Submitted: 29 November, 2025 Accepted: 20 January, 2026 Published: 03 May, 2026 DOI:10.22514/jocpd.2026.082

ORIGINAL RESEARCH

Journal of Clinical Pediatric Dentistry

Chatbots in clinical decision-making for dental trauma: from diagnosis to references

Derya Sarıoğlu[1,*], Büşra Yücetürk[1], Zehra Güner[1], Zübeyde Uçar Gündoğar[1]

[1]Department of Paediatric Dentistry, Faculty of Dentistry, Gaziantep University Gaziantep, 27310 Gaziantep, Türkiye

***Correspondence**
dsarioglu@gantep.edu.tr
(Derya Sarıoğlu)

Abstract

Background: Traumatic dental injuries are a significant public health issue, yet intelligent chatbots offering potential solutions may broaden both patients and clinicians' horizons. This study aimed to evaluate the accuracy, temporal reliability, and reference reliability of the diagnosis and treatment responses provided by artificial intelligence (AI) chatbots to created dental trauma. **Methods**: 45 dental trauma scenarios based on the International Association of Dental Traumatology (IADT) guidelines were presented to four generative large language models (LLMs): ChatGPT-4o, Claude Sonnet 3.7, Gemini Advance, and DeepSeek R1. Three questions (diagnosis, treatment, and references) were asked for each scenario, and the responses were scored using a modified Global Quality Scale (mGQS) developed by the authors, allowing quantitative comparison of the results. The obtained data were analyzed using IBM SPSS Statistics version 27.0. Differences between diagnosis and treatment scores were evaluated using the Analysis of Variance (ANOVA) test, and the temporal reliability of chatbots was evaluated using the intraclass correlation coefficient (ICC). The sources provided by the LLMs were cross-checked via Google Scholar and PubMed and classified as real or fake references. **Results**: No significant differences were found among LLMs in diagnosis and treatment scores ($p > 0.05$). In the overall evaluation, DeepSeek R1 received the highest scores, while Claude Sonnet 3.7 showed the lowest average scores. When temporal reliability was assessed, ChatGPT-4o demonstrated good, clinically acceptable temporal reliability (ICC = 0.80). In contrast, Claude Sonnet showed poor reliability, while Gemini Advance and DeepSeek R1 exhibited moderate reliability. In terms of reference reliability, the highest true source rate was observed in DeepSeek R1 (84.78%), while the lowest rate was seen in Claude Sonnet 3.7 (52.57%). **Conclusions**: Although LLMs provided partially accurate and consistent responses for simple dental trauma cases, they are not yet suitable for clinical use, particularly in terms of treatment recommendations and source reliability.

Keywords

Artificial intelligence; Chatbot; Dental trauma; Dentistry; Large language models; Pediatric dentistry

1. Introduction

Traumatic dental injuries are among the most common injuries worldwide and are a universal problem that affects quality of life [1]. These types of injuries are more common in adolescents and children [2]. If not treated urgently, they can cause irreversible damage to dental and periodontal tissues in the short and long term [3].

In such injuries, parents and patients' awareness regarding emergency approaches and the doctor's quick decision-making in diagnosis and treatment ensure the continuity of teeth and periodontal tissues and visibly improve the child's psychology [4]. Therefore, it is vital that parents and children are informed about traumatic dental injuries, and that competent dentists are available in their place of residence for emergency intervention [5]. In this context, having a clinical assistant whom doctors and parents can consult and who can approach problems objectively and quickly can speed up emergency intervention.

Large language model (LLM) software developed with today's technology shows promise in providing information to people who have difficulty reaching hospitals and ensuring that doctors make diagnoses with equal information [6]. LLMs' ability to understand natural language and solve problems as humans do make them practical and invaluable in many areas [7]. Providing people, especially in the healthcare field, with meaningful information has increased these models' use [8–10]. While a single model type dominated the infrastructure in the early development stages of these models, LLMs using

multiple models (human feedback, constitution-based, multimodal transformers, *etc.*) were later released to the market [11]. The most popular LLMs in the healthcare field in recent years are ChatGPT-4o, Claude Sonnet 3.7, Gemini Advance, and DeepSeek R1 [12].

ChatGPT-4o, a version of ChatGPT developed by OpenAI, was released in 2024. ChatGPT uses an intensive transformer model based on reinforcement learning from human feedback. Claude was created by Antropic in 2023, and Claude Sonnet 3.7 was released in 2024. Claude is a constitution-based language model that aims to increase reliability through human feedback. Gemini Advance was founded in 2023 and released its Advance model in 2024. It uses a multimodal transformation architecture to convert data such as text, images, and code into a framework. DeepSeek was created in 2023, and the DeepSeek R1 model was developed in 2025. According to current data, DeepSeek uses the most suitable model of expert mixture (Model of Expertise) and performs well in domain-specific tasks [11–13].

Today, LLMs are thought to assist healthcare professionals in emergency response, diagnosis, and treatment planning [14, 15]. They can provide guidance on public health issues such as dental trauma. While the literature indicates that researchers have evaluated LLM responses to frequently asked questions (FAQs) about dental trauma, no researchers have examined LLMs' adequacy in assisting with diagnosis and treatment of dental trauma [6, 15]. The objective of this study is to determine the reliability and consistency of diagnostic and treatment support methods identified by LLMs in different traumatic dental injuries, to investigate whether they can assist patients and clinicians in the diagnostic and treatment process, and to question the realism of the references from which this information is derived. In line with these objectives, the null hypothesis of the study is that there is no significant difference between the diagnostic and treatment scores provided by LLMs.

2. Materials and methods

2.1 Study design

Two pediatric dentists (DS, BY) designed 45 dental injury scenarios in accordance with the International Association of Dental Traumatology (IADT) guidelines [16] (**Supplementary material**). In cases where inconsistencies were identified, two additional pediatric dentists (ZG, ZUG) with 10 and 15 years of clinical experience, respectively, were consulted, and any residing discrepancies were resolved through evidence-based discussion. These steps were repeated until a single, consensus-based diagnosis and treatment follow-up guide was produced for each scenario and approved by all participants. Subsequently, the scenarios were grouped according to Andreasen's classification of traumatic dental injuries [1], and a total of 45 distinct scenarios for each trauma type were reviewed and validated by specialists.

All scenarios were submitted to four LLMs (ChatGPT-4o, Claude Sonnet 3.7, Gemini Advance, and DeepSeek R1) in May 2025, and each model was asked three questions for each scenario. These questions addressed the generation of up to three possible diagnoses (Question A—*What do you think is the possible problem with this patient? Can you tell me the three most probable answers, from the answer you find most probable to the answer you find least probable?*), treatment recommendations (Question B—*If you were a dentist, what would be your treatment? Is splinting necessary?*), and the sources supporting the provided information (Question C—*Which sources did you base this information on? Can you write your references for me in APA 7 format without explanation?*). Fig. 1 illustrates an example of the traumatic dental injury scenario and the structured diagnostic and treatment questions presented to the LLMs [16]. No predefined diagnostic categories were provided to the LLMs. Each question was asked in a new chat session to prevent memory bias, and the process was repeated at the same time on the following day. The responses were saved in a Word document (Microsoft, Redmond, Washington, USA) and converted into a survey for evaluation via Google Forms. To minimize potential bias, LLM names were anonymized and replaced with numerical codes [1–4].

A 12-year-old boy falls off his bike and hits his upper front tooth on the asphalt. The fracture is on the incisal edge of the tooth, and there is no bleeding from the tooth. The patient feels mild thermal sensitivity and complains about the appearance. One-third of the tooth breaks parallel to the incisal edge, and the tooth fragment is found on the ground.
A) *What do you think is the possible problem with this patient? Can you tell me the three most probable answers, from the answer you find most probable to the answer you find least probable?*
B) *If you were a dentist, what would be your treatment? Is splinting necessary?*
C) *Which sources did you base this information on? Can you write your references for me in APA 7 format without explanation?*

FIGURE 1. Sample question style.

2.2 Evaluation of LLM responses

The responses provided for Questions A and B in all scenarios were evaluated using the mGQS [6]. This scale employs a five-point Likert framework in which higher scores indicate better quality. A score of 1 represents very poor quality characterized by inaccurate, unclear, or insufficient information, while a score of 2 represents poor quality with limited accuracy and substantial informational gaps. A score of 3 represents moderate quality, indicating generally acceptable yet partially incomplete content. A score of 4 represents good quality with clear, relevant, and mostly comprehensive information showing only minor deficiencies. Finally, a score of 5 represents excellent quality, defined by highly accurate, comprehensive, and clearly presented information that is fully useful for clinical assessment. For both Question A (diagnosis) and Question B (treatment), the maximum attainable score was 5, corresponding to a perfectly correct response. To minimize subjectivity in quality assessment, study-specific scoring criteria were developed and applied. For Question A, scoring was based on a GQS-derived framework; 2 points were awarded if the primary diagnosis was correct, 1 point for each additional diagnosis deemed reasonable (total 2 points), and 1 point if the diagnoses were supported by appropriate justification and sufficient explanation. The total score was then recorded accordingly. For Question B, the scoring criteria were derived from the GQS framework as follows: 2 points

were awarded if the proposed treatment was based on the correct diagnosis, 1 point if radiographic examination was incorporated into the clinical assessment and antibiotics were prescribed when soft-tissue injuries warranted it, 1 point if the treatment steps were clearly described and logically ordered, and 1 point if the splinting criteria were correctly applied.

2.3 Reference reliability

All responses provided for Question C (reference lists from Day 1–Day 2) were verified through cross-referencing using Google Scholar and PubMed. Authentic and fabricated references were evaluated separately for each LLM. Verified references were coded as "real", whereas unverified or nonexistent references were coded as "fake", and all data were recorded in an Excel file (Microsoft, Redmond, WA, USA).

2.4 Temporal reliability

Reliability assessment was performed only on diagnostic (A) questions. This is because treatment recommendations are directly dependent on the diagnostic decision. Thus, the internal stability of diagnostic decisions was examined; it was deemed inappropriate to include treatment scores in the analysis due to their dependent nature derived from the diagnosis.

Temporal reliability of diagnosis scores was assessed using the intraclass correlation coefficient (ICC). A two-way mixed-effects model with absolute agreement and single measures (ICC (3,1)) was applied because the same AI models generated diagnosis scores at two time points. ICC values were interpreted as follows: <0.40 poor, 0.40–0.75 moderate, and >0.75 good reliability. Ninety-five percent confidence intervals (CI) were calculated. All statistical analyses were performed using IBM SPSS Statistics 27.0 (IBM Corp., Armonk, NY, USA), and a p-value < 0.05 was considered statistically significant.

2.5 Statistical analysis

The data obtained in this study were analyzed using the licensed IBM SPSS Statistics 27.0 (IBM Corp., Armonk, NY, USA) software package. In the first stage, the mean (Mean), standard deviation (SD), minimum (Min), and maximum (Max) values for each group were calculated. To investigate whether the variables were normally distributed, skewness and kurtosis coefficients were used. According to Tabachnik and Fidell, if the skewness and kurtosis values are between −1.50 and +1.50, the variables are considered to be normally distributed. The Levene test was used for the assumption of variance homogeneity. While examining the differences between the groups, the ANOVA test was used because the variables came from a normal distribution, and the ICC was used to examine the temporal reliability of AI LLMs.

An overview of the complete methodological workflow of the study is presented in Fig. 2.

3. Results

The statistical analysis compared the response quality, temporal reliability, reference reliability, and diagnosis and treatment scores according to trauma types for four different LLMs.

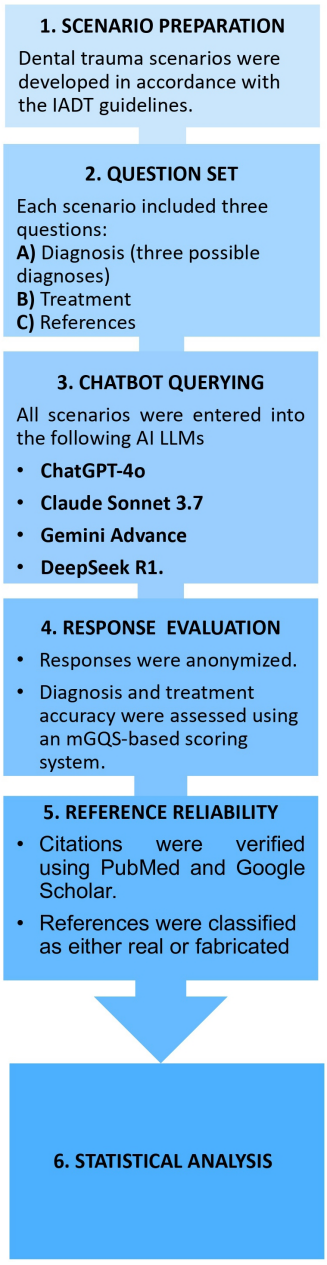

FIGURE 2. Work flow. IADT: International Association of Dental Traumatology; LLM: large language models; mGQS: modified Global Quality Scale; AI: Artificial intelligence.

3.1 Response quality

Descriptive statistics for all question categories are presented in Table 1. Overall, numerical differences were observed among the LLMs across diagnosis (Day 1–Day 2) and treatment domains. DeepSeek R1 showed higher mean scores in these domains, whereas Claude Sonnet 3.7 and ChatGPT-4o showed lower mean scores in diagnosis and treatment, respectively. However, none of these differences reached statistical significance.

3.2 Temporal reliability results

Table 2 presents temporal reliability of diagnosis scores evaluated using the ICC. A two-way mixed-effects model with absolute agreement and single measures (ICC (3,1)) was applied to

TABLE 1. Comparison of LLMs' average total scores.

	AI Model									
	ChatGPT-4o		Claude Sonnet 3.7		Gemini Advance		Deepseek R1		*F* value	*p* value
	Mean ± SD	Min–Max	Mean ± SD	Min–Max	Mean ± SD	Min–Max	Mean ± SD	Min–Max		
D1	3.78 ± 1.24	2–5	3.42 ± 1.29	2–5	3.71 ± 1.20	2–5	3.84 ± 0.82	1–5	1.170	0.323
D2	3.84 ± 1.02	3–5	3.64 ± 1.11	2–5	3.82 ± 1.07	1–5	4.02 ± 0.89	2–5	1.018	0.386
T	3.29 ± 1.60	0–5	3.53 ± 1.67	0–5	3.33 ± 1.67	0–5	3.60 ± 1.40	0–5	0.359	0.783
Total	3.64		3.53		3.62		3.82			

ANOVA test was applied.
D1: diagnostic responses on Day 1; D2: diagnostic responses on Day 2; T: treatment responses on Day 1.
SD: standard deviation; Min: minimum; Max: maximum; AI: artificial intelligence.

TABLE 2. Intraclass correlation coefficient (ICC) for Day 1 and Day 2 diagnosis scores.

AI Model	ICC (3,1)	95% Confidence Interval
ChatGPT-4o	0.80	0.65–0.91
Claude Sonnet	0.26	−0.10–0.56
Gemini Advance	0.49	0.20–0.71
DeepSeek R1	0.44	0.21–0.62

ICC values were calculated using a two-way mixed-effects model with absolute agreement and single measures (ICC (3,1)) to assess temporal reliability between Day 1 and Day 2 diagnosis scores. AI: Artificial intelligence.

compare Day 1 and Day 2 scores for each AI model. ChatGPT-4o demonstrated good temporal reliability (ICC = 0.80; 95% CI: 0.65–0.91). In contrast, Claude Sonnet showed poor agreement (ICC = 0.26; 95% CI: −0.10–0.56), while Gemini Advance (ICC = 0.49; 95% CI: 0.20–0.71) and DeepSeek R1 (ICC = 0.44; 95% CI: 0.21–0.62) exhibited moderate reliability. Only ChatGPT-4o reached the threshold generally considered acceptable for clinical reliability.

3.3 Reference reliability results

As shown in Table 3, reference reliability varies substantially across chatbots, indicating differences in their citation-generation strategies. DeepSeek R1 achieved the highest reliability rate (84.78%), followed closely by Gemini Advance (80.64%) and ChatGPT-4o (78.04%), suggesting that these models are more effective in producing verifiable references. Notably, Gemini Advance reached a high reliability level despite generating fewer total references, implying a more conservative and accuracy-oriented approach. In contrast, Claude Sonnet 3.7 demonstrated considerably lower reliability (52.57%), with nearly half of its references classified as fake, highlighting a greater tendency toward reference hallucination. Overall, the findings suggest that higher reference reliability is associated not with the quantity of references produced, but with more controlled and selective citation practices.

3.4 Diagnosis and treatment scores according to trauma types

Table 4 shows the diagnosis and treatment scores obtained from AI LLMs according to question types. Based on mean scores, numerically higher scores were observed for enamel fracture, subluxation, and alveolar fracture, whereas lower mean scores were observed for uncomplicated root fracture, lateral luxation, and concussion.

4. Discussion

Traumatic dental injury is one of the major health problems that burdens public health both financially and emotionally [2]. The prognosis of a traumatized tooth is closely related to the level of knowledge possessed by parents, patients, and dentists, which is particularly critical in pediatric cases. Limited first aid knowledge in environments where access to adequate healthcare services is restricted adversely affects oral and dental health, ultimately compromising overall dental integrity. When trauma results in an unfavorable prognosis in a child, the psychological impact can be profound for both the child and their parents [4]. Moreover, even when healthcare services are readily available, insufficient clinical experience in dental trauma may lead the dentist to make incorrect diagnostic or treatment decisions, exacerbating the condition. Therefore, an LLM-based AI system accessible to both clinicians and patients and well informed about traumatic dental injuries may provide appropriate first aid guidance for families and serve as a valuable clinical support tool and as a diagnostic aid for practitioners [8].

Researchers have measured the level of knowledge of AI LLMs on the subject of dental trauma [6, 15, 17–19]. Their studies are accompanied by the answers generated by different LLM software to different types of questions. Ozden *et al.* [15] presented 25 yes–no questions regarding general aspects of dental trauma to ChatGPT-4o and Google Bard (Gemini), thereby evaluating the LLMs' ability to provide

TABLE 3. Reference reliability scores.

AI Model	Total Reference	Category	n	%
ChatGPT-4o	186 (RR: 145)	Real	32	78.04
		Fake	9	21.95
Claude Sonnet 3.7	164 (RR: 67)	Real	51	52.57
		Fake	46	47.42
Gemini Advance	72 (RR: 41)	Real	25	80.64
		Fake	6	19.35
DeepSeek R1	187 (RR: 141)	Real	39	84.78
		Fake	7	15.21

RR: Recurring reference; AI: Artificial intelligence.

TABLE 4. Mean scores of AI models by question type across diagnostic (Day 1 and Day 2) and treatment responses.

Question Type	ChatGPT-4o Mean (D1/D2/T)	Claude Sonnet 3.7 Mean (D1/D2/T)	Gemini Advance Mean (D1/D2/T)	Deepseek R1 Mean (D1/D2/T)
Enamel crack	3.67/4.67/4.33	2.67/5.00/2.67	3.33/3.00/3.67	3.67/4.00/4.67
Enamel fracture	4.67/4.67/4.67	4.00/3.67/5.00	5.00/4.67/5.00	4.00/4.67/5.00
Enamel-dentin fracture	4.33/4.00/4.00	2.67/3.33/2.67	4.67/4.67/4.67	3.33/3.33/2.67
Complex enamel-dentin fracture	4.00/3.33/4.00	4.33/3.33/4.33	4.33/4.33/3.67	3.67/3.67/5.00
Complex crown-root fracture	4.67/4.00/4.00	4.33/3.33/4.33	4.67/4.00/3.00	4.00/4.00/4.00
Root fracture	4.25/4.00/3.25	3.25/3.50/3.00	3.50/3.00/4.00	4.00/4.00/4.00
Subluxation	4.50/4.50/5.00	4.00/3.50/5.00	5.00/4.50/5.00	4.00/3.25/3.00
Extrusion	3.00/3.00/4.00	3.67/4.67/4.67	2.67/4.67/1.67	3.00/4.00/3.67
Lateral luxation	2.00/2.67/1.00	2.67/4.00/3.33	2.33/4.00/1.67	3.33/3.67/1.00
Intrusion	4.33/4.33/3.00	3.00/3.00/3.67	3.67/3.00/2.00	4.67/4.67/3.33
Avulsion	4.25/4.00/3.25	4.00/3.50/3.50	4.00/3.50/4.25	3.25/3.25/3.75
Alveolar fracture	4.50/4.50/4.50	5.00/5.00/4.50	4.00/5.00/2.00	4.50/5.00/3.50
Mixed trauma problems	4.00/4.20/3.20	4.00/4.00/4.00	3.40/4.00/3.20	3.80/4.22/3.40
Uncomplicated crown-root fracture	1.00/2.00/0.00	1.00/2.00/0.00	2.00/2.00/2.67	3.67/3.67/5.00
Concussion	3.00/3.00/0.00	2.00/2.00/0.00	3.40/4.00/3.20	3.80/4.20/3.40

D1: diagnostic responses on Day 1; D2: diagnostic responses on Day 2; T: treatment responses on Day 1.

accurate and reliable responses. Taraç *et al.* [17] asked LLMs 31 binary questions. Mustuloğlu *et al.* [18] asked LLMs 18 true–false questions related to avulsion. Johnson *et al.* [19] asked LLMs 20 FAQs related to dental trauma. Guven *et al.* [6] asked LLMs 59 FAQs and evaluated the answers. In a study on parents' concerns about dental trauma, Taraç *et al.* [17] submitted frequently asked parental questions to LLMs to evaluate their responses. In addition to these studies, researchers have also used multiple-choice, fill-in-the-blank, and binary questions together [20, 21]. In contrast to previous studies in the literature, the present study did not focus on general trauma-related questions. Instead, the diagnostic responses generated by LLMs for trauma scenarios constructed in accordance with the IADT guidelines were evaluated [16]. Additionally, by requesting three possible diagnoses for each scenario, the we aimed to expand the LLMs' clinical reasoning capacity. Treatment recommendations were also assessed. Although some researchers have examined the reference reliability of LLMs, the present research represents the first reference-reliability analysis conducted specifically within the context of dental trauma [22].

To ensure the objective evaluation of data obtained across studies, various assessment scales have been employed in the relevant literature. The tools used to assess LLM performance in dental trauma research vary considerably. Some researchers have utilized categorical coding approaches due to the presence of binary, multiple-choice, or fill-in-the-blank question formats [15, 18, 21, 22]. Others have evaluated the quality and accuracy of LLM outputs using established instruments, such as DISCERN and the GQS, which are widely applied

in similar contexts [17, 19, 23]. In studies involving open-ended questions, Likert-type scales derived from the GQS have frequently been preferred [19].

In accordance with this methodological background, the GQS was selected in the present study because it provides a practical, reliable, and comprehensive framework for evaluating the overall quality of LLM responses. The success rates of LLMs have varied in studies in the literature.

The literature demonstrates substantial variability in LLM performance, which can be attributed to differences in the models evaluated, the thematic scope of the studies, the types of questions posed, and the number and structure of the assessment items. For example, Ozden *et al.* [15] reported an accuracy rate of 57.5% when evaluating ChatGPT-4o and Google Bard, with reliability scores of 0.266 and 0.419, respectively. Taraç *et al.* [17], using 31 binary questions, found that Bing achieved the highest accuracy rate (96.34%), whereas Claude Sonnet 3.7 demonstrated the lowest (88.17%). Kuru *et al.* [20] examined five LLMs using 30 mixed-format questions and observed descending success rates for ChatGPT-3.5, Copilot Pro, Copilot Free, ChatGPT-4o, and Google Gemini. Johnson *et al.* [19] assessed the validity and reliability of four LLMs and identified Claude AI as the highest-performing model. In contrast, Sezer *et al.* [12] evaluated four LLMs (ChatGPT-4o, DeepSeek R1, Gemini Advance, and Claude Sonnet 3.7) using 25 open-ended questions related to primary teeth and found ChatGPT-4o to have the highest accuracy rate, although the differences were not statistically significant; DeepSeek R1 and Claude Sonnet 3.7 were reported as more reliable than Gemini Advance [12].

In the present study, numerical variability in diagnostic performance was observed across LLMs. While DeepSeek R1 showed higher mean diagnostic scores in the current analysis, these findings differed from those of Sezer *et al.* [12], who reported higher accuracy for ChatGPT-4o. The differences observed between results may be explained by the rapid and continuous evolution of LLMs, which can lead to substantial performance variations even within the same LLM family. Furthermore, similar to Taraç *et al.*'s [17] results, Claude showed the lowest diagnostic accuracy, a finding that contrasts with the conclusions of Johnson *et al.* [19], who identified Claude AI as the most reliable model. Gemini, consistent with most studies except that of Taraç *et al.* [17], generally performed at a comparatively lower level across diagnostic evaluations. Although no statistically significant differences were found between the LLMs in this study, certain numerical trends—specifically, higher scores for DeepSeek R1 and lower scores for Claude Sonnet 3.7—are noteworthy. While these trends did not reach statistical significance, they may still have clinically meaningful implications. Accordingly, the null hypothesis (H_0) was accepted and the alternative hypothesis (H_1) rejected.

DeepSeek R1 generally succeeded in identifying the correct diagnosis; however, it was unable to provide three differential diagnoses as requested. In contrast, other LLMs—particularly ChatGPT-4o—were able to generate three possibilities by employing different noun phrases with equivalent meanings. For instance, while the first diagnosis was stated as a complex crown fracture, the second was expressed as an enamel-dentin fracture with pulp exposure. This tendency to reformulate diagnostic terminology may partially explain ChatGPT-4o's higher diagnostic score compared with Claude Sonnet 3.7 and Gemini Advance. When prompted to list three diagnostic possibilities, Claude Sonnet 3.7 typically presented its predictions clearly, but did not provide the underlying rationale. Consequently, its score decreased in cases where explanatory reasoning was required. However, unlike ChatGPT-4o, it tended to avoid producing alternative labels for the same diagnosis. A further limitation observed in Claude Sonnet 3.7 was its occasional substitution of symptoms for diagnoses in its secondary and tertiary predictions (*e.g.*, providing "aesthetic concerns" as a diagnosis).

The scenarios presented in this study were formulated using terminology appropriate for dental professionals. Accordingly, ChatGPT-4o, DeepSeek R1, and Claude Sonnet 3.7 interpreted the user as a clinician and responded in a professional register. Gemini Advance, however, consistently interpreted the user as a patient and produced responses using more lay-appropriate terminology. Distinct from the other LLMs, Gemini Advance also incorporated elements of psychological reassurance in its answers—a strategy that may be beneficial in reducing parental anxiety. This finding aligns with the observations reported by Taraç *et al.* [17] regarding LLM use among parents concerned about dental trauma.

There are currently no studies in the literature aside from providing basic oral-health guidance; no researchers have assessed the capacity of LLMs to support or formulate dental trauma treatment plans [24]. When treatment quality was evaluated, numerical trends in mean scores were observed, with DeepSeek R1 showing higher mean treatment scores, followed by Claude Sonnet 3.7, Gemini Advance, and ChatGPT-4o. DeepSeek R1 generally produced coherent and clinically appropriate treatment suggestions. Claude Sonnet 3.7 tended to articulate its management strategies in a sequential manner across the three proposed diagnostic possibilities, which contributed to its relatively high treatment-quality score. In contrast, ChatGPT-4o typically provided a detailed explanation only for the primary diagnosis and did not elaborate on management options for the alternative differential diagnoses.

Another noteworthy finding of this study was the tendency of LLMs to generate terminology that is not included in established classifications of traumatic dental injuries and is absent from the scientific literature. Examples of such unsupported terms include pulp contusion–concussion, partial avulsion, crown contusion, tooth infusion, and fixed subluxation. The use of diagnostic terminology that is not part of internationally accepted classifications may lead to confusion, particularly within referral chains and multidisciplinary clinical settings. Furthermore, misinterpretation of the diagnosis may result in inappropriate or delayed treatment.

When reference reliability was evaluated, DeepSeek R1 demonstrated the highest reliability in reference generation, with 84.78% of its citations corresponding to real sources, whereas Claude Sonnet 3.7 exhibited the lowest reference reliability among the evaluated LLMs. In contrast, Gemini Advance provided the fewest references and frequently avoided citing sources by stating that its responses were based on "clinical experience". Reference fabrication most commonly

involved incorrect combinations of existing bibliographic elements or the inclusion of nonacademic sources.

In clinical practice, reliance on non-verifiable sources may lead to misinformation and contribute to the normalization of inappropriate clinical practices, raising serious ethical concerns related to accountability and responsibility. These findings highlight the need for retrieval-augmented generation approaches integrated with authoritative medical databases in the clinical application of LLMs.

Taken together, the generation of nonexistent terminology and the widespread use of inaccurate or fabricated references highlight critical weaknesses in current AI LLMs. These findings underscore the need for further development to improve the factual accuracy, reference integrity, and terminological reliability of AI systems intended for use in dental trauma guidance.

5. Limitations of the study

The present study approached treatment evaluation in a diagnosis-dependent manner. From this perspective, the treatment of misdiagnosis received a low score. The fundamental reason is that, in clinical practice, a "correct" treatment based on an "incorrect" diagnosis is still regarded as clinically inappropriate for the patients. Especially in emergencies requiring careful management, such as dental trauma, there is no meaningful equivalent to a correct treatment plan based on an incorrect diagnosis. However, future researchers that evaluate treatment content as an independent factor, separate from diagnosis, may help determine whether LLMs can support the treatment process in clinical settings where a definitive diagnosis can be established.

At the same time, objectively evaluating treatment stages in this study was challenging because treatment decisions in real-life clinical practice are shaped by the clinician's judgment. These decisions may vary depending on factors such as the level of the child and parents' cooperation and of oral hygiene's adequacy. Because these variables were excluded and only the IADT guidelines were followed, the responses may not fully reflect real-life clinical scenarios.

Treatment scores were derived from diagnostic decisions, rather than from an independent treatment decision-making process; therefore, temporal reliability was assessed solely based on diagnosis. Furthermore, due to the large number of questions and LLM evaluated, all questions were administered within a two-day period. Two days represent the minimum duration required for temporal reliability assessment [25].

Although no statistically significant differences were found between the LLMs in this study, certain numerical trends were observed. These trends may be related to differences in the LLMs' underlying model architectures. The lack of reported effect sizes and power analyses limits the clinical interpretation of these numerical differences. Future studies incorporating these analyses may provide a more robust understanding of such trends' potential clinical relevance.

6. Conclusions

LLMs' use of nonstandard terminology, the generation of unreliable or fabricated references, potential disruptions in the diagnosis–treatment continuum, and underlying algorithms' limited transparency pose significant ethical risks related to accountability and clinical responsibility. Therefore, until LLMs are capable of consistently adhering to international clinical guidelines and ensuring verifiable source reliability, they should be regarded solely as supportive tools, rather than clinical decision-makers. In the future, AI-based applications that successfully address these limitations may enable dentists worldwide to access standardized, professional, and evidence-based information on a global scale.

AVAILABILITY OF DATA AND MATERIALS

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

AUTHOR CONTRIBUTIONS

DS—prepared the main draft of the study; did the main writing of the article. DS and BY—worked together to create the data. ZG and ZUG—evaluated and written statistical analysis; performed the final review of the manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Since the study did not use human data, ethical approval was not required.

ACKNOWLEDGMENT

The authors would like to thank Hasan Gündoğar and Ömer Faruk Kaygısız for their contributions to this study.

FUNDING

This research received no external funding.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY MATERIAL

Supplementary material associated with this article can be found, in the online version, at https://oss.jocpd.com/files/article/2049764474418544640/attachment/Supplementary%20material.docx.

REFERENCES

[1] Andreasen JO, Andreasen FM, Andersson L. Textbook and color atlas of traumatic injuries to the teeth. 5th edn. John Wiley & Sons: Munksgaard. 2018.

[2] Petti S, Glendor U, Andersson L. World traumatic dental injury prevalence and incidence: a meta-analysis. Dental Traumatology. 2018; 34: 71–86.

[3] Lin S, Pilosof N, Karawani M, Wigler R, Kaufman AY, Teich ST. Occurrence and timing of complications following traumatic dental injuries. Journal of Clinical and Experimental Dentistry. 2016; 8: e429–e436.

[4] Lee JY, Divaris K. Hidden consequences of dental trauma: the social and psychological effects. Pediatric Dentistry. 2009; 31: 96–101.

[5] Antipovienė A, Narbutaitė J, Virtanen JI. Traumatic dental injuries, treatment, and complications in children and adolescents. European Journal of Dentistry. 2021; 15: 557–562.

[6] Guven Y, Ozdemir OT, Kavan MY. Performance of artificial intelligence chatbots in responding to patient queries related to traumatic dental injuries: a comparative study. Dental Traumatology. 2024; 41: 338–347.

[7] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, *et al*. Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology. 2017; 2: 230–243.

[8] Lee H. The rise of ChatGPT: exploring its potential in medical education. Anatomical Sciences Education. 2024; 17: 926–931.

[9] Zhang J, Zhang Z. Ethics and governance of trustworthy medical artificial intelligence. BMC Medical Informatics and Decision Making. 2023; 23: 7.

[10] Zielinski C, Winker MA, Aggarwal R, Ferris LE, Heinemann M, Lapeña JF Jr, *et al*. Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications. Colombia Médica. 2023; 54: e1015868.

[11] Marr B. A short history of ChatGPT: how we got to where we are today. 2023. Available at: https://www.forbes.com/ (Accessed: 20 May 2025).

[12] Sezer B, Aydoğdu T. Performance of advanced artificial intelligence models in traumatic dental injuries in primary dentition. Applied Sciences. 2025; 15: 7778.

[13] Anthropic. Claude 3 model family. 2024. Available at: https://www.anthropic.com/ (Accessed: 20 May 2025).

[14] Shan T, Tay FR, Gu L. Application of artificial intelligence in dentistry. Journal of Dental Research. 2021; 100: 232–244.

[15] 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.

[16] Levin L, Day PF, Hicks L, O'Connell A, Fouad AF, Bourguignon C, *et al*. International association of dental traumatology guidelines for the management of traumatic dental injuries: general introduction. Dental Traumatology. 2020; 36: 309–313.

[17] Taraç MG. Evaluation of artificial intelligence chatbots in the management of primary tooth traumas: a comparative analysis. Journal of International Dental Sciences. 2025; 11: 22–31.

[18] Mustuloğlu Ş, Deniz BP. Evaluation of chatbots in the emergency management of avulsion injuries. Dental Traumatology. 2025; 41: 437–444.

[19] Johnson AJ, Singh TK, Gupta A, Sankar H, Gill I, Shalini M, *et al.* Evaluation of validity and reliability of AI chatbots as public sources of information on dental trauma. Dental Traumatology. 2025; 41: 187–193.

[20] Kuru HE, Aşık A, Demir DM. Can artificial intelligence language models effectively address dental trauma questions? Dental Traumatology. 2025; 41: 567–580.

[21] Tokgöz Kaplan T, Cankar M. Evidence-based potential of generative artificial intelligence large language models on dental avulsion: ChatGPT versus Gemini. Dental Traumatology. 2025; 41: 178–186.

[22] Kaygisiz ÖF, Teke MT. Can DeepSeek and ChatGPT be used in the diagnosis of oral pathologies? BMC Oral Health. 2025; 25: 638.

[23] Öztürk Z, Bal C, Çelikkaya BN. Evaluation of information provided by ChatGPT versions on traumatic dental injuries for dental students and professionals. Dental Traumatology. 2025; 41: 427–436.

[24] Moeini A, Torabi S. The role of artificial intelligence in dental diagnosis and treatment planning. Journal of Oral and Dental Health Nexus. 2025; 2: 14–26.

[25] Suárez A, Jiménez J, Llorente de Pedro M, Andreu-Vázquez C, Díaz-Flores García V, Gómez Sánchez M, *et al*. Beyond the scalpel: assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Computational and Structural Biotechnology Journal. 2024; 24: 46–52.

How to cite this article: Derya Sarıoğlu, Büşra Yücetürk, Zehra Güner, Zübeyde Uçar Gündoğar. Chatbots in clinical decision-making for dental trauma: from diagnosis to references. Journal of Clinical Pediatric Dentistry. 2026; 50(3): 274-281. doi: 10.22514/jocpd.2026.082.