DOI: 10.37990/medr.1446396 Med Records 2024;6(2):235-8

MEDICAL RECORDS-International Medical Journal

Research Article

Evaluation of ChatGPT as a Multiple-Choice Question Generator in Dental Traumatology

Yagiz Ozbay

Karabük University, Faculty of Dentistry, Department of Endodontics, Karabük, Türkiye

Abstract

Aim: This study aims to evaluate the ability of ChatGPT-4 to generate clinical case-based multiple-choice questions in dental traumatology.

Material and Method: International Association of Dental Traumatology guidelines for the management of traumatic dental injuries were introduced to ChatGPT-4 as an information source and ChatGPT-4 was prompted to 20 generate questions in fractures and luxations, avulsion of permanent teeth, injuries in the primary dentition. Sixty questions in total were generated by ChatGPT and the questions were evaluated by 2 endodontists using a 3-point scale. The One-way analysis of variance and Post Hoc Tukey test were used to analyze the data and the significance was $P<0.05$.

Results: The average time to generate 20 questions was 1 min 55 sec. It was noted that 52% of the questions were usable without modification or with minor changes, while 28% were incorrect or completely useless.

Conclusion: Despite the flaws, ChatGPT can be useful for creating multiple-choice questions in dental traumatology after a rigorous evaluation, elimination, and development procedure.

Keywords: Artificial intelligence, clinical practice guidelines, dental education, endodontics, treatment planning

INTRODUCTION

Dental education is a meticulously structured journey that prepares students for the challenging responsibilities of the occupation. It combines rigorous academic coursework with practical training. The goal is to ensure that upon graduation, these students are academically and practically competent. Dental traumatology (DT) education is an essential component of dental school curricula, addressing the diagnosis, treatment, and prevention of trauma to the teeth and surrounding oral structures. Moreover, DT is one of the important topics of both endodontics and pediatric dentistry, as it is a wide area ranging from an enamel fracture to tooth avulsion and may require complex treatments depending on the diagnosis. Traumatic dental injuries, often seen in sports, accidents, or falls, require immediate and effective intervention to save teeth and maintain oral function. By integrating DT into dental education, future dentists are equipped with the necessary skills and knowledge to manage these injuries effectively.

Assessment is a crucial part of understanding students' competence. Multiple choice questions (MCQs) are a common method of assessment of medical examinations because they are objective, standardized, and time-efficient (1). Preparation of MCQs that evaluate both the application and interpretation of knowledge rather than the recall of information only can be challenging (2). Furthermore, it was found that students were inclined to engage more thoroughly in their studies when faced with test questions that necessitated advanced analytical thinking (3). The university staff has various tasks including treatment of patients, research, and teaching, etc (4). Therefore, reducing the workload through automation of appropriate tasks can be useful for the productive use of time and to avoid staff burnout.

Education is among the fields that potentially be revolutionized by artificial intelligence (AI) (5). AI is a rapidly developing phenomenon that includes many different technologies, such as machine learning and natural language processing. AI technologies can assist

CITATION

Ozbay Y. Evaluation of ChatGPT as a Multiple-Choice Question Generator in Dental Traumatology. Med Records. 2024;6(2):235-8. DOI:1037990/medr.1446396

Received: 03.03.2024 **Accepted:** 19.04.2024 **Published:** 08.05.2024

Corresponding Author: Yagiz Ozbay, Karabük University, Faculty of Dentistry, Department of Endodontics, Karabük, Türkiye

E-mail: yagiz_ozbay@hotmail.com

healthcare professionals by automating tasks that are repetitive and consume a lot of time (6). Large Language Models (LLMs) are advanced AI systems that emulate human language processing skills by training on extensive datasets. These models are capable of comprehending and generating the content, allowing them to perform tasks like article generation and answering questions in a near-human manner (7,8).

ChatGPT (Chat Generative Pre-Trained Transformer) (OpenAI, San Francisco, CA, USA) is a promising LLM, utilizing deep learning AI techniques to generate articulate and human-resembling texts (9). Although studies (10-13) questioning the ability of ChatGPT to answer questions correctly related to different fields of dentistry and its use as an information source are increasing, it has been found in different studies that it can produce texts containing unrealistic and misleading information, which is called *"hallucination"*. ChatGPT's ability to generate questions has also been evaluated in various studies (14,15) and different results have been reported. Cheung et al. (14) found that ChatGPT can generate medical MCQs with comparable quality to university staff. ChatGPT-3.5, the older version, was found to generate suboptimal MCQs without distractor options in dermatology (15). ChatGPT-4, as the newest version, has features that its predecessors did not have, such as enabling image and document input and internet access. Thus, when a text is uploaded as a document, it can understand, summarise, and answer questions about that text (16,17). Given these features, it can be speculated that ChatGPT-4 can produce high-quality and reliable MCQs based on an existing text. Deriving questions by introducing documents representing consensus in dentistry, such as position statements of dental associations and clinical guidelines, into ChatGPT-4 can shorten the question preparation time and improve the quality of the evaluated questions. Thereby, the production of texts containing unrealistic references and information can be prevented.

To the best of our knowledge, there is no study in the literature questioning the ability of LLMs such as ChatGPT to generate questions in any other field of dentistry. Thus, the aim of this study is to evaluate the ability of ChatGPT-4 to generate clinical case-based MCQs in DT. The null hypothesis was as follows: There is no association between the type of dental trauma (fractures and luxations, avulsion of permanent teeth, injuries in the primary dentition) and the usability of MCQs generated by ChatGPT-4.

MATERIAL AND METHOD

The study was conducted under the Declaration of Helsinki. Since the study had no human subjects involved, ethical approval was not required. ChatGPT-4 was logged in on 14 January 2024. International Association of Dental Traumatology (IADT) guidelines for the management of traumatic dental injuries represent the consensus of experts in the field and are up-to-date as a resource prepared by the working group of the IADT as a result of a comprehensive literature review and approved by the board of directors (18-21). In addition, since these are open-access articles, they were used as a source to derive questions in our study.

The first part of the guidelines, *"fractures and luxations"*, was uploaded to the ChatGPT console and the following prompt was written exactly as follows: *"Can you write a full set of 20 multiple-choice questions based on the document I provided, each with 5 options and focusing solely on clinical case scenarios in dental traumatology? An answer key is required at the end."* (Figure 1). Then the second section, avulsion of permanent teeth, and the third section, injuries in the primary dentition, were uploaded in separate tabs and questions were generated with the same prompt. Question generation times for all sections were noted and as a result, 60 questions were obtained, and these questions were copied and saved. The two reviewers jointly assessed and scored the questions based on the guidelines. In case of disagreement, the decision of a third reviewer was decisive. In the evaluation, the evaluation criteria used in a similar study by Ngo et al. (22) was adopted (Table 1).

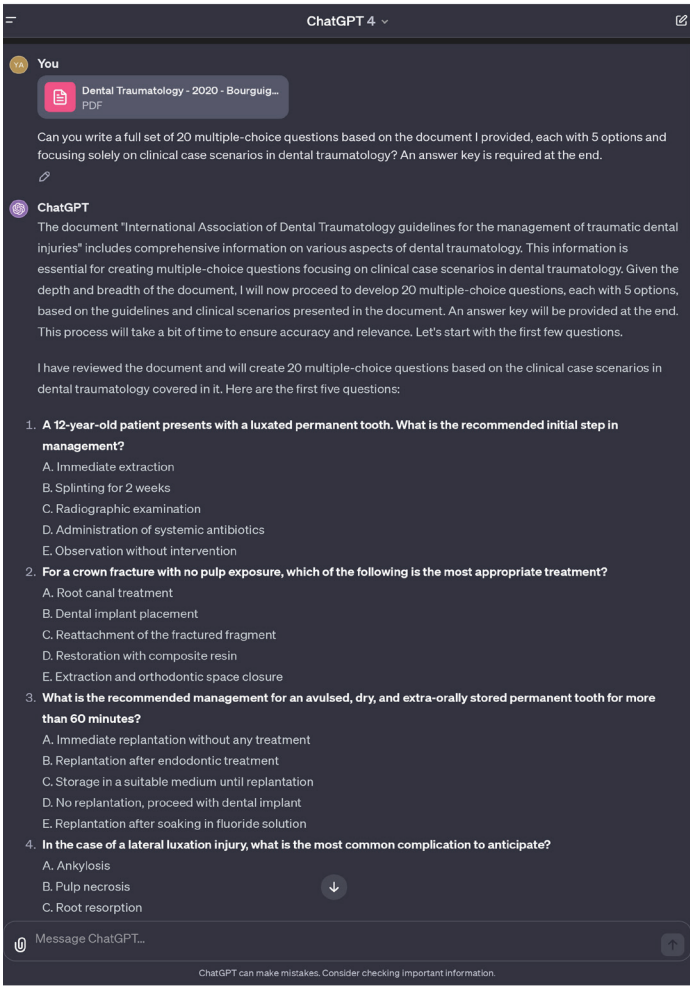

Figure 1. Generation of multiple-choice questions via ChatGPT-4

Table1. Scoring definition	
Score	**Description**
1	Questions and answers were correct and may be used with only minor formatting
2	Questions and answers would require substantial modifications to be appropriate
3	Answers were incorrect or significantly misleading

Statistical Analysis

The statistical analysis was performed via MiniTab 17 (Minitab Inc., PA, USA). The normality of the data was assessed with Ryan-Joiner test and the normal distribution of the data was confirmed. One-way analysis of variance (ANOVA) and Post Hoc Tukey test were performed. A significance level of P<0.05 was used to determine statistical significance.

RESULTS

The mean values and standard deviations of each group are presented in Table 2. Average score of all DT questions was 1,55. *"Avulsion of permanent"* teeth demonstrated the best score followed by *"injuries in the primary dentition"* and *"fractures and luxations"* respectively. The difference between *"fractures and luxations"* and *"avulsion of permanent teeth"* was statistically significant (P=0.029). After the prompt was input, the average time to generate 20 questions was 1 min 55 sec. It was noted that 52% of the questions were usable without modification or with minor changes, while 28% were incorrect or completely useless.

Table 2. Scores of questions by topic						
Topic	**Number of questions**	**Mean**	**Standard deviation**	**Score-1**	**Score-2**	**Score-3**
Fractures and luxations[A]	20	2.00	0.858	7	6	7
Avulsion of permanent teeth[B]	20	1.35	0.587	14	5	1
Injuries in the primary dentition[A,B]	20	1.95	0.999	10	1	9
Dental traumatology	60	1.55	0.790	31 (52%)	12 (20%)	17 (28%)
Different letters indicate statistically significant difference (P<0.05)						

DISCUSSION

The null hypothesis was that there is no association between the type of dental trauma and the usability of MCQs generated by ChatGPT-4. The null hypothesis can be partially rejected as the analysis revealed a significant association between the type of dental trauma and question usability.

It was found that all the questions were within the scope of the subject. It was also concluded that there were no spelling or grammatical errors in the questions and all questions were understandable. The ability to prepare 20 questions within the scope of the subject in a short time such as an average of 1 min 55 sec shows that ChatGPT is a potential tool that can save time for the faculty members. In agreement with our study, Cheung et al. (14) also reported that ChatGPT prepared 50 medical MCQs in 20 min 25 sec, which is much faster compared to the professional staff who prepared 50 medical MCQs in 211 min 33 sec.

The reason why questions with case scenarios were included in the study is that it is important to measure the usability of information in addition to the recall of information. The common deficiency observed in some of the questions produced by ChatGPT was that the case stories were quite superficial and lacked detail. For example, one of the prepared questions was as follows: *"A 15-year-old patient presents with a luxated and non-responsive tooth. The most likely diagnosis is:..."* It can be expected that more relevant and high-quality questions would contain more details about the cases. This may make it necessary to revise the ChatGPT questions rather than using them as they are. Therefore, it is not possible to claim that all the questions produced by ChatGPT are high quality clinical case-based questions, although the dental trauma questions are within the context and relevant.

When preparing a case question, depending on the question, the presence of periapical X-rays or intraoral photographs may be essential or may improve the quality of the question. Unlike its predecessors, ChatGPT-4 has the feature of creating images as instructed and within the limits of skill. However, it is apparent that ChatGPT-4's inability to create medical images reduces the question quality and this is a deficiency in its use for MCQ preparation. Considering the pace of development of LLMs, it may be useful to re-evaluate the usability of ChatGPT and other language models in dental education when more advanced versions become available.

ChatGPT and other LLMs producing false information and references have been observed before in various studies (10-13,22) in various fields. In this study, to avoid this situation, which is called hallucination, and to obtain questions based on accurate and up-to-date information, IADT's guidelines were introduced to ChatGPT, and it was aimed to prepare the questions within the framework of these guidelines. The results showed that introducing sources to ChatGPT did not prevent completely the production of text containing false information.

In previous studies (14,22) in which MCQs were prepared based on a selected source, the researchers introduced the information source to ChatGPT by selecting a text, copying and pasting it into ChatGPT. The reason for using ChatGPT-4 in the present study is that, unlike its

predecessors and other LLMs, it allows documents to be uploaded as files. Thus, in this study, unlike the previous studies, sources were directly uploaded as files. The direct uploadability of the source on which the questions are based may offer ease of use.

According to the results of this study, only 52% of the questions produced by ChatGPT can be used without major revision. Ngo et al. (22) reported that only 32% of the MCQs generated by ChatGPT-3.5 in pathology were correct. However, it should be taken into consideration that the previous version of ChatGPT was used in the study of Ngo et al. However, these results are in line with our study in terms of the presence of completely wrong or useless questions.

CONCLUSION

This is the first study to evaluate the question preparation ability of an LLM in dentistry and evaluated the MCQs produced by ChatGPT-4 on DT and considered ChatGPT-4 as a potential question bank. Although ChatGPT has the potential to be used as a question bank in dental education, it can only remain as a "potential" unless the production of misinformation and the lack of creative medical writing are completely eliminated. The questions produced by ChatGPT can only be used in their current form after a serious evaluation and revision.

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: *Since the study had no human subjects involved, ethical approval was not required.*

REFERENCES

1. Javaeed A. Assessment of higher ordered thinking in medical education: multiple choice questions and modified essay questions. MedEdPublish. 2018;7:128.
2. Scully D. Constructing multiple-choice items to measure higher-order thinking. PARE. 2019;22:4.
3. Scouller K. The influence of assessment method on students' learning approaches: multiple choice question examination versus assignment essay. Higher Education; Dordrecht. 1998;35:453-72.
4. Rao SK, Kimball AB, Lehrhoff SR, et al. The impact of administrative burden on academic physicians: results of a hospital-wide physician survey. Acad Med. 2017;92:237-43.
5. Chen L, Chen P, Lin Z. Artificial intelligence in education: a review. IEEE Access. 2020;8:75264-78.
6. Cardoso MJ, Houssami N, Pozzi G, Séroussi B. Artificial intelligence (AI) in breast cancer care-leveraging multidisciplinary skills to improve care. Artif Intell Med. 2022;123:102215.
7. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ. 2023;9:e48291.
8. Mesko B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023;6:120.
9. Fatani B. ChatGPT for future medical and dental research. Cureus. 2023;15:e37285.
10. Giannakopoulos K, Kavadella A, Stamatopoulos V, Kaklamanos E. Evaluation of generative artificial intelligence large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: a comparative mixed-methods study. J Med Internet Res. 2023;25:e51580.
11. Acar AH. Can natural language processing serve as a consultant in oral surgery?. J Stomatol Oral Maxillofac Surg. 2024;125:101724.
12. Suarez A, Diaz-Flores Garcia V, Algar J, et al. Unveiling the ChatGPT phenomenon: evaluating the consistency and accuracy of endodontic question answers. Int Endod J. 2024;57:108-13.
13. Ali K, Barhom N, Tamimi F, Duggal M. ChatGPT-A double-edged sword for healthcare education? Implications for assessments of dental students. Eur J Dent Educ. 2024;28:206-11.
14. Cheung BHH, Lau GKK, Wong GTC, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions-A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS One. 2023;18:e0290691.
15. Ayub I, Hamann D, Hamann CR, Davis MJ. Exploring the potential and limitations of chat generative pre-trained transformer (ChatGPT) in generating board-style dermatology questions: a qualitative analysis. Cureus. 2023;15:e43717.
16. Kim HW, Shin DH, Kim J, et al. Assessing the performance of ChatGPT's responses to questions related to epilepsy: a cross-sectional study on natural language processing and medical information retrieval. Seizure. 2024;114:1-8.
17. Spallek S, Birrell L, Kershaw S, et al. Can we use ChatGPT for mental health and substance use education? Examining its quality and potential harms. JMIR Med Educ. 2023;9:e51243.
18. Bourguignon C, Cohenca N, Lauridsen E, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: 1. fractures and luxations. Dent Traumatol. 2020;4:314-30.
19. Levin L, Day PF, Hicks L, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: general introduction. Dent Traumatol. 2020;4:309-13.
20. Fouad AF, Abbott PV, Tsilingaridis G, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: 2. avulsion of permanent teeth. Dent Traumatol. 2020;36:331-42.
21. Day PF, Flores MT, O'Connell AC, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: 3. injuries in the primary dentition. Dent Traumatol. 2020;36:343-59.
22. Ngo A, Gupta S, Perrine O, et al. ChatGPT 3.5 fails to write appropriate multiple choice practice exam questions. Acad Pathol. 2023;11:100099.