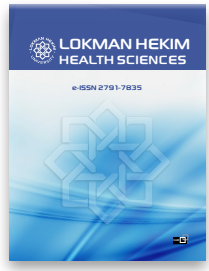
LOKMAN HEKIM HEALTH SCIENCES

DOI: 10.14744/lhhs.2026.38881
Lokman Hekim Health Sci 2026;6(2):255–263



lokmanhekimhs.com

ORIGINAL ARTICLE

Evaluation of AI Chatbots in Tooth Avulsion Management According to the International Association of Dental Traumatology Guidelines

Merve Özdemir[1], Esra Yıldırım Manav[2]

[1]Department of Pediatric Dentistry, Faculty of Dentistry, Lokman Hekim University, Ankara, Türkiye
[2]Department of Restorative Dentistry, Faculty of Dentistry, Lokman Hekim University, Ankara, Türkiye

Abstract

Introduction: This study aimed to evaluate the extent to which widely used artificial intelligence (AI)-based chatbots adhere to the 2020 International Association of Dental Traumatology (IADT) guidelines for the management of tooth avulsion and to assess the accuracy of the bibliographic references (i.e., complete citation details including title, authors, journal, year, and DOI) they generate.

Methods: This cross-sectional observational study assessed four AI-based chatbots (ChatGPT-5.2, Perplexity AI, Gemini 2.5 Flash, and DeepSeek-v3.2) using ten standardized, clinician-directed avulsion scenarios aligned with the 2020 IADT guidelines. Each scenario was submitted once per chatbot, without iterative prompting, on 3 January 2026. Scenarios varied by extra-oral dry time, storage medium, apex maturity, dentition type, and replantation timing. Responses were evaluated using the 9-item IADT Compliance Index. Bibliographic accuracy was assessed using the reference hallucination score (RHS).

Results: No statistically significant difference was observed in overall normalized compliance scores among the chatbots ($p=0.089$). However, significant between-model differences emerged in technically critical domains, including root surface cleaning ($p=0.017$), and splint type and duration ($p<0.001$). ChatGPT-5.2 and Perplexity AI consistently outperformed Gemini 2.5 Flash and DeepSeek-v3.2. Although RHS values did not differ significantly between models ($p=0.114$), all chatbots demonstrated occasional reference hallucinations.

Discussion and Conclusion: Performance was higher in simpler scenarios, such as immediate replantation, whereas more complex conditions – particularly prolonged dry time and primary tooth avulsion – showed lower compliance and greater variability. Although chatbots reproduce general principles, limitations restrict reliability; thus, they should be used with clinician supervision.

Keywords: Artificial intelligence; Dental trauma; International Association of Dental Traumatology guidelines; Tooth avulsion

Dental avulsion is one of the most severe and time-critical traumatic dental injuries (TDIs), accounting for up to 16% of dental trauma cases in children and adolescents.[1] The prognosis of an avulsed permanent tooth is determined primarily by the duration of extra-oral dry time, the condition of the periodontal ligament (PDL), and the appropriateness

Cite this article as: *Özdemir M, Yıldırım Manav E. Evaluation of AI Chatbots in Tooth Avulsion Management According to the International Association of Dental Traumatology Guidelines. Lokman Hekim Health Sci 2026;6(2):255–263.*

Correspondence: Merve Ozdemir, M.D. Lokman Hekim Üniversitesi, Diş Hekimliği Fakültesi, Pediatrik Diş Hekimliği Bölümü, Ankara, Türkiye
E-mail: merveozdemir@lokmanhekim.edu.tr **Submitted:** 16.02.2026 **Revised:** 08.05.2026 **Accepted:** 18.05.2026 **Available Online:** 02.06.2026

of immediate management steps performed at the scene of injury or in the emergency clinic.[2,3] Even minor deviations from established protocols, such as using an inappropriate storage medium or delaying replantation, can lead to ankylosis-related replacement resorption and eventual tooth loss. Furthermore, errors in other key management steps – such as inappropriate replantation decisions, incorrect splint duration, or suboptimal antibiotic recommendations – may further compromise periodontal healing, increase the risk of infection or inflammatory root resorption, and negatively affect long-term tooth survival.[4]

To ensure standardized, evidence-based management, the International Association of Dental Traumatology (IADT) published comprehensive guidelines in 2020 for both permanent and primary dentitions.[5,6] Management is influenced by factors such as apex maturity, extent of PDL damage, and appropriate follow-up for detecting complications. Despite the availability of these guidelines, adherence in clinical practice remains inconsistent. Studies indicate that many general practitioners and emergency physicians are unfamiliar with less common scenarios, such as open-apex avulsion, delayed replantation, or management of contaminated root surfaces.[7,8] In addition, real-time emergencies may limit access to clinical protocols, specialist consultation, or institutional resources, particularly in resource-limited settings. Challenges such as determining appropriate splinting duration, assessing tetanus status, and distinguishing between permanent and primary dentition further increase the risk of error.

In recent years, artificial intelligence (AI) chatbots have emerged as tools to address this accessibility gap. With retrieval-augmented generation (RAG) and access to academic databases, they can provide real-time, guideline-based recommendations and bibliographic references. Surveys suggest that more than half of clinicians use AI for decision support, particularly in time-sensitive trauma settings.[9–11] Although IADT guidelines are publicly available, accessing them during emergencies is often impractical. In contrast, AI chatbots offer rapid, on-demand responses, making them attractive for acute decision support.[9,10] However, their outputs remain unregulated and variable, and it is unclear whether they provide consistent recommendations across diverse avulsion scenarios, including variations in dry time, storage conditions, apex maturity, dentition type, and follow-up requirements.

One emerging concern is reference hallucination, in which AI systems generate citations that appear valid but do not correspond to real publications. Recent evaluations have shown hallucination rates ranging from 30% to 60% in medical and dental AI-generated reference lists, posing risks for misinformation, incorrect clinical decision-making, and compromised academic integrity.[12–14] Ensuring accurate citation of the 2020 IADT guidelines is especially important because management differs substantially depending on factors such as extra-oral time thresholds (≤30 min vs. ≥90 min), apex maturity, and appropriate follow-up intervals.

Recent studies have demonstrated that AI-based chatbots are increasingly being applied across various dental domains, including diagnosis, treatment planning, and patient education.[15–17] Previous studies have shown that AI-based systems can generally reproduce broad management principles for TDIs, including avulsion.[18–20] However, these studies have primarily focused on overall accuracy rather than compliance with guideline-sensitive, time-dependent clinical steps. Key aspects emphasized in the 2020 IADT guidelines – such as root surface management after prolonged dry time, socket preparation, and splinting protocols – have not been systematically evaluated across diverse, clinically realistic scenarios.

In addition, although some AI tools provide bibliographic references, the accuracy of these citations, particularly regarding the 2020 IADT avulsion guidelines, remains largely unexplored.

This study therefore aimed to evaluate the ability of AI chatbots to follow IADT-recommended avulsion management by applying a standardized prompt encompassing all essential guideline components across diverse clinical scenarios. The null hypothesis was that there would be no significant difference among the evaluated AI chatbots in terms of their compliance with the 2020 IADT avulsion guidelines and the accuracy of their cited references.

Materials and Methods

Study Design

This cross-sectional observational study evaluated the adherence of four AI-based chatbots to the 2020 IADT guidelines when responding to clinician-directed avulsion management questions. No human subjects or biological samples were involved; therefore, ethics approval was not required.

The study assessed four widely used English-language AI-based chatbots equipped with real-time literature retrieval systems:

- ChatGPT-5.2 (OpenAI, USA)
- Perplexity AI (Perplexity Inc., USA)
- Gemini 2.5 Flash (Google DeepMind, UK)
- DeepSeek-v3.2 (DeepSeek Lab, China).

Question Preparation Process

A total of 10 clinically realistic avulsion scenarios were created in alignment with the IADT Guidelines by two dental traumatology specialists with at least 5 years of clinical experience (Supplementary Material 1).

To reflect age-dependent differences in root development, open-apex scenarios were designed to represent 7-year-old children, whereas closed-apex scenarios represented 12-year-old patients. The scenarios incorporated variations in extra-oral dry time (30 vs. 90 min), storage medium (milk for 30 or 90 min), apex maturity (open vs. closed), primary tooth avulsion, and immediate replantation at the accident site.

The final scenario list consisted of:

1. Closed apex – 30 min dry (12 years)
2. Open apex – 30 min dry (7 years)
3. Closed apex – 90 min dry (12 years)
4. Open apex – 90 min dry (7 years)
5. Closed apex – milk 30 min (12 years)
6. Open apex – milk 30 min (7 years)
7. Closed apex – milk 90 min (12 years)
8. Open apex – milk 90 min (7 years)
9. Avulsed primary maxillary incisor
10. Immediately replanted the tooth at the accident site (12 years).

Following each clinical scenario, all chatbots were consistently asked the standardized fixed question: "How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, web link, and DOI." This approach was used to ensure methodological standardization and comparability across all scenarios.

Question Redirection Process

To minimize potential biases and eliminate memory effects, each clinical scenario was presented in a separate chat session. Thus, ten independent sessions were created for each chatbot, with one scenario asked per session. Each scenario and prompt pair was submitted individually to each chatbot. Before initiating each session, the chat history was cleared, the "delete our chat history" command was executed when applicable, and any remaining messages were manually removed. In addition, the browser cache, cookies, and browsing history were cleared before each session to prevent any carryover effects.

All chatbot interactions were performed using publicly accessible web-based user interfaces (ChatGPT through the OpenAI platform, Perplexity AI through perplexity.ai, Gemini through gemini.google.com, and DeepSeek through chat.deepseek.com), rather than application programming interfaces, under free-tier user access conditions. Model versions were identified based on the labels displayed in the user interface at the time of access. Exact backend version identifiers are not publicly disclosed by the providers; therefore, model identification relied on publicly visible version information. No custom system prompts, developer instructions, or role-based priming were applied. Default system settings were used for all models, and no adjustable generation parameters (e.g., temperature, top-p sampling, or maximum token limits) were modified. Real-time literature retrieval or browsing features (i.e., RAG), when available, were not manually activated, deactivated, or modified, and all models were evaluated under their default system configurations. No plugins, developer modes, or auxiliary tools were activated during data collection. Each prompt was submitted once per scenario, and no follow-up questions or iterative refinements were performed. All scenarios were submitted on the same day (January 03, 2026) to minimize temporal variability. All responses were exported as plain text, anonymized, and coded for evaluation. Each chatbot received the same ten prompts in identical order and formatting, and all prompts were submitted verbatim without modification. No cross-session information sharing occurred between queries. All interactions were carried out by a single experienced investigator (MO) to ensure standardization and consistency.

Evaluation Process

Although the chatbots were only asked a single clinician-directed question per scenario, their responses were independently evaluated using a predefined 9-item IADT Compliance Index derived from the 2020 IADT guidelines. The evaluation focused on the presence and accuracy of the following management components: (1) Indication for replantation, (2) root surface cleaning, (3) socket preparation, (4) endodontic timing and intracanal medicaments, (5) systemic antibiotic therapy, (6) tetanus prophylaxis, (7) type and duration of splinting, (8) post-operative patient instructions, and (9) clinical and radiographic follow-up schedule. Items not applicable to a given scenario (e.g., replantation for primary teeth) were excluded from the denominator. A compliance percentage was calculated for each response.

Reference Verification (Reference Hallucination Score [RHS] Framework)

The RHS is a structured framework developed to systematically quantify the severity of hallucinated citations in AI-generated content and was applied in the present study to evaluate each bibliographic reference generated by the chatbots. The term "bibliographic reference" refers to the complete citation details, including title, authors, journal name, year of publication, volume, issue, page numbers, web link, and DOI. Each reference was systematically evaluated across seven predefined bibliographic identifiers, including the accuracy of the title, authors' names, journal name, publication year, digital object identifier (DOI), and web link (URL), as well as its relevance to the dental trauma topic.[12] Each major hallucination (e.g., incorrect or missing title, author list, journal, or DOI) received 2 points, while minor hallucinations (e.g., wrong year, invalid link, or irrelevant topic) received 1 point. Thus, the total RHS per reference ranged from 0 (fully accurate) to 11 (completely hallucinated). In addition, all clinical response components defined within the 9-item IADT Compliance Index, as well as all bibliographic references generated by the chatbots, were independently evaluated by the same two calibrated dental traumatology experts, who were blinded to chatbot identities throughout the assessment process. Clinical items were scored dichotomously as correctly addressed or incorrect/missing, while references were assessed according to predefined RHS criteria. In cases of disagreement, responses and references were re-evaluated, and consensus was achieved through discussion. Inter-rater reliability was assessed using the intraclass correlation coefficient (ICC), demonstrating excellent agreement for both clinical scoring (ICC=0.93) and RHS assessment (ICC=0.89).

Statistical Analysis

Statistical analyses were performed using IBM Statistical Package for the Social Sciences Statistics (version 27; IBM Corp., Armonk, NY, USA). Descriptive statistics were calculated for all variables. Total clinical adherence scores were derived from a 9-item IADT-based index and converted to percentage values (0–100). Given that all chatbot models were evaluated using identical clinical scenarios, the data were treated as repeated measures. The Shapiro–Wilk test indicated non-normal distribution of total scores. Accordingly, non-parametric tests appropriate for repeated-measures designs were applied. Comparisons of total clinical adherence scores across chatbot models were performed using the Friedman test. For individual clinical criteria, responses were dichotomized as correct or incorrect and expressed as percentages. Between-model comparisons for these binary variables were conducted using Cochran's Q test. For criteria demonstrating statistically significant overall differences, post hoc pairwise comparisons were performed using McNemar tests with Bonferroni correction to account for multiple comparisons. RHS values were compared across the four chatbot models using the Kruskal–Wallis test. A two-tailed $p<0.05$ was considered statistically significant.

Table 1. Comparison of total clinical adherence scores across chatbot models

Chatbot models	Mean±SD (%)	p
ChatGPT-5.2	70.6±15.7	0.089
Perplexity AI (%)	72.7±14.0	
Gemini 2.5 Flash	75.0±19.5	
DeepSeek-v3.2	60.0±13.0	

Values are presented as mean±standard deviation (SD) percentages based on a 9-item IADT clinical adherence index. Total scores were normalized to a 0–100 scale. Between-model comparisons were performed using the Friedman test. A $p<0.05$ was considered statistically significant.

Results

Comparison of percentage-based total scores across chatbot models showed no statistically significant difference among the models ($p=0.089$), indicating comparable overall clinical performance. In contrast, statistically significant differences emerged in several technical domains requiring precise interpretation of the IADT guidelines. Significant between-group differences were observed for root surface cleaning ($p=0.017$). The greatest divergence among chatbot models was identified in the splint type and duration criterion, which demonstrated a highly significant difference ($p<0.001$). Mean clinical adherence scores for each chatbot model are presented in Table 1, whereas detailed item-level performance across all evaluated IADT-based criteria is shown in Table 2.

Post hoc pairwise comparisons were performed for clinical criteria demonstrating significant overall differences using McNemar tests with Bonferroni correction (adjusted significance level $p<0.008$). For root surface cleaning, ChatGPT-5.2 and Perplexity AI showed higher accuracy compared to DeepSeek-v3.2, with differences reaching statistical significance after correction ($p<0.008$). No statistically significant difference was observed between ChatGPT-5.2 and Perplexity AI. Comparisons involving Gemini 2.5 Flash showed a trend

Table 2. Comparison of chatbot performance across IADT-based criteria

Outcome/Criterion	ChatGPT-5.2 (%)	Perplexity AI (%)	Gemini 2.5 Flash (%)	DeepSeek-v3.2 (%)	p	Effect size
Indication for replantation	100.0	100.0	90.0	90.0	0.572	0.11
Root surface cleaning	75.0	75.0	37.5	0.0	0.017	0.42
Replantation site preparation	44.4	100.0	77.8	55.6	0.530	0.28
Endodontic considerations	55.6	66.7	88.9	66.7	0.438	0.15
Antibiotic recommendation	80.0	70.0	90.0	90.0	0.194	0.22
Tetanus prophylaxis	80.0	80.0	90.0	80.0	0.875	0.04
Splint type and duration	88.9	0.0	88.9	55.6	0.001	0.58
Post-operative care	30.0	80.0	40.0	40.0	0.690	0.26
Follow-up schedule	80.0	80.0	70.0	80.0	0.912	0.03

Values represent percentages of correct responses for each clinical criterion. Between-model comparisons were performed using Cochran's Q test. A $p<0.05$ was considered statistically significant. Effect size: Kendall's W interpreted as: 0.1–0.3 (Small), 0.3–0.5 (Moderate), >0.5 (Large). IADT: International Association of Dental Traumatology.

toward lower performance; however, these differences did not consistently reach the adjusted significance threshold. The most pronounced between-model differences were observed for splint type and duration. Both ChatGPT-5.2 and Gemini 2.5 Flash demonstrated significantly higher accuracy than Perplexity AI ($p<0.008$). No statistically significant difference was detected between ChatGPT-5.2 and Gemini 2.5 Flash. Comparisons involving DeepSeek-v3.2 indicated lower performance but did not consistently reach statistical significance after Bonferroni correction. The comparative performance of the chatbot models across these discriminative criteria is shown in Appendix Figure 1.

Follow-up performance differed across chatbot models. Correct reporting of the recommended clinical and radiographic follow-up schedule was observed in 80% of responses generated by ChatGPT-5.2, Perplexity AI, and DeepSeek-v3.2, whereas Gemini 2.5 Flash demonstrated a lower accuracy of 70%. Although overall compliance scores were comparable, these findings indicate that follow-up recommendations constitute a domain in which clinically relevant discrepancies persist across models, particularly in scenarios requiring structured long-term monitoring. However, the difference in follow-up schedule accuracy among chatbot models did not reach statistical significance ($p=0.912$).

RHS values across chatbot models are presented in Table 3. Although ChatGPT-5.2 and Perplexity AI exhibited wider score distributions with occasional high outliers, no statistically significant difference in RHS values was observed among ChatGPT-5.2, Perplexity AI, Gemini 2.5 Flash, and DeepSeek-v3.2 ($H=5.95$, $p=0.114$).

Table 3. Comparison of RHS values across chatbot models

Chatbot	n	Mean±SD	Median	Min–Max
ChatGPT-5.2	10	1.6 ± 2.46	1	0–8
DeepSeek-v3.2	10	1.6 ± 0.84	2	0–2
Gemini 2.5 Flash	10	1.6 ± 0.84	2	0–2
Perplexity AI	10	1.1 ± 2.51	0	0–8

Values are presented as mean±standard deviation (SD), median, and minimum–maximum. RHS values were compared across chatbot models using the Kruskal–Wallis test. A $p<0.05$ was considered statistically significant. RHS: Reference hallucination score.

Discussion

The present study evaluated the extent to which four widely used AI-based chatbots comply with the 2020 IADT avulsion guidelines when responding to standardized, clinician-directed scenarios. In accordance with the study hypothesis, no statistically significant difference was observed in overall normalized performance scores among the chatbot models; therefore, the null hypothesis was accepted. However, meaningful disparities emerged in specific guideline-sensitive domains. These findings suggest that overall performance metrics may mask potential deficiencies in technically demanding aspects of avulsion management that could influence prognosis.[2–6]

The absence of statistically significant differences in total performance scores suggests that current AI chatbots are generally capable of reproducing broad avulsion management frameworks. Most models consistently addressed general components such as replantation indications, antibiotic recommendations, tetanus assessment, and follow-up instructions. However, these elements represent general principles rather than scenario-

dependent clinical decisions. Similar observations have been reported in previous studies evaluating clinical decision-support tools and AI-assisted emergency triage systems.[10,11,21,22] Avulsion management is inherently algorithmic and unforgiving, and success depends not on general principles but on precise execution of time-dependent and scenario-specific steps, as emphasized in the IADT guidelines.[5,6] Recent studies evaluating large language models in dentistry have similarly demonstrated that AI systems may achieve acceptable overall performance while showing variability in clinically sensitive decision points requiring contextual interpretation.[15–17]

One of the key findings of this study was the presence of statistically significant between-model differences in root surface cleaning, splint type, and duration. These domains require a nuanced interpretation of the IADT guidelines and careful differentiation based on extra-oral dry time, storage conditions, and PDL viability.[3–5] These findings suggest that AI models may differ in their ability to integrate biological healing principles with procedural recommendations, rather than simply retrieving guideline information. This variability may reflect differences in how AI models prioritize procedural details versus general clinical principles, which has also been reported in previous evaluations of AI-based clinical decision-support tools.[15–17]

ChatGPT-5.2 and Perplexity AI consistently outperformed Gemini 2.5 Flash and DeepSeek-v3.2 in root surface cleaning recommendations. Correct differentiation between gentle saline rinsing, avoidance of mechanical scraping, and the use of fluoride treatment in delayed replantation scenarios is essential to reduce the risk of inflammatory or replacement resorption.[3,4] Inaccurate guidance at this step may irreversibly compromise periodontal healing, even if subsequent management steps are appropriate.

The most pronounced differences between chatbot models were observed in splint type and duration, which represent highly guideline-sensitive components of avulsion management. Flexible splinting for up to 2 weeks is a cornerstone of modern trauma care,[5] yet rigid splinting or inappropriate splint duration remains a common clinical error among practitioners.[7,8] The markedly lower accuracy observed for Gemini 2.5 Flash and DeepSeek-v3.2 in this criterion is clinically concerning, as incorrect splinting has been directly associated with increased rates of ankylosis and replacement resorption.[1,3] Splint selection represents a particularly sensitive decision point in avulsion management, as it requires integration of biological healing principles with mechanical stabilization strategies. Similar variability in AI performance has been reported in recent dental AI studies, particularly in tasks requiring structured clinical judgment and protocol-based decision-making.[23–25]

Although some of the observed effect sizes were statistically moderate, their clinical implications may nevertheless be substantial. In avulsion management, even seemingly limited deviations from guideline recommendations in highly sensitive steps, such as splint type and duration, root surface management, or treatment decisions based on extra-oral dry time, can significantly influence PDL healing and pulpal prognosis. Unlike less critical supportive recommendations, errors in these domains may increase the risk of complications including inflammatory root resorption, replacement resorption, ankylosis, or long-term tooth loss. Therefore, moderate inter-model differences should not be interpreted solely as statistical variations, but rather in the context of their potential impact on clinical outcomes and evidence-based trauma management.

Although no statistically significant differences were observed in RHS values, the presence of high outliers indicates that even higher-performing models may occasionally generate inaccurate or fabricated references. Nevertheless, the presence of high outliers – particularly in ChatGPT-5.2 and Perplexity AI – underscores that even high-performing systems may occasionally generate severely inaccurate or fabricated citations. This finding is consistent with earlier reports documenting reference hallucination rates ranging from 30% to 60% in AI-generated medical and dental content.[12–14] These findings further emphasize that fluent and confident AI-generated outputs do not necessarily reflect factual accuracy, particularly in reference generation.

Given that the 2020 IADT guidelines serve as a definitive reference in dental traumatology, misquotation or fabrication of guideline citations poses a tangible risk, especially if AI-generated outputs are accepted uncritically in emergency settings. Even a single hallucinated or misattributed reference may mislead clinicians, educators, or trainees, potentially compromising evidence-based trauma management.[23,24]

From a clinical perspective, these findings suggest that AI chatbots should not be used as standalone decision-support tools for avulsion management. While they may assist in recalling general management principles, their variable performance in technically critical steps limits their reliability in real-time trauma care. Because the evaluations were based on adherence to IADT guidelines, responses categorized as inaccurate or partially compliant may also reflect recommendations that could be clinically inappropriate or potentially risky in real-life trauma management scenarios. Therefore, inter-model differences

should be interpreted not only as numerical performance variations but also in terms of their possible clinical consequences. This concern aligns with prior evaluations of AI use in emergency medicine and dentistry, which emphasize the continued necessity of clinician oversight and guideline literacy.[10,11,25–27] Therefore, AI-generated recommendations should be interpreted cautiously and verified against established clinical guidelines.

These findings also have implications for dental education and training. AI-generated responses that are partially correct may create a false sense of competence among students or junior clinicians, particularly when errors occur in less intuitive aspects of care such as splint selection or root surface management. Previous studies in dental education have highlighted the risk that AI tools may inadvertently reinforce superficial learning unless explicitly integrated into curricula emphasizing critical appraisal and guideline verification.[9,28–30] This highlights the importance of incorporating AI literacy into dental education curricula.

Several limitations should be acknowledged. First, the use of a dichotomous (0/1) scoring system may oversimplify partially correct responses and does not capture nuances in clinical reasoning. In addition, all items were equally weighted, although certain criteria (e.g., splint type and duration, or management following prolonged dry time) may have greater prognostic significance. Second, the study evaluated a limited number of predefined scenarios, which, although carefully designed to represent key avulsion conditions, cannot encompass the full spectrum of clinical variability encountered in practice. Although the scenarios were developed according to the 2020 IADT guidelines, the inclusion of only 10 predefined scenarios may limit the robustness and generalizability of the findings. A larger and more diverse scenario set could reveal additional performance variability and potentially detect subtle inter-model differences that were not identified in the present study. Third, chatbot performance reflects a snapshot in time; ongoing model updates may alter compliance patterns. In addition, only a single response per scenario was evaluated, and potential variability across repeated queries was not assessed. Fourth, responses were assessed in English only, and performance may differ across languages. Fifth, the relatively small sample size and limited number of scenarios may have reduced the statistical power to detect subtle differences between chatbot models. Therefore, the findings should be interpreted with caution, and future studies with larger datasets are warranted. Finally, the study evaluated textual recommendations rather than real-world clinical behavior, which may be influenced by contextual and experiential factors not captured in prompt-based interactions.

Future research should expand scenario complexity, include additional trauma categories, and evaluate longitudinal changes in AI performance following model updates. Investigating clinician–AI interaction patterns and assessing whether AI-assisted decision-making improves or impairs adherence to trauma guidelines in simulated or real clinical environments would also be valuable, as suggested by previous work on digital decision-support systems in dental traumatology.

Conclusion

Performance was higher in simpler scenarios, such as immediate replantation, whereas more complex conditions – particularly prolonged dry time and primary tooth avulsion – were associated with lower compliance and greater variability. Errors were most frequently observed in detailed clinical parameters, including splinting protocols and endodontic management. Reference accuracy also varied across scenarios, with higher hallucination scores observed in certain conditions, particularly in primary tooth cases, indicating inconsistencies in generating accurate bibliographic information. Overall, while AI chatbots demonstrate comparable overall performance in managing avulsion scenarios according to the 2020 IADT guidelines, notable discrepancies were observed in technically critical components that may directly influence clinical decision-making and prognosis. These findings suggest that aggregate performance scores may obscure clinically relevant differences in guideline-sensitive steps. Given that the analysis was based on a single response per scenario collected at a single time point, the results should be interpreted with caution, as chatbot outputs may vary across sessions and model updates. Accordingly, AI-generated guidance should be considered as a supportive educational resource rather than a substitute for established clinical protocols in dental trauma management.

Ethics Committee Approval: No human subjects or biological samples were involved; therefore, ethics approval was not required.

Conflict of Interest: The author declare that there is no conflict of interest.

Financial Disclosure: The authors declared that this study received no financial support.

Use of AI for Writing Assistance: The authors declared that artificial intelligence was not used in the study.

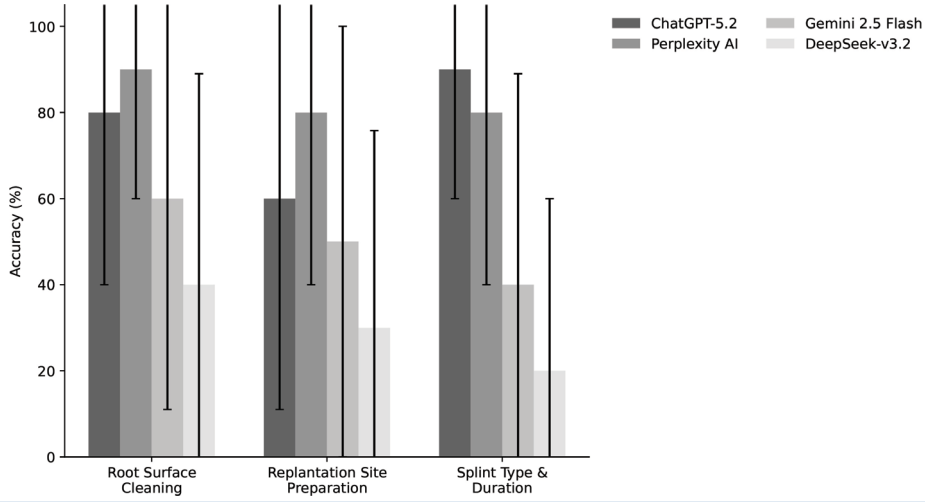
Authorship Contributions: Concept: EYM; Design: MO; Supervision: EYM, MO; Resource: EYM, MO; Data Collection and/or Processing: MO; Analysis and/or Interpretation: EYM; Literature Search: EYM, MO; Writing: EYM, MO; Critical Reviews: EYM, MO.

Peer-review: Double blind peer-reviewed.

References

1. Bourguignon C, Cohenca N, Lauridsen E, Flores MT, O'Connell AC, Day PF, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: 1. Fractures and luxations. Dent Traumatol 2020;36(4):314-30. [CrossRef]
2. Andreasen JO, Andreasen FM. Essentials of traumatic injuries to the teeth: a step-by-step treatment guide. 2nd ed. United States: Wiley-Blackwell; 2010.
3. Pohl Y, Filippi A, Kirschner H. Results after replantation of avulsed permanent teeth. II. Periodontal healing and the role of physiologic storage and antiresorptive-regenerative therapy. Dent Traumatol 2005;21(2):93-101. [CrossRef]
4. Trope M. Clinical management of the avulsed tooth: present strategies and future directions. Dent Traumatol 2002;18(1):1-11. [CrossRef]
5. Fouad AF, Abbott PV, Tsilingaridis G, Cohenca N, Lauridsen E, Bourguignon C, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: 2. Avulsion of permanent teeth. Dent Traumatol 2020;36(4):331-42. [CrossRef]
6. Day PF, Flores MT, O'Connell AC, Abbott PV, Tsilingaridis G, Fouad AF, 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(4):343-59. [CrossRef]
7. Al-Huthaifi BH, Ghwainem AA, Alqarni AS, Alshehri BY, Almnea RA, Alelyani AA, et al. Knowledge, perception, and management toward traumatic tooth avulsion among dental professionals: a cross-sectional study. BMC Med Educ 2025;25(1):1206. [CrossRef]
8. Mustuloğlu Ş, Deniz BP. Evaluation of Chatbots in the Emergency Management of Avulsion Injuries. Dent Traumatol 2025;41(4):437-44. [CrossRef]
9. Çege EE, Cömert H, Akal N, Ölmez A. Evaluation of the Performance of Artificial Intelligence Based Chatbots in Providing First Aid Information on Dental Trauma According to the ToothSOS Application. Dent Traumatol 2025;41(6):696-705. [CrossRef]
10. 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. Dent Traumatol 2025;41(2):187-93. [CrossRef]
11. Tokgöz Kaplan T, Cankar M. Evidence-Based Potential of Generative Artificial Intelligence Large Language Models on Dental Avulsion: ChatGPT Versus Gemini. Dent Traumatol 2025;41(2):178-86. [CrossRef]
12. Aljamaan F, Temsah MH, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. JMIR Med Inform 2024;12:e54345. [CrossRef]
13. Hua HU, Kaakour AH, Rachitskaya A, Srivastava S, Sharma S, Mammo DA. Evaluation and comparison of ophthalmic scientific abstracts and references by current artificial intelligence chatbots. JAMA Ophthalmol 2023;141(9):819-24. [CrossRef]
14. Sharun K, Banu SA, Pawde AM, Kumar R, Akash S, Dhama K, Pal A. ChatGPT and artificial hallucinations in stem cell research: assessing the accuracy of generated references - a preliminary study. Ann Med Surg (Lond) 2023;85(10):5275-8. [CrossRef]
15. Demir Cicek B, Cicek O. Evaluating the Response of AI-Based Large Language Models to Common Patient Concerns About Endodontic Root Canal Treatment: A Comparative Performance Analysis. J Clin Med 2025;14(21):7482. [CrossRef]
16. Yıldırım A, Cicek O, Genç YS. Can AI-Based ChatGPT Models Accurately Analyze Hand-Wrist Radiographs? A Comparative Study. Diagnostics (Basel) 2025;15(12):1513. [CrossRef]
17. Yıldırım A, Cicek O. Assessment of AI-Driven Large Language Models for Orthodontic Aesthetic Scoring Using the IOTN-AC. Diagnostics (Basel) 2025;15(23):3048. [CrossRef]
18. Guven Y, Ozdemir OT, Kavan MY. Performance of Artificial Intelligence Chatbots in Responding to Patient Queries Related to Traumatic Dental Injuries: A Comparative Study. Dent Traumatol 2025;41(3):338-47. [CrossRef]
19. Grinberg N, Arbel S, Boyadjiev YY, Ianculovici C, Kleinman S, Peleg O. The Performance of Artificial Intelligence in Providing Real-Time Aid in Emergency Dental Trauma: A Clinical Validation Study. Dent Traumatol 2026;42(3):356-62. [CrossRef]
20. Keleş ÖK, Arslan ZB. Performance of artificial intelligence chatbots in the diagnosis and management of simulated dental trauma cases: an evaluation based on IADT guidelines. Clin Oral Investig 2025;30(1):26. [CrossRef]
21. Kueper JK, Terry AL, Zwarenstein M, Lizotte DJ. Artificial Intelligence and Primary Care Research: A Scoping Review. Ann Fam Med 2020;18(3):250-8. [CrossRef]
22. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2023;55(12):1-38. [CrossRef]
23. Miller DD, Brown EW. Artificial Intelligence in Medical Practice: The Question to the Answer? Am J Med 2018;131(2):129-33. [CrossRef]
24. Borji A. A categorical archive of ChatGPT failures. arXiv 2023;2302.03494. [CrossRef]
25. Najeeb M, Islam S. Artificial intelligence (AI) in restorative dentistry: current trends and future prospects. BMC Oral Health 2025;25(1):592. [CrossRef]
26. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44-56. [CrossRef]
27. Lysaght T, Lim HY, Xafis V, Ngiam KY. AI-Assisted Decision-making in Healthcare: The Application of an

Ethics Framework for Big Data in Health and Research. Asian Bioeth Rev 2019;11(3):299-314. [CrossRef]
28. Tiwari A, Kumar A, Jain S, Dhull KS, Sajjanar A, Puthenkandathil R, et al. Implications of ChatGPT in Public Health Dentistry: A Systematic Review. Cureus 2023;15(6):e40367. [CrossRef]
29. Thurzo A, Strunga M, Urban R, Surovková J, Afrashtehfar KI. Impact of artificial intelligence on dental education: a review and guide for curriculum update. Educ Sci (Basel) 2023;13(2):150. [CrossRef]
30. Masters K. Artificial intelligence in medical education. Med Teach 2019;41(9):976-80. [CrossRef]



Appendix Figure 1. Comparison of chatbot performance across discriminative International Association of Dental Traumatology-based clinical criteria. Bars represent mean accuracy (%) and error bars indicate standard deviation. The figure displays performance across root surface cleaning, replantation site preparation, and splint type and duration for ChatGPT-5.2, Perplexity AI, Gemini 2.5 Flash, and DeepSeek-v3.2.

Supplementary Material

1. **Permanent maxillary incisor with closed apex – 30 min dry time**
   Prompt: A 12-year-old patient presents with an avulsed permanent maxillary incisor with a closed apex that has remained dry for 30 min. How should this case be managed according to the IADT Guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
2. **Permanent maxillary incisor with open apex – 30 min dry time**
   Prompt: A 7-year-old child presents with an avulsed permanent maxillary incisor with an open apex that has remained dry for 30 min. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
3. **Permanent maxillary incisor with closed apex – 90 min dry time**
   Prompt: A 12-year-old patient presents 90 min after avulsion of a permanent maxillary incisor with a closed apex, left completely dry. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
4. **Permanent maxillary incisor with open apex – 90 min dry time**
   Prompt: A 7-year-old child presents with an open-apex permanent incisor avulsed and dry for 90 min. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
5. **Permanent maxillary incisor with closed apex – stored in milk for 30 min**
   Prompt: A 12-year-old patient with a closed apex presents after an avulsion injury involving a permanent maxillary incisor, with the tooth stored in milk for 30 min. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
6. **Permanent maxillary incisor with open apex – stored in milk for 30 min**
   Prompt: A 7-year-old child presents with an open-apex incisor that was stored in milk for 30 min. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
7. **Permanent maxillary incisor with closed apex – stored in milk for 90 min**
   Prompt: A 12-year-old patient presents with an avulsed permanent maxillary incisor with a closed apex that was stored in milk for 90 min. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
8. **Permanent maxillary incisor with open apex – stored in milk for 90 min**
   Prompt: A 7-year-old presents with an open-apex incisor stored in milk for 90 min. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
9. **Avulsed primary maxillary incisor**
   Prompt: A young child presents with a completely avulsed primary maxillary incisor. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
10. **Immediately replanted tooth at the accident site (12-year-old patient)**
    Prompt: A 12-year-old patient presents with an avulsed permanent maxillary incisor with a closed apex that was immediately replanted at the accident site. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.

Appendix Table 1. Scenario-level IADT clinical compliance scoring matrix for all chatbot models across all evaluated avulsion scenarios

Scenario	Chatbots	IfR	RSC	RSP	EC	AR	TP	STaD	Post-op	FUS	TS (0–9)
Closed apex – 30 min dry (12 years)	ChatGPT-5.2	1	1	0	1	1	1	0	0	1	6
Open apex – 30 min dry (7 years)	ChatGPT-5.2	1	0	1	1	1	1	1	1	1	8
Closed apex – 90 min dry (12 years)	ChatGPT-5.2	1	1	1	0	1	1	1	0	1	7
Open apex – 90 min dry (7 y)	ChatGPT-5.2	1	1	0	0	1	1	1	0	1	6
Closed apex – milk 30 min (12 years)	ChatGPT-5.2	1	1	0	1	1	0	1	0	1	6
Open apex – milk 30 min (7 years)	ChatGPT-5.2	1	1	0	1	0	1	1	0	0	5
Closed apex – milk 90 min (12 years)	ChatGPT-5.2	1	1	1	0	1	1	1	0	1	7
Open apex – milk 90 min (7 years)	ChatGPT-5.2	1	0	0	0	1	1	1	0	1	5
Avulsed primary maxillary incisor	ChatGPT-5.2	1	NA	1	NA	0	0	NA	1	0	3
Immediately replanted the tooth at the accident site (12 years)	ChatGPT-5.2	1	NA	NA	1	1	1	1	1	1	7
Closed apex – 30 min dry (12 years)	Perplexity AI	1	1	1	1	1	1	0	1	1	8
Open apex – 30 min dry (7 years)	Perplexity AI	1	1	1	0	0	1	0	1	1	6
Closed apex – 90 min dry (12 years)	Perplexity AI	1	1	1	0	1	1	0	1	0	6
Open apex – 90 min dry (7 years)	Perplexity AI	1	1	1	1	1	1	0	1	1	8
Closed apex – milk 30 min (12 years)	Perplexity AI	1	1	1	1	1	1	0	1	1	8
Open apex – milk 30 min (7 years)	Perplexity AI	1	1	1	1	0	0	0	0	0	4
Closed apex – milk 90 min (12 years)	Perplexity AI	1	0	1	1	1	1	0	0	1	6
Open apex – milk 90 min (7 years)	Perplexity AI	1	0	1	1	1	1	0	1	1	7
Avulsed primary maxillary incisor	Perplexity AI	1	NA	1	NA	0	0	NA	1	1	4
Immediately replanted tooth at the accident site (12 years)	Perplexity AI	1	NA	NA	0	1	1	0	1	1	5
Closed apex – 30 min dry (12 years)	Gemini 2.5 Flash	1	0	1	1	1	1	1	0	0	6
Open apex – 30 min dry (7 years)	Gemini 2.5 Flash	1	0	1	1	1	1	1	0	1	7
Closed apex – 90 min dry (12 years)	Gemini 2.5 Flash	1	0	1	1	1	1	1	0	0	6
Open apex – 90 min dry (7 years)	Gemini 2.5 Flash	1	0	1	0	1	1	0	0	1	5
Closed apex – milk 30 min (12 years)	Gemini 2.5 Flash	1	1	0	1	1	1	1	0	1	7
Open apex – milk 30 min (7 years)	Gemini 2.5 Flash	1	0	0	1	1	0	1	0	1	5
Closed apex – milk 90 min (12 years)	Gemini 2.5 Flash	1	1	1	1	1	1	1	1	1	9
Open apex – milk 90 min (7 years)	Gemini 2.5 Flash	1	1	1	1	1	1	1	1	1	9
Avulsed primary maxillary incisor	Gemini 2.5 Flash	0	NA	1	NA	0	1	NA	1	0	3
Immediately replanted tooth at the accident site (12 years)	Gemini 2.5 Flash	1	NA	NA	1	1	1	1	1	1	7
Closed apex – 30 min dry (12 years)	DeepSeek-v3.2	1	0	0	1	1	1	1	0	1	6
Open apex – 30 min dry (7 years)	DeepSeek-v3.2	1	0	0	1	1	0	1	0	1	5
Closed apex – 90 min dry (12 years)	DeepSeek-v3.2	1	0	0	0	1	1	0	1	1	5
Open apex – 90 min dry (7 years)	DeepSeek-v3.2	1	0	1	1	1	1	0	0	1	6
Closed apex – milk 30 min (12 years)	DeepSeek-v3.2	1	0	0	1	1	1	1	0	1	6
Open apex – milk 30 min (7 years)	DeepSeek-v3.2	1	0	1	1	1	1	1	0	1	7
Closed apex – milk 90 min (12 years)	DeepSeek-v3.2	1	0	1	1	1	1	0	1	0	6
Open apex – milk 90 min (7 years)	DeepSeek-v3.2	0	0	1	0	1	1	0	0	1	4
Avulsed primary maxillary incisor	DeepSeek-v3.2	1	NA	1	NA	0	0	NA	1	0	3
Immediately replanted tooth at the accident site (12 years)	DeepSeek-v3.2	1	NA	NA	0	1	1	1	1	1	6

IADT: International Association of Dental Traumatology; IfR: Indication for replantation; RSC: Root surface cleaning; RSP: Replantation site preparation; EC: Endodontic considerations; AR: Antibiotic recommendation; TP: Tetanus prophylaxis; STaD: Splint type and duration; Post-op: Post-operative care; FUS: Follow-up schedule; TS: Total score.

Appendix Table 2. Reference hallucination score for each chatbot-generated bibliographic reference across all scenarios

Scenario	ChatGPT-5.2	DeepSeek-v3.2	Gemini 2.5 Flash	Perplexity AI
Avulsed primary maxillary incisor	8	2	0	8
Closed apex – 30 min dry (12 years)	0	0	2	0
Closed apex – 90 min dry (12 years)	2	2	2	0
Closed apex – milk 30 min (12 years)	0	2	2	0
Closed apex – milk 90 min (12 years)	2	2	2	0
Immediately replanted tooth at the accident site (12 years)	2	2	0	2
Open apex – 30 min dry (7 years)	0	0	2	0
Open apex – 90 min dry (7 years)	0	2	2	1
Open apex – milk 30 min (7 years)	0	2	2	0
Open apex – milk 90 min (7 years)	2	2	2	0