Submitted: 10 September, 2025 Accepted: 31 October, 2025 Published: 03 March, 2026 DOI:10.22514/jocpd.2026.049

Journal of Clinical Pediatric Dentistry

ORIGINAL RESEARCH

Artificial intelligence applications in tooth avulsion: comparative accuracy of ChatGPT and DeepSeek

Gizem Karagöz Doğan[1,*], Yelda Polat Yavuz[2], İzzet Yavuz[2]

[1]Department of Pediatric Dentistry, Faculty of Dentistry, Iğdır University, 76000 Iğdır, Turkey
[2]Department of Pediatric Dentistry, Faculty of Dentistry, Dicle University, 21000 Diyarbakır, Turkey

***Correspondence**
gizem.dogan@igdir.edu.tr
(Gizem Karagöz Doğan)

Abstract

Background: The accuracy and performance of artificial intelligence (AI)-based chatbots in clinical applications can directly influence healthcare outcomes. In cases of dental trauma, adherence to the International Association of Dental Traumatology (IADT) guidelines is essential for clinical success. Although the use of AI in healthcare is increasing, few studies have evaluated the ability of chatbots to provide accurate information in dental trauma. This study aimed to evaluate and compare the performance of the ChatGPT and DeepSeek platforms in providing guideline-based information on the management of dental avulsion, using the IADT guidelines as a reference standard. **Methods**: Based on the IADT guidelines, 25 questions (12 yes/no and 13 open-ended) were posed to ChatGPT-3.5 and DeepSeek over the course of one week. Two independent researchers asked each question three times daily. Responses were classified as correct, incorrect, or insufficient according to the guidelines. Statistical analyses were conducted to assess agreement and accuracy. **Results**: A total of 1050 responses were analyzed. DeepSeek demonstrated moderate agreement with the guideline-based answers ($\kappa \approx 0.52$; 95% confidence interval (CI): 0.48–0.55; $p < 0.001$), whereas ChatGPT showed weak-to-moderate agreement ($\kappa \approx 0.44$; 95% CI: 0.40–0.48; $p < 0.001$). The mean accuracy difference between the two platforms was approximately 7% (p = 0.001). **Conclusions**: ChatGPT and DeepSeek have potential as knowledge resources for healthcare applications. However, their accuracy and consistency in addressing dental avulsion-related questions remain limited. Clinicians should consider these systems as complementary tools that support, but do not replace, clinical expertise and decision-making. Further research should explore AI models specifically trained in dental trauma to determine their clinical utility.

Keywords

Artificial intelligence; Chatbots; ChatGPT; Tooth avulsion; DeepSeek; Large language models

1. Introduction

The application of artificial intelligence (AI) in healthcare has evolved since the 1950s, when computers were first used to analyze medical data and assist in diagnosis. Today, advanced medical systems equipped with deep learning algorithms are capable of achieving remarkably high levels of accuracy. Contemporary AI applications in healthcare encompass diagnostic imaging, personalized treatment planning, patient monitoring, and disease risk assessment. These technologies streamline clinical workflows, enhancing both the efficiency and effectiveness of medical services. In parallel with technological advances, the multifaceted nature of human perception has inspired the development of large language models (LLMs). Designed to emulate human language processing capabilities, LLMs function as AI-driven conversational agents. These models utilize deep learning techniques, such as neural networks, and are trained on vast text corpora derived from books, scholarly articles, and online sources. Through extensive training, LLMs acquire the ability to generate highly coherent and contextually appropriate texts. By analyzing linguistic patterns and contextual cues within their training data, LLMs can predict the most probable words or expressions in a given context [1, 2].

ChatGPT (GPT-3.5 Turbo, OpenAI Inc., San Francisco, CA, USA) and DeepSeek (V3.2-Exp, Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd., Hangzhou, Zhejiang, China) are among the LLMs developed for this purpose. While ChatGPT has demonstrated remarkable capabilities, it also presents several challenges, including high computational costs, scalability limitations, and difficulties in adapting to novel or domain-specific tasks. These constraints have highlighted the need to make AI systems more efficient, cost-effective, and adaptable. Consequently, researchers have sought solutions through the development of lightweight models and the enhancement of continuous learning capacities. In this context, DeepSeek was introduced as a next-generation LLM, designed to overcome the fundamental obstacles that restrict the applicability of traditional models. It integrates a series of innovative approaches in architectural design, learning paradigms, and optimization techniques, aiming to improve both performance and adaptability [3, 4].

Tooth avulsion is a severe form of traumatic dentoalveolar injury characterized by the complete displacement of a tooth from its socket. Traumatic tooth avulsion injuries (TTAIs) account for approximately 16% of all traumatic dental injuries, with a notably higher prevalence among children [5].

A substantial proportion of adverse outcomes following trauma arises from inadequate or inappropriate emergency management, which may lead to both functional and aesthetic complications. The success of treatment depends not only on immediate care at the site of injury, but also on the clinician's knowledge and expertise. Therefore, it is essential that dentists, as well as parents, teachers, sports coaches, and other professionals who frequently interact with children, possess adequate knowledge regarding the emergency management of dental trauma [6]. However, numerous studies have demonstrated that both the general population and professionals involved in the management of dentoalveolar injuries often lack sufficient understanding of this topic [7–15].

In the healthcare domain, patients frequently seek information online due to factors such as limited access to healthcare providers or curiosity about the experiences of individuals with similar medical histories. Clinicians may also rely on AI-driven conversational agents in managing time-sensitive cases, such as dental trauma, where prompt and accurate intervention is critical. The accuracy and performance of these models in clinical applications can have a direct impact on patient outcomes. Although LLMs continue to advance rapidly across diverse domains, their application in medicine and dentistry remains limited, underscoring the need for further empirical research in these fields [2].

Adherence to the International Association of Dental Traumatology (IADT) guidelines is essential for achieving optimal clinical outcomes in traumatic injury management. However, despite the growing integration of AI in healthcare, few studies have investigated whether AI-driven conversational agents can provide accurate and guideline-consistent responses in the context of dental trauma [6].

To date, no study in pediatric dentistry has specifically evaluated the effectiveness of AI-driven conversational agents in delivering accurate information to clinicians and patients regarding dental avulsion, nor compared the informational performance of different platforms. Accordingly, the present study seeks to fill this gap by providing insights into the reliability of AI-assisted knowledge systems and addressing existing uncertainties, thereby offering guidance for future clinical and educational applications.

The aim of this study was to evaluate and compare the abilities of ChatGPT and DeepSeek to provide accurate information on the management of dental avulsion cases, using the IADT guidelines as the reference standard. The first null hypothesis was that "there is no significant difference between ChatGPT and DeepSeek in their ability to provide accurate information to parents and clinicians regarding dental avulsion". The second null hypothesis was that "the accuracy of LLM-generated responses to questions concerning avulsion management would fall below the diagnostic accuracy threshold commonly accepted in clinical research".

2. Materials and methods

2.1 Study design and ethics

This cross-sectional study was conducted in accordance with the principles of the Declaration of Helsinki. Ethical approval was not required because only publicly available data were analyzed, and no biological materials from humans or animals were involved.

2.2 Study protocols

To obtain information generated by AI models, ChatGPT-3.5 and DeepSeek were utilized. DeepSeek, a relatively new platform, has been introduced as a competitor to ChatGPT, claiming to overcome limitations such as high computational costs, scalability challenges, and reduced adaptability to novel tasks. Both models were included in this study because they are freely accessible to the public.

The "Guidelines for the Evaluation and Management of Traumatic Dental Injuries" published by the IADT in 2020 served as the reference standard. Based on these guidelines, a total of 25 questions were developed—12 binary (yes/no) and 13 open-ended—to simulate inquiries that parents or dentists might pose to an AI system regarding dental avulsion (Table 1). All questions were designed according to the 2020 IADT guidelines to ensure both scientific accuracy and clinical relevance.

Clinical Practice-Oriented Questions: Some questions reflected scenarios parents might face at the scene of an accident resulting in avulsion, such as: "*What is the most appropriate way to hold an avulsed tooth when found?*", "*In an avulsion case, is immediate replantation at the accident site the best treatment?*", "*What are the most suitable storage media for an avulsed tooth?*". Others were targeted dentists seeking rapid and accurate guidance from AI systems during case management, including: "*What are the ideal characteristics of a splint in dental avulsion cases?*", "*What should a patient be mindful of during the splinting process of an avulsed tooth?*", "*Under what conditions should endodontic treatment be initiated in cases of avulsion involving open-apex teeth?*". All questions were independently reviewed by two pediatric dentistry specialists to confirm their accuracy, clarity and clinical applicability. Revisions were made based on their feedback.

Topic-Centered Analysis: Rather than addressing the entire scope of dental trauma, this study focused exclusively on avulsion, a type of injury that demands the most immediate and precise clinical decision-making among traumatic dental injuries.

The prepared questions were submitted by each researcher to the AI platforms three times daily between 03 February and 09 February 2025. To minimize the potential influence of response timing, both researchers conducted each session simultaneously. The platforms were accessed via Google's search

TABLE 1. Questions.

No.	Questions
1	What is the most appropriate part of an avulsed tooth to hold when found?
2	In an avulsion case, is the best treatment to replant the tooth at the accident site?
3	What are the most suitable storage conditions for an avulsed tooth?
4	Should systemic antibiotic use be recommended after replantation of an avulsed tooth?
5	What are the ideal characteristics of a splint in dental avulsion cases?
6	What should the patient be mindful of during the splinting process of an avulsed tooth?
7	In cases of avulsion of open-apex teeth, under what conditions should endodontic treatment be initiated?
8	What should be the follow-up schedule for replanted closed-apex teeth?
9	Is more frequent follow-up required in avulsion cases of open-apex teeth?
10	What are the favorable replantation outcomes for closed-apex teeth?
11	What are the favorable replantation outcomes for open-apex teeth?
12	What are the unfavorable replantation outcomes for closed-apex teeth?
13	What are the unfavorable replantation outcomes for open-apex teeth?
14	What treatment options are available for infra-position in replanted teeth?
15	What is the ideal splinting duration for avulsed teeth in the absence of other complications?
16	What is the ideal splinting duration when there are complications such as alveolar bone fractures in avulsed teeth?
17	If there is a fracture in the socket wall of a permanent avulsed tooth, should replantation be delayed until the fracture heals?
18	Since the blood clot in the socket of a permanent avulsed tooth supports healing, should it be left untouched even if it obstructs ideal tooth positioning during replantation?
19	Should root canal treatment be initiated within the first two weeks after replantation of a closed-apex permanent avulsed tooth?
20	Should root canal treatment be initiated within the first two weeks after replantation of an open-apex permanent avulsed tooth?
21	Is the apical status (open or closed) of an avulsed permanent tooth important in its repositioning?
22	Is the elapsed time before replantation of an avulsed permanent tooth important?
23	Should tetanus vaccination be recommended for every avulsion case?
24	Should an avulsed primary tooth be replanted?
25	Is avulsion the type of dental injury with the highest risk of ankylosis?

engine interface (Google LLC, Mountain View, CA, USA). To prevent bias from search algorithms, ChatGPT was accessed using a newly created account (https://chat.openai.com) established exclusively for this study during the specified dates. Prior to each question–answer session, all browsing history and cookies were cleared from the computer. To further reduce the potential effect of previous responses, a "new conversation" window was initiated for each question category. The initial responses of the platforms were recorded without additional prompting or follow-up queries. All responses were documented in Microsoft Excel (Microsoft Excel 365, Microsoft Corporation, Redmond, WA, USA). At the end of the seven-day period, the collected responses were compared with the correct answers outlined in the IADT guidelines. Responses were coded as 1 = correct, 0 = incorrect, and 2 = insufficient when the answers were only partially aligned with the guideline criteria. In this study, the term "insufficient" was operationally defined based on the recommendations outlined in the IADT guidelines. The chatbot responses were compared against the correct procedural steps specified in the IADT guidelines, and any response that did not fully align with these reference steps was considered "insufficient". For instance, if the guideline described a four-step management protocol and the chatbot accurately addressed only three of those steps, the response was categorized as insufficient due to incomplete information. Therefore, the adequacy of responses was objectively determined according to the scientific accuracy and comprehensiveness of the IADT guidelines.

2.3 Statistical analysis

Before data analysis, the evaluators were calibrated using a pilot response set of 10 questions. Inter-evaluator agreement was assessed using Cohen's kappa coefficient (κ = 0.88), indicating excellent reliability. Additionally, three pediatric dentistry experts independently evaluated all questions for clarity, accuracy, and redundancy using a 4-point Likert scale.

The Scale-level Content Validity Index (S-CVI) was calculated as 0.87, confirming good content validity. All statistical analyses were performed using IBM SPSS Statistics version 29.0 (IBM Corp., Armonk, NY, USA). To examine the effects of the program factor and the program × time interaction on the dependent variable(s), a repeated measures multivariate analysis of variance (Repeated Measures MANOVA) was conducted. Four multivariate test statistics were employed—Pillai's Trace, Wilks' Lambda, Hotelling's Trace, and Roy's Largest Root—to test the same hypothesis through different mathematical approaches and assess the consistency of the results. A significance level of $p < 0.05$ was considered statistically significant.

Fig. 1 illustrates the sequential stages of the study, encompassing question development, expert validation, AI response collection, classification, and statistical analysis.

3. Results

A total of 42 responses were collected for each question, yielding 1050 responses in total. Statistical analyses were conducted on the responses obtained across three time intervals—morning, noon, and evening. The distribution of responses by platform is presented in Table 2. According to the analysis, ChatGPT produced 38.3% correct, 49.3% incorrect, and 12.4% insufficient responses. In comparison, DeepSeek yielded 42.7% correct, 39.6% incorrect, and 17.7% insufficient responses. Collectively, both applications generated 40.47% correct responses (Table 2).

The accuracy rates of the platforms across three different time intervals (morning, noon, and evening) over the course of one week are presented in Fig. 2.

The agreement between DeepSeek and the guideline-based answers was found to be moderate ($\kappa \approx 0.52$; 95% confidence interval (CI): 0.48–0.55; $p < 0.001$). In contrast, the agreement between ChatGPT and the guideline answers was weak to moderate ($\kappa \approx 0.44$; 95% CI: 0.40–0.48; $p < 0.001$). The mean accuracy difference between the two programs was 7%, which was statistically significant ($p = 0.001$).

In the time-based analysis, DeepSeek produced significantly more accurate responses during the morning sessions of Days 1 and 2, as well as the noon sessions of Days 1 and 2. ChatGPT, however, demonstrated significantly higher accuracy only during the noon session of Day 7 (Table 3).

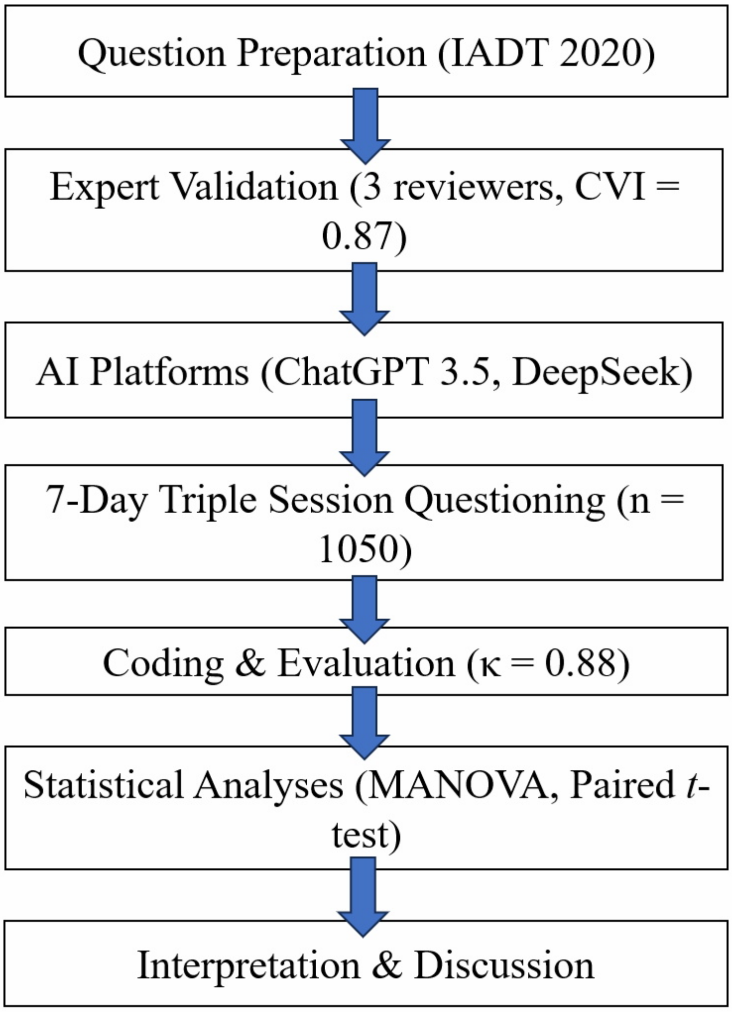
Question Preparation (IADT 2020)
Expert Validation (3 reviewers, CVI = 0.87)
AI Platforms (ChatGPT 3.5, DeepSeek)
7-Day Triple Session Questioning (n = 1050)
Coding & Evaluation (κ = 0.88)
Statistical Analyses (MANOVA, Paired t-test)
Interpretation & Discussion

FIGURE 1. Flowchart illustrating the overall study design and analytical process. IADT: International Association of Dental Traumatology; CVI: Content Validity Index; AI: artificial intelligence; MANOVA: multivariate analysis of variance. κ: Cohen's kappa coefficient.

TABLE 2. Distribution of responses regardless of time according to LLMs.

	DeepSeek		ChatGPT		Total	
	n	%	n	%	n	%
Responses						
Incorrect	208	39.6%	259	49.3%	467	44.5%
Correct	224	42.7%	201	38.3%	425	40.5%
Insufficient	93	17.7%	65	12.4%	158	15.0%
Total	525	100.0%	525	100.0%	1050	100.0%

Values are presented as n (%).

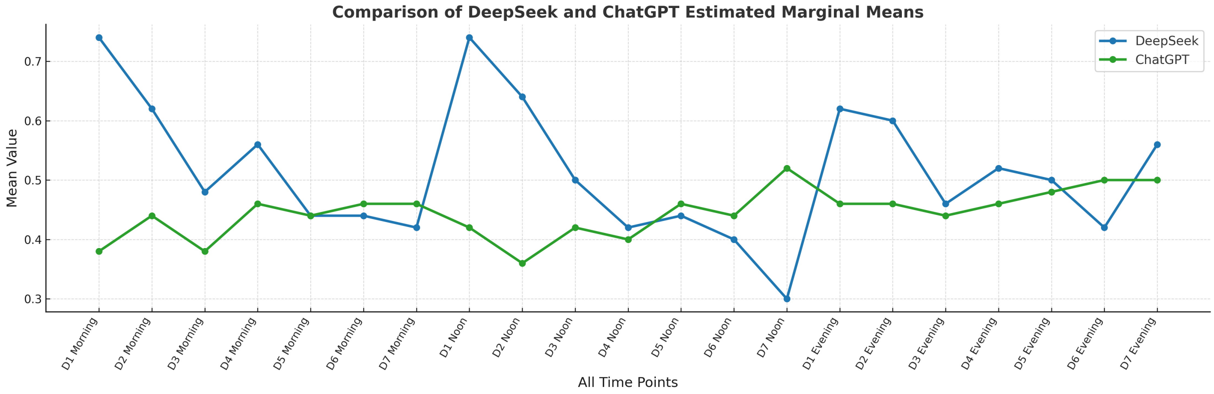

FIGURE 2. The accuracy rates of the platforms across three different time intervals over the course of one week. D: Day.

TABLE 3. Analysis of the differences between DeepSeek and ChatGPT across all time intervals.

	n	Mean	Median	Min	Max	Std.	Paired *T* Test	
							t	*p*
D1 Morning								
DeepSeek	25	0.74	1.0	0	1	0.357	3.524	0.002*
ChatGPT	25	0.38	0.0	0	1	0.463		
D2 Morning								
DeepSeek	25	0.62	1.0	0	1	0.440	2.377	0.026*
ChatGPT	25	0.44	0.0	0	1	0.486		
D3 Morning								
DeepSeek	25	0.48	0.5	0	1	0.467	1.414	0.170
ChatGPT	25	0.38	0.0	0	1	0.463		
D4 Morning								
DeepSeek	25	0.56	0.5	0	1	0.464	1.095	0.284
ChatGPT	25	0.46	0.5	0	1	0.477		
D5 Morning								
DeepSeek	25	0.44	0.5	0	1	0.464	0	1.000
ChatGPT	25	0.44	0.5	0	1	0.464		
D6 Morning								
DeepSeek	25	0.44	0.5	0	1	0.464	−0.196	0.846
ChatGPT	25	0.46	0.5	0	1	0.477		

TABLE 3. Continued.

	n	Mean	Median	Min	Max	Std.	Paired *T* Test	
							t	*p*
D7 Morning								
DeepSeek	25	0.42	0.0	0	1	0.472	−0.492	0.627
ChatGPT	25	0.46	0.5	0	1	0.477		
D1 Noon								
DeepSeek	25	0.74	1.0	0	1	0.357	3.216	0.004*
ChatGPT	25	0.42	0.0	0	1	0.472		
D2 Noon								
DeepSeek	25	0.64	1.0	0	1	0.421	3.412	0.002*
ChatGPT	25	0.36	0.0	0	1	0.468		
D3 Noon								
DeepSeek	25	0.50	0.5	0	1	0.456	1.000	0.327
ChatGPT	25	0.42	0.0	0	1	0.472		
D4 Noon								
DeepSeek	25	0.42	0.5	0	1	0.449	0.238	0.814
ChatGPT	25	0.40	0.0	0	1	0.479		
D5 Noon								
DeepSeek	25	0.44	0.5	0	1	0.464	−0.238	0.814
ChatGPT	25	0.46	0.5	0	1	0.477		
D6 Noon								
DeepSeek	25	0.40	0.0	0	1	0.456	−0.440	0.664
ChatGPT	25	0.44	0.0	0	1	0.486		
D7 Noon								
DeepSeek	25	0.30	0.0	0	1	0.433	−2.529	0.018*
ChatGPT	25	0.52	0.5	0	1	0.467		
D1 Evening								
DeepSeek	25	0.62	1.0	0	1	0.440	1.693	0.103
ChatGPT	25	0.46	0.5	0	1	0.477		
D2 Evening								
DeepSeek	25	0.60	0.5	0	1	0.408	1.899	0.070
ChatGPT	25	0.46	0.5	0	1	0.477		
D3 Evening								
DeepSeek	25	0.46	0.5	0	1	0.455	0.238	0.814
ChatGPT	25	0.44	0.5	0	1	0.464		
D4 Evening								
DeepSeek	25	0.52	0.5	0	1	0.467	0.768	0.450
ChatGPT	25	0.46	0.5	0	1	0.477		
D5 Evening								
DeepSeek	25	0.50	0.5	0	1	0.479	0.371	0.714
ChatGPT	25	0.48	0.5	0	1	0.467		
D6 Evening								
DeepSeek	25	0.42	0.0	0	1	0.493	−1.072	0.294
ChatGPT	25	0.50	0.5	0	1	0.456		

TABLE 3. Continued.

		n	Mean	Median	Min	Max	Std.	Paired *T* Test	
								t	*p*
D7 Evening									
	DeepSeek	25	0.56	1.0	0	1	0.486	0.681	0.503
	ChatGPT	25	0.50	0.5	0	1	0.479		
Total									
	DeepSeek	525	0.52	0.5	0	1	0.454	3.689	0.001*
	ChatGPT	525	0.44	0.5	0	1	0.465		

Max: maximum; Min: minimum; Std.: standard deviation; D: day. Values are presented as n (%). Group comparisons were performed using Paired T test.
**Level of significance: $p \leq 0.05$ is considered statistically significant.*

4. Discussion

Artificial intelligence (AI) tools and systems possess the capacity to learn, adapt to new inputs, gain experience, and emulate human intelligence to perform tasks traditionally associated with humans. Today, AI is applied across diverse fields, including healthcare, mathematics, information technology, and education. In healthcare, patients often turn to the internet for information, either due to limited accessibility to physicians or curiosity about experiences of individuals with similar medical histories. The integration of AI technologies has facilitated the workflows of medical professionals, enabling more efficient processes and supporting faster and more accurate treatment for patients [6]. However, human perception is inherently multimodal, shaped by the interaction across language, visual input, video, and sound. Recently, text-based LLMs have been augmented with additional comprehension and perceptual capabilities, giving rise to multimodal LLMs. ChatGPT and DeepSeek are widely used LLM-based conversational agents; in the present study, however, we evaluated only their text-based responses. Medicine is inherently multifaceted, and clinical decision-making often involves multimodal reasoning. Consequently, multimodal LLMs hold considerable potential in the medical domain. Nevertheless, given that clinical reasoning is typically acquired through years of training and experience, the notion that AI could fully replace clinicians warrants careful consideration [2].

Traumatic dental injuries account for approximately 5–17% of all injuries worldwide and are recognized as the fifth most common disease. In such cases, emergency intervention is critical. The knowledge of parents, who may provide first aid at the time of trauma, and the clinician's ability to manage the situation are both of paramount importance.

Parents are not always able to access dental professionals promptly regarding their children's oral health. In some cases, they may resort to online resources merely to satisfy curiosity. Consequently, accessing information through AI-driven systems has become a contemporary reality. This study focused on avulsion cases, a type of dental trauma in which immediate and accurate intervention is crucial for prognosis. The aim was both to address clinically relevant questions and to evaluate multimodal LLM-based conversational agents, specifically ChatGPT and DeepSeek, for their potential use in clinical practice.

ChatGPT and DeepSeek were selected for this study because of their multimodal structures, which support effective management of multidisciplinary clinical scenarios, as well as their current popularity and ease of accessibility as AI-driven conversational agents [2].

Given the multifaceted nature of clinical practice and the necessity of considering multiple factors in decision-making, it was insufficient to rely solely on binary question formats. Therefore, open-ended questions were also incorporated. In previous research involving LLMs, various question formats, such as multiple-choice, open-ended, and binary (yes/no), have been employed. For instance, in a study on regenerative endodontic procedures, Ekmekçi *et al.* [2] evaluated the responses of conversational agents to 23 questions (14 open-ended and 9 yes/no), reporting an accuracy rate of 86.2% for ChatGPT-4, compared with 48% for Gemini.

In another study on dental trauma, 25 questions (all in yes/no format) were posed to conversational agents, and the responses were evaluated. The reported accuracy rates were 64% for Gemini and 51% for ChatGPT [6]. In this study, the accuracy rates for the 25 questions (13 open-ended and 12 yes/no) were 38.3% for ChatGPT and 42.7% for DeepSeek. The relatively low overall accuracy observed for both conversational agents may be attributed to the higher proportion of open-ended questions compared with previous studies. Open-ended questions inherently reflect the complexity of clinical decision-making, requiring the simultaneous consideration of multiple possibilities and factors, which likely contributed to the observed outcome. In some studies [2, 16, 17], guidelines published by international scientific associations have been used both in the preparation of questions and in the evaluation of responses, ensuring scientific accuracy. In this study, all questions were developed in accordance with the IADT guidelines.

In the literature, the acceptable threshold for diagnostic accuracy is generally reported to be above 90%, a benchmark considered critical for patient safety and clinical effectiveness [2, 18]. Responses generated by the conversational agents in this study were classified as correct, incorrect, or insufficient with reference to the IADT guidelines. DeepSeek achieved an accuracy rate of 42.7%, whereas ChatGPT achieved 38.3%. A statistically significant difference in overall accuracy performance was observed between the two programs ($p < 0.05$).

Consequently, the first null hypothesis—that “there is no significant difference between ChatGPT and DeepSeek in their ability to provide accurate information to parents and clinicians regarding dental avulsion”—was rejected.

Both platforms exhibited performance levels below the accepted diagnostic accuracy threshold, consistent with the results of Suárez *et al.* [17], who evaluated ChatGPT’s responses to clinical questions in endodontics. Similarly, Díaz-Flores García reported an accuracy rate of 37.11% for a different AI-driven conversational agent in endodontic clinical questions [16]. Other studies in the literature support these findings [6, 19–21]. Although DeepSeek produced slightly more consistent results than ChatGPT, the difference was not substantial. Based on these findings, the study’s second null hypothesis—“the accuracy of LLM-generated responses to questions concerning avulsion management would fall below the diagnostic accuracy threshold commonly accepted in clinical research”—was supported by the findings. Overall, while both AI applications demonstrated promising outcomes, their performance remained below the accuracy threshold generally accepted in diagnostic studies.

Conversely, in a study based on frequently asked questions from the public regarding fluoride, ChatGPT’s responses were reported to be adequate and comprehensive [20]. Likewise, ChatGPT provided sufficient patient education in oral and maxillofacial surgery; however, its responses to academically oriented questions were considered insufficient [21]. Given the increasing use of AI-powered applications by patients and their relatives for health information—particularly in emergency situations—these findings carry significant ethical implications. It is important to emphasize that not all responses generated by LLMs are grounded in scientific evidence, and thus they may not always be reliable or accurate. Misleading information could negatively influence clinical decision-making and adversely affect treatment prognosis [6]. Accordingly, given the generally insufficient accuracy for clinical practice, the use of LLMs should be approached with caution, especially in situations requiring sensitive medical decisions. These findings also highlight that the knowledge base and response capabilities of current AI systems remain an area open to improvement for clinical applications.

In a study conducted by Nahir *et al.* [22] in pediatric dentistry, statistically significant differences were observed in ChatGPT’s responses to academic questions, with the lowest accuracy reported in the domain of dental trauma. Given the high density of technical knowledge and case-specific complexities inherent to this area, ChatGPT was found particularly inadequate in addressing technical and case-dependent questions, such as those concerning “decoronation” [22]. In the present study, both platforms demonstrated insufficient accuracy when responding to the question concerning decoronation. In this context, Balel *et al.* [21] proposed the potential development of an academic-focused version of ChatGPT, termed “ChatGPT-Academic”, capable of generating responses supported by scientific evidence. Such initiatives underscore the importance of developing academic AI conversational agents, which could enhance the reliability of clinical decision support systems in future studies.

Question-based analysis revealed that DeepSeek produced fewer incorrect responses for open-ended questions, whereas ChatGPT generated fewer incorrect responses for yes/no questions. Considering the multifaceted nature of clinical decision-making, the accuracy of responses to open-ended questions may be of greater importance for clinical applications.

When examining the accuracy of responses to the first three questions—most likely to be asked by parents—as well as to Question 24, both platforms tended to provide more accurate answers, with the exception of the third question (open-ended). These findings are consistent with results reported in previous studies [20, 22, 23].

Several limitations of this study should be noted. First, the LLMs used were not specifically trained in dental traumatology or avulsion, which may have affected their performance. Future research could assess models explicitly trained in this domain. Second, the study focused solely on AI-generated responses and did not compare these with the knowledge levels of pediatric dentists regarding avulsion. Incorporating such comparisons would provide valuable insight into the potential of AI systems to enhance clinical accuracy and effectiveness. Third, the readability of AI-generated responses was not quantitatively evaluated using standardized readability indices. Future studies should incorporate measures, such as the Flesch–Kincaid Grade Level, to enhance the interpretability of readability findings. Finally, the absence of prompt engineering may have influenced both the quality and efficiency of AI responses.

Despite the ongoing advancements in AI models, clinical reasoning is typically acquired through years of education and experience. The notion that AI could replace clinicians requires comprehensive and critical discussion. Excessive reliance on AI, together with the potential for system errors, may lead to adverse outcomes [2, 24]. Similar limitations in accuracy have been documented in other medical disciplines, including orthopedics, gastroenterology, and bariatric surgery [3, 23, 25]. These findings suggest that the challenges observed in dental avulsion reflect a broader issue affecting AI-based clinical support systems. Consequently, although LLMs offer potential advantages, they are currently inadequate as the primary or sole source of medically relevant information and cannot replace academic references. Although the accuracy of AI-generated responses can be improved through additional training, these systems continue to exhibit notable limitations. They cannot provide personalized recommendations that account for the unique circumstances or local conditions of individual patients, further underscoring their insufficiency in clinical contexts [3]. Temporal variations in AI responses may arise from the stochastic nature of generative models, server load fluctuations, or system updates implemented by developers. Such variability can influence token probability distribution and response coherence, highlighting the current limitations of LLM-based systems in maintaining consistent accuracy, especially in clinical contexts where reproducibility is essential.

The clinical significance of this study lies in the potential of AI-based conversational agents to facilitate rapid access to information during emergencies, such as traumatic dental injuries. Accurate and timely guidance may play a decisive role in determining tooth prognosis, particularly for parents and non-professional responders. However, the current limitations of AI systems—such as inadequate clinical reasoning, lack of personalization, and inability to fully interpret context-specific clinical variables—indicate that these models should presently be regarded as complementary tools, rather than replacements for clinical expertise. Future research should focus on developing domain-specific large language models trained on comprehensive datasets in dental traumatology. Such models could improve accuracy, enable standardized responses, and integrate ethical safeguards, thereby enhancing the overall reliability of AI systems in dental practice and patient education.

Overall, this study demonstrates that while LLMs may offer useful guidance on specific questions within a defined subject area, they may also deliver inaccurate information depending on the quality and scope of their training data. Therefore, caution is warranted when integrating these systems into clinical decision-making processes.

5. Conclusions

LLMs can provide patients and parents with a certain level of essential information regarding dental avulsion. However, clinicians and pediatric dentists should not adopt these systems as primary decision-making tools. Instead, they should be regarded as complementary resources that support, rather than replace, the practitioner's clinical expertise and knowledge base. Before integrating AI-generated information into treatment planning, practitioners must critically evaluate its scientific accuracy and relevance against the current literature. Moreover, future research should focus on assessing the performance of AI models specifically trained in the field of dental trauma to determine their suitability and reliability for clinical applications.

AVAILABILITY OF DATA AND MATERIALS

The datasets generated and/or analyzed during this study are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

GKD, YPY—Conception, data collection, and statistical analysis. GKD—Manuscript writing. GKD, YPY and İY—Review and editing the manuscript. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study was conducted in accordance with the principles of the Declaration of Helsinki, and ethical approval was not required as it did not involve any materials obtained from humans or animals.

ACKNOWLEDGMENT

The authors would like to thank Iğdır University for its technical contributions to this study.

FUNDING

This research received no external funding.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. Journal of Medical Systems. 2023; 47: 1–5.

[2] Ekmekçi E, Durmazpinar PM. Evaluation of different artificial intelligence applications in responding to regenerative endodontic procedures. BMC Oral Health. 2025; 25: 1–7.

[3] Dubin JA, Bains SS, Chen Z, Hameed D, Nace J, Mont MA, *et al.* Using a Google web search analysis to assess the utility of ChatGPT in total joint arthroplasty. The Journal of Arthroplasty. 2023; 38: 1195–1202.

[4] Hayder W, Hayder WA. Highlighting DeepSeek-R1: architecture, features and future implications. International Journal of Computer Science and Mobile Computing. 2025; 14: 1–13.

[5] Ulusoy AT, Önder H, Çetin B, Kaya Ş. Knowledge of medical hospital emergency physicians about the first-aid management of traumatic tooth avulsion. International Journal of Paediatric Dentistry. 2012; 22: 211–216.

[6] Ozden I, Gokyar M, Ozden ME, Sazak Ovecoglu H. Assessment of artificial intelligence applications in responding to dental trauma. Dental Traumatology. 2024; 40: 722–729.

[7] Ghaderi F, Adl A, Ranjbar Z. Effect of a leaflet given to parents on knowledge of tooth avulsion. European Journal of Paediatric Dentistry. 2013; 14: 13–16.

[8] Al-Jundi SH. Knowledge of Jordanian mothers with regards to emergency management of dental trauma. Dental Traumatology. 2006; 22: 291–295.

[9] Al-Jame Q, Andersson L, Al-Asfour A. Kuwaiti parents' knowledge of first-aid measures of avulsion and replantation of teeth. Medical Principles and Practice. 2007; 16: 274–279.

[10] Jain A, Kulkarni P, Kumar S, Jain M. Knowledge and attitude of parents towards avulsed permanent tooth of their children and its emergency management in Bhopal city. Journal of Clinical and Diagnostic Research. 2017; 11: ZC40.

[11] Ozer S, Yilmaz EI, Bayrak S, Tunc E Sen. Parental knowledge and attitudes regarding the emergency treatment of avulsed permanent teeth. European Journal of Dentistry. 2012; 6: 370–375.

[12] Loo TJ, Gurunathan D, Somasundaram S. Knowledge and attitude of parents with regard to avulsed permanent tooth of their children and their emergency management—Chennai. Journal of Indian Society of Pedodontics and Preventive Dentistry. 2014; 32: 97–110.

[13] Hu LW, Prisco CRD, Bombana AC. Knowledge of Brazilian general dentists and endodontists about the emergency management of dento-alveolar trauma. Dental Traumatology. 2006; 22: 113–117.

[14] Kostopoulou MN, Duggal MS. A study into dentists' knowledge of the treatment of traumatic injuries to young permanent incisors. International Journal of Paediatric Dentistry. 2005; 15: 10–19.

[15] Santos MESMI, Habecost APZ, Gomes FV, Weber JBB, De Oliveira MG. Parent and caretaker knowledge about avulsion of permanent teeth. Dental Traumatology. 2009; 25: 203–208.

[16] Díaz-Flores García V, Freire Y, Tortosa M, Tejedor B, Estevez R, Suárez A. Google Gemini's performance in endodontics: a study on answer precision and reliability. Applied Sciences. 2024; 14: 6390.

[17] Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: evaluating the

consistency and accuracy of endodontic question answers. International Endodontic Journal. 2024; 57: 108–113.

[18] Umer F, Habib S. Critical analysis of artificial intelligence in endodontics: a scoping review. Journal of Endodontics. 2022; 48: 152–160.

[19] Arqub SA, Al-Moghrabi D, Allareddy V, Upadhyay M, Vaid N, Yadav S. Content analysis of AI-generated (ChatGPT) responses concerning orthodontic clear aligners. The Angle Orthodontist. 2024; 94: 263–272.

[20] Buldur M, Sezer B. Can artificial intelligence effectively respond to frequently asked questions about fluoride usage and effects? A qualitative study on ChatGPT. Fluoride. 2023; 56: 201–216.

[21] Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? Journal of Stomatology Oral and Maxillofacial Surgery. 2023; 124: 101471.

[22] Bayraktar Nahir C. Can ChatGPT be guide in pediatric dentistry? BMC Oral Health. 2025; 25: 1–8.

[23] Lahat A, Shachar E, Avidan B, Glicksberg B, Klang E. Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: are we there yet? Diagnostics. 2023; 13: 1950.

[24] Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. Journal of Dental Research. 2020; 99: 769–774.

[25] Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, *et al.* Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obesity Surgery. 2023; 33: 1790–1796.

How to cite this article: Gizem Karagöz Doğan, Yelda Polat Yavuz, İzzet Yavuz. Artificial intelligence applications in tooth avulsion: comparative accuracy of ChatGPT and DeepSeek. Journal of Clinical Pediatric Dentistry. 2026; 50(2): 199-208. doi: 10.22514/jocpd.2026.049.