021;100:1452-60.

22. Chen YW, Stanley K, Att W. Artificial intelligence in dentistry: current applications and future perspectives. Quintessence Int. 2020;51:248-57. Erratum in: Quintessence Int. 2020;51:430.

23. Duggal I, Tripathi T. Ethical principles in dental healthcare: Relevance in the current technological era of artificial intelligence. J Oral Biol Craniofac Res. 2024;14:317-21.

24. Sahin MC, Sozer A, Kuzucu P. et al. Beyond human in neurosurgical exams: ChatGPT's success in the Turkish neurosurgical society proficiency board exams. Comput Biol Med. 2024;169:10780