Imaging Science in Dentistry 2024; 54: 271-5
https://doi.org/10.5624/isd.20240037

Performance of ChatGPT 3.5 and 4 on U.S. dental examinations: the INBDE, ADAT, and DAT

Mahmood Dashti[1,*], Shohreh Ghasemi[2], Niloofar Ghadimi[3], Delband Hefzi[4], Azizeh Karimian[5], Niusha Zare[6], Amir Fahimipour[7], Zohaib Khurshid[8], Maryam Mohammadalizadeh Chafjiri[9], Sahar Ghaedsharaf[10]

[1] *Dentofacial Deformities Research Center, Research Institute of Dental Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran*
[2] *Department of Trauma and Craniofacial Reconstruction, Queen Mary College, London, England*
[3] *Department of Oral and Maxillofacial Radiology, Dental School, Islamic Azad University of Medical Sciences, Tehran, Iran*
[4] *School of Dentistry, Tehran University of Medical Science, Tehran, Iran*
[5] *Department of Biostatistics, Dental Research Center, Golestan University of Medical Sciences, Gorgan, Iran*
[6] *Department of Operative Dentistry, University of Southern California, CA, USA*
[7] *Discipline of Oral Surgery, Medicine and Diagnostics, School of Dentistry, Faculty of Medicine and Health, Westmead Centre for Oral Health, The University of Sydney, Sydney, Australia*
[8] *Department of Prosthodontics and Dental Implantology, King Faisal University, Al Ahsa, Kingdom of Saudi Arabia*
[9] *Department of Oral and Maxillofacial Pathology, School of Dentistry, Shahid Beheshti University of Medical Sciences, Tehran, Iran*
[10] *Department of Oral and Maxillofacial Radiology, School of Dentistry, Shahid Beheshti University of Medical Sciences, Tehran, Iran*

ABSTRACT

Purpose: Recent advancements in artificial intelligence (AI), particularly tools such as ChatGPT developed by OpenAI, a U.S.-based AI research organization, have transformed the healthcare and education sectors. This study investigated the effectiveness of ChatGPT in answering dentistry exam questions, demonstrating its potential to enhance professional practice and patient care.

Materials and Methods: This study assessed the performance of ChatGPT 3.5 and 4 on U.S. dental exams - specifically, the Integrated National Board Dental Examination (INBDE), Dental Admission Test (DAT), and Advanced Dental Admission Test (ADAT) - excluding image-based questions. Using customized prompts, ChatGPT's answers were evaluated against official answer sheets.

Results: ChatGPT 3.5 and 4 were tested with 253 questions from the INBDE, ADAT, and DAT exams. For the INBDE, both versions achieved 80% accuracy in knowledge-based questions and 66-69% in case history questions. In ADAT, they scored 66-83% in knowledge-based and 76% in case history questions. ChatGPT 4 excelled on the DAT, with 94% accuracy in knowledge-based questions, 57% in mathematical analysis items, and 100% in comprehension questions, surpassing ChatGPT 3.5's rates of 83%, 31%, and 82%, respectively. The difference was significant for knowledge-based questions ($P = 0.009$). Both versions showed similar patterns in incorrect responses.

Conclusion: Both ChatGPT 3.5 and 4 effectively handled knowledge-based, case history, and comprehension questions, with ChatGPT 4 being more reliable and surpassing the performance of 3.5. ChatGPT 4's perfect score in comprehension questions underscores its trainability in specific subjects. However, both versions exhibited weaker performance in mathematical analysis, suggesting this as an area for improvement. ***(Imaging Sci Dent 2024; 54: 271-5)***

KEY WORDS: Artificial Intelligence; Deep Learning; Dentistry; Education; Dental

Introduction

Artificial intelligence (AI) has become a trending topic in recent years, and its applications are expanding exponentially across various fields worldwide.[1] AI, or artificial

Received February 29, 2024; Revised April 21, 2024; Accepted April 27, 2024
Published online July 2, 2024

*Correspondence to : Dr. Mahmood Dashti
School of Dentistry, Shahid Beheshti University of Medical Sciences, Daneshju Blv., Velenjak St., Chamran Highway, Tehran 1983969411, Iran
Tel) 98-922-274-6706, E-mail) Dashti.mahmood72@gmail.com

intelligence, involves using computers to mimic human intelligence, performing tasks that typically require human capabilities such as understanding, reasoning, and decision-making.[1,2]

AI software will eventually provide an efficient way for the world of education and learning. AI programs are frequently used across a variety of fields such as engineering, marketing, and medicine.[3] Numerous tasks have become easier as a result of using AI in daily life.[2] Despite the advantages of AI, many people are still unfamiliar with its principles.

OpenAI is a private research laboratory that was founded in December 2015,[4] with the goal of making rapid progress in AI technologies.[4] One of its notable innovations is ChatGPT (short for "Generative Pre-Trained Transformer"), a publicly accessible tool.[4-6] ChatGPT is based on the GPT language model.[7] Alongside Microsoft Bing and Google Bard, ChatGPT is recognized as a prominent AI chatbot. It utilizes deep learning AI techniques to generate responses that closely mimic human interaction in natural language.[8] ChatGPT is an example of a large language model (LLM), which is trained on text and produces textual content.[9,10] LLMs have been applied to various areas in the medical field, and these models represent a notable advancement in the field of AI.[11]

In dentistry, ChatGPT offers a variety of services for medical personnel, including diagnosis, disease prevention, medication management, and reduction of medical errors.[11] These applications have the potential to significantly improve healthcare and dentistry by promoting patient engagement and self-determination.

ChatGPT is a chatbot that utilizes the GPT-3 and GPT-4 language models.[4,7] This tool is primarily designed to generate human-like responses to text inputs and fulfill a wide range of text-based requests, from answering complex questions to generating brief texts that mimic human language.[5,12] Users can access it through various platforms, including mobile apps and websites, either by text or voice.[5,6,10]

Recent publications have demonstrated that ChatGPT can accurately answer exam questions, including those from the United States Medical Licensing Examination. The authors have also challenged it to respond to a variety of other exam questions.[8]

This study aimed to further test and evaluate both the qualitative and quantitative performance of ChatGPT on exam questions in the field of dentistry in the U.S., to determine if ChatGPT is capable of reaching the passing threshold. The null hypothesis was that ChatGPT cannot assist dental students and dentists in answering dental questions.

Material and Methods

Official question samples from 3 U.S. dental examinations - specifically, the Integrated National Board Dental Examination (INBDE), Dental Admission Test (DAT), and Advanced Dental Admission Test (ADAT) - were collected.

INBDE and ADAT questions were categorized into two types: Knowledge-based questions and Case history questions. Additionally, questions that included images were excluded. Similarly, DAT questions were divided into three categories: Knowledge-based questions, Mathematical analysis questions, and Comprehension questions, with image-based questions also being excluded. For the INBDE and ADAT examinations, the following prompts were used to solicit responses from ChatGPT 3.5 and 4. For knowledge-based questions, the prompt was: "You are a dentist, which is taking a dental examination, please chose the best answer for the following question." For case history questions, the prompt was: "You are a dentist, which is taking a dental examination, based on the following case history, chose the best answer."

For the DAT examination, the following prompts were used to ask ChatGPT 3.5 and 4 to answer the questions. For knowledge-based questions, the prompt was: "You want to apply for dental school and need to pass the DAT (Dental admission test) exam, please answer the following questions from the DAT exam as best as you can." For mathematical analysis questions and chemical equation questions, the prompt was: "You want to apply for dental school and need to pass the DAT (Dental admission test) exam, please answer the following mathematical questions from the DAT exam." For comprehension questions, the prompt was: "You want to apply for dental school and need to pass the DAT (Dental admission test) exam, please answer the following questions from the DAT exam as best as you can. Based on the passage provided in the next prompt please answer the questions." The distribution of each question type for each examination is shown in Table 1.

Both ChatGPT 3.5 and 4 were asked to answer the ques-

Table 1. Breakdown of each examination

	Knowledge-based questions	Case history questions
INBDE	10	39
ADAT	59	21
DAT	124 (17 comprehension) (54 mathematical, and chemical equation questions) (53 knowledge-based question)	None

tions. The accuracy of their responses was checked based on the answer sheet that was provided by the examiner.

Each examination was evaluated separately according to the categories of the questions. Initially, the percentage of correct answers was assessed for each of the four categories: knowledge-based, case history, mathematical, and comprehension questions. The chi-square and Fisher exact tests were used to evaluate the performance of ChatGPT 3.5 compared to ChatGPT 4. The significance level was set at 0.05.

Results

The ChatGPT 3.5 and ChatGPT 4 models were used to examine a total of 49 questions from the INBDE examination, 80 questions from the ADAT examination, and 124 questions from the DAT examination.

ChatGPT 3.5 and ChatGPT 4 performed well on the INBDE, particularly in the knowledge-based questions, where both versions answered correctly 80% of the time. In the case history questions, ChatGPT 4 had a slightly higher success rate, answering 69% correctly, compared to ChatGPT 3.5, which answered 66% correctly (Fig. 1).

In the ADAT questions, ChatGPT 4 correctly answered 83% of the queries, while ChatGPT 3.5 had a success rate of 66%. Regarding the case history questions, both ChatGPT 3.5 and ChatGPT 4 achieved a correct response rate of 76% (Fig. 1).

For the DAT, in the knowledge-based questions, Chat GPT 4 correctly answered 94% of the questions, while ChatGPT 3.5 had a correct response rate of 83%. In mathematical analysis, ChatGPT 4 achieved a correct answer rate of 57%, compared to 31% for ChatGPT 3.5. The chi square test indicated that ChatGPT 4's improved performance was statistically significant ($P=0.021$). In chemistry calculations, ChatGPT 4 answered 68% of the questions correctly, whereas ChatGPT 3.5 answered 56% correctly; however, this difference was not statistically significant ($P=0.465$). For comprehension questions, ChatGPT 4 had a 100% correct response rate, significantly outperforming ChatGPT 3.5, which answered 82% correctly. The Fisher exact test revealed no significant difference in performance on these questions ($P=0.227$) (Fig. 1).

Overall, in the knowledge-based question section, Chat GPT 4 correctly answered 88% of the questions, while ChatGPT 3.5 correctly answered 78%. The chi-square test revealed that ChatGPT 4's performance was significantly better ($P=0.009$). In the case history questions section, ChatGPT 4 correctly answered 71% of the questions, compared to 70% for ChatGPT 3.5. The chi-square test indicated no significant difference in performance between the two versions ($P=0.841$) (Fig. 2).

Interestingly, both ChatGPT 3.5 and ChatGPT 4 provided the same incorrect answers for 6 out of the 9 knowledge-based questions they answered incorrectly, representing a 66% overlap. In the case history questions section, they both answered 12 questions incorrectly, with identical errors in 7 of these, indicating a 58% overlap. In the mathematical analysis section, they gave the same incorrect responses to 6 out of 11 incorrectly answered questions, a 54% overlap. Lastly, in the chemistry calculation questions section, they both answered 4 questions incorrectly, with 2 being the same, showing a 50% overlap. This pattern suggests that ChatGPT 3.5 and ChatGPT 4 tend to make similar errors (Fig. 3).

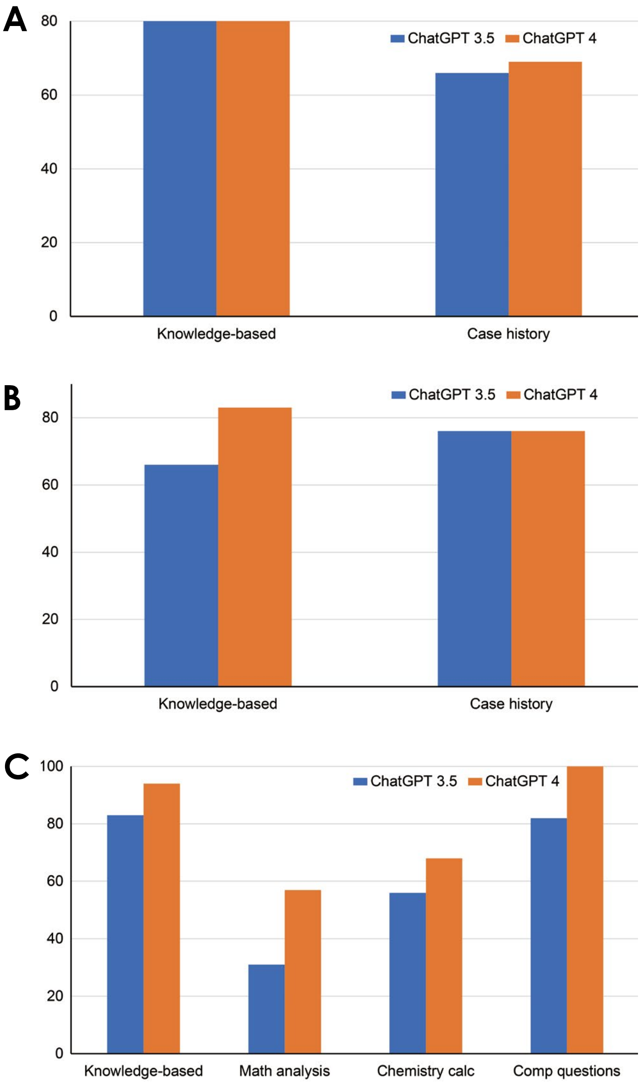

Fig. 1. Performance of ChatGPT 3.5 and 4 on the INBDE (A), ADAT (B), and DAT (C) examinations.

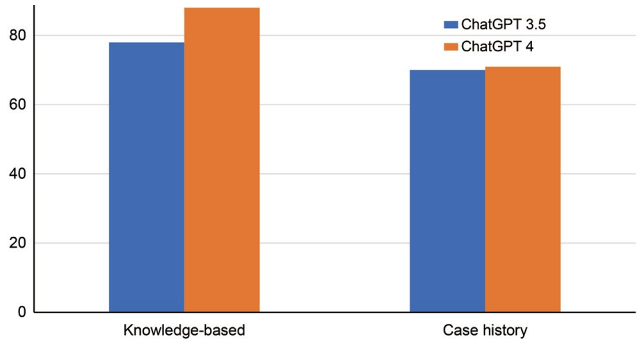

Fig. 2. Overall topics performance on question types.

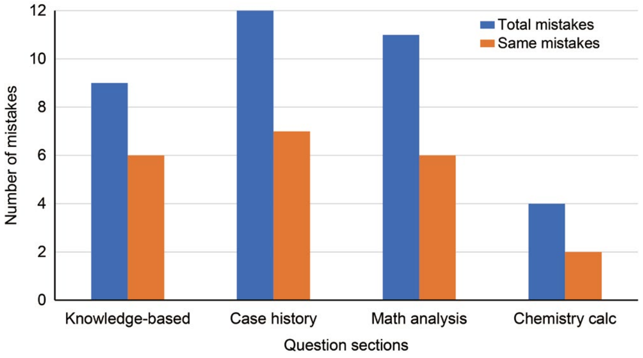

Fig. 3. Overlap in mistakes between different sections.

Discussion

Since its establishment in December 2015,[4] OpenAI has become a key player in the development of AI technology. A significant aspect of this advancement is the creation of the ChatGPT series,[7] which is based on the Generative Pre-Trained Transformer architecture. Notable advancements were made with the release of ChatGPT 3.5 and ChatGPT 4, with each version improving performance through training on increasingly larger text datasets.

To enhance the accuracy and intricacy of its responses, ChatGPT 3.5 was trained on a broader variety of materials, serving as a transitional enhancement over its earlier versions. The most current version, ChatGPT 4, builds on this foundation by incorporating an even larger and more diverse training sample. As a result, the model now delivers answers that are far more nuanced and contextually rich.[4]

In this study, ChatGPT - particularly ChatGPT 4 - demonstrated promising results in taking U.S.-based dental examinations. In the knowledge-based questions, ChatGPT 4 correctly answered 88% of the items, showing a statistically significant improvement over ChatGPT 3.5. Additionally, in the case history questions section, ChatGPT 4 correctly answered 71% of the questions. These results suggest that ChatGPT 4 can easily pass the dental board examination tests.

In the comprehension section, text was provided to ChatGPT 4 and it was asked questions. The model answered all questions correctly, demonstrating its capability to be trained effectively to meet the specific needs of this field.

Although ChatGPT demonstrated proficiency in knowledge-based, case history, and comprehension questions, its performance on mathematical questions was less impressive. In the ADAT mathematical questions section, Chat GPT 4 correctly answered 57% of the questions, while ChatGPT 3.5 managed only 31%. Despite the improved performance of ChatGPT 4 over ChatGPT 3.5 in mathematical analysis questions, it still cannot be considered a reliable resource for assisting students with these types of questions.

Another point to consider is the possibility of hallucination in LLMs, such as ChatGPT. Hallucination in an LLM is defined as “generated content that is nonsensical or unfaithful to the provided source content,” which can degrade system performance and fail to meet user expectations in many real-world scenarios.[12] This issue may arise when a user questions the LLM with prompts like, “Are you sure?” or “I think your answer is incorrect, can you double-check?” Such inquiries can confuse the LLM, leading it to change a correct response to an incorrect one. Hallucination can also occur in contexts such as dental examinations, highlighting the need for caution to prevent it when using these systems.

ChatGPT 3.5 and 4 have been developed and trained using data available on the internet up to the September 2021, and January 2023 respectively. It can address a wide range of questions effectively, demonstrating strong performance in knowledge-based and case history queries. Notably, it achieves a perfect score of 100% in the comprehension questions section. This underscores the capability of LLMs like ChatGPT to be trained extensively in specific subject areas, thereby maximizing their potential.

Many aspects of natural language processing software and LLMs remain to be tested. However, the current study demonstrates the potential of ChatGPT in assisting dentists and dental students during dental examinations. However, there is a need to develop more LLMs that are specifically trained and tailored for medical and dental subjects.

In conclusion, both ChatGPT 3.5 and ChatGPT 4 are capable of answering knowledge-based, case history, and comprehension questions effectively, achieving good scores. ChatGPT 4 outperformed ChatGPT 3.5 in most aspects, demonstrating greater reliability. Notably, ChatGPT 4 achieved a perfect score on the comprehension questions, reflecting its robust training in the user-selected subject matter. However, both versions exhibited weaker performance in mathematical analysis questions, underscoring the need for further improvement in this area.

Conflicts of Interest: None

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work the Z.Kh. and M.D. used ChatGPT 4 to paraphrase, improve the readability and enhance the language. After using this tool/service, all of the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Dashti M, Londono J, Ghasemi S, Khurshid Z, Khosraviani F, Moghaddasi N, et al. Attitudes, knowledge, and perceptions of dentists and dental students toward artificial intelligence: a systematic review. J Taibah Univ Med Sci 2024; 19: 327-37.
2. Yüzbaşıoğlu E. Attitudes and perceptions of dental students towards artificial intelligence. J Dent Educ 2021; 85: 60-8.
3. 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: 193-8.
4. Livberber T, Ayvaz S. The impact of Artificial Intelligence in academia: views of Turkish academics on ChatGPT. Heliyon 2023; 9: e19688.
5. Alhaidry HM, Fatani B, Alrayes JO, Almana AM, Alfhaed NK. ChatGPT in dentistry: a comprehensive review. Cureus 2023; 15: e38317.
6. Tiwari A, Kumar A, Jain S, Dhull KS, Sajjanar A, Puthenkandathil R, et al. Implications of ChatGPT in public health dentistry: a systematic review. Cureus 2023; 15: e40367.
7. Egli A. ChatGPT, GPT-4, and other large language models: the next revolution for clinical microbiology? Clin Infect Dis 2023; 77: 1322-8.
8. 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 Digit Health 2023; 2: e0000198.
9. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023; 11: 1166120.
10. Dashti M, Londono J, Ghasemi S, Moghaddasi N. How much can we rely on artificial intelligence chatbots such as the ChatGPT software program to assist with scientific writing? J Prosthet Dent (in press).
11. Chakravorty S, Aulakh BK, Shil M, Nepale M, Puthenkandathil R, Syed W. Role of Artificial Intelligence (AI) in dentistry: a literature review. J Pharm Bioallied Sci 2024; 16(Suppl 1): S14-6.
12. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2023; 55: 248.