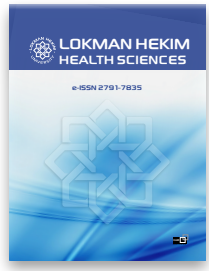
LOKMAN HEKIM HEALTH SCIENCES

DOI: 10.14744/lhhs.2026.84647
Lokman Hekim Health Sci 2026;6(2):221–227



lokmanhekimhs.com

ORIGINAL ARTICLE

Accuracy of Artificial Intelligence-Generated References in Dental Trauma Management

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

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

Abstract

Introduction: The objective of the study is to evaluate the accuracy of scientific references generated by artificial intelligence (AI) chatbots in response to clinical scenarios related to traumatic dental injuries (TDIs) and to determine the potential impact of reference errors on clinical decision-making.

Methods: This cross-sectional observational study analyzed 400 references generated by four AI chatbots (ChatGPT, Perplexity AI, Gemini, DeepSeek) in response to ten clinical prompts representing internationally recognized TDI categories. Each chatbot was instructed to retrieve recent PubMed-indexed studies and provide full bibliographic data. Reference authenticity and accuracy were verified using PubMed, Scopus, and Google Scholar. Hallucination severity was quantified using the reference hallucination score (RHS) scale (0–11). Non-parametric statistics and generalized linear modeling were applied (α=0.05).

Results: Significant differences in RHS were observed between chatbots ($p<0.001$). ChatGPT and Perplexity AI demonstrated significantly lower hallucination severity compared with Gemini and DeepSeek ($p<0.001$). Trauma category showed no significant effect on RHS ($p>0.05$). Internal consistency for RHS components was acceptable to excellent (Cronbach's α=0.82).

Discussion and Conclusion: Although AI chatbots may provide rapid guidance for TDI management, the reliability of their generated references varies considerably across models. The presence of fabricated or inaccurate citations represents a potential risk for evidence-based clinical decision-making.

Keywords: Artificial intelligence; Bibliographic accuracy; Chatbot; Dental trauma; Evidence-based dentistry; Reference hallucination

Traumatic dental injuries (TDIs) represent one of the most challenging emergencies in dental practice, often requiring rapid diagnosis and evidence-based intervention to preserve pulp vitality, periodontal support, tooth function, and esthetics. The prognosis of traumatized teeth is strongly influenced by the accuracy of initial assessment, correct classification of the injury, and adherence to current recommendations such as those issued by the International Association of Dental Traumatology (IADT).[1–3] However, because TDIs encompass a wide spectrum of clinical presentations – including avulsion, luxation, and root fractures – uncertainty regarding optimal management is common, even among experienced clinicians.[4]

Cite this article as: *Yıldırım Manav E, Özdemir M. Accuracy of Artificial Intelligence-Generated References in Dental Trauma Management. Lokman Hekim Health Sci 2026;6(2):221–227.*

Correspondence: Esra Yıldırım Manav, M.D. Lokman Hekim Üniversitesi, Diş Hekimliği Fakültesi, Restoratif Diş Hekimliği Anabilim Dalı, Ankara, Türkiye
E-mail: esra.manav@lokmanhekim.edu.tr **Submitted:** 04.02.2026 **Revised:** 24.02.2026 **Accepted:** 03.04.2026 **Available Online:** 21.05.2026

In parallel with advances in digital health, artificial intelligence (AI) chatbots based on large language models (LLMs) are increasingly being used by clinicians and students as accessible sources of clinical information. Systems such as ChatGPT (OpenAI, USA), Perplexity AI (Perplexity Inc., USA), Gemini (Google DeepMind, USA), and DeepSeek (DeepSeek AI, China) are capable of generating structured responses to diagnostic and treatment-related queries and frequently provide supporting scientific references, thereby mimicking expert consultation.[5–7] While these tools have the potential to enhance clinical decision-making and facilitate rapid access to knowledge, their bibliographic reliability remains uncertain – particularly in acute conditions such as dental trauma, where treatment outcomes are time-sensitive.

A major concern is the phenomenon of reference hallucination, whereby AI models generate fabricated or partially incorrect citations that appear scientifically credible but do not correspond to verifiable publications.[8,9] Such inaccuracies may involve nonexistent article titles, incorrect author names, or invalid journal and digital object identifier (DOI) information. In medicine and dentistry, hallucinated references may mislead clinicians, distort evidence interpretation, and undermine academic integrity.[10–12] Reported hallucination rates exceeding 50–60% in general biomedical settings suggest that this issue may be widespread.[13] However, to date, there is no structured assessment of AI-generated reference accuracy specifically within dental traumatology – a discipline where erroneous evidence may directly affect emergency management.

Bibliographic accuracy may also vary across AI systems depending on their architecture. Retrieval-augmented models, such as ChatGPT-5 and Perplexity AI, incorporate live search or database grounding, which may reduce hallucination risk. In contrast, closed-source generative systems such as DeepSeek rely primarily on probabilistic language prediction and may therefore produce fabricated references more frequently.[14,15] Whether these architectural differences translate into measurable disparities in citation reliability during trauma-related clinical querying remains unknown.

To address this evidence gap, the primary aim of the present study was to evaluate the accuracy and reliability of AI-generated references for standardized TDI scenarios using the reference hallucination score (RHS) framework.[1,2] A secondary objective was to compare four widely used AI chatbots in terms of the frequency and severity of reference hallucinations and to determine whether hallucination patterns differ according to chatbot model or trauma category. The null hypothesis was that no statistically significant differences would exist in RHS values among the evaluated chatbots, irrespective of trauma type or system architecture.

Materials and Methods

Study Design

This cross-sectional observational study was conducted in December 2025 to evaluate the accuracy of bibliographic references generated by AI chatbots in response to standardized dental trauma management scenarios. Ethics committee approval was not required because the study did not include human subjects, patient information, or biological materials. The methodological framework was based on the validated RHS system proposed by Aljamaan et al.[13] and later adapted for dental research reliability assessment.[2]

Selection of AI Chatbots

Four publicly accessible, English-language AI chatbots were selected based on their global academic usage, technological diversity, and public availability. The evaluated systems included ChatGPT (version 5.2), Perplexity AI, Gemini (3 Flash), and DeepSeek, all of which were accessed in their most recent publicly available versions at the time of data collection.

Each chatbot was accessed through its official web interface using a verified account on a secure institutional network. All sessions were conducted in incognito mode to prevent previous chat memory from influencing outputs.

Sample Size and Power Calculation

An a priori power analysis was performed using GPower 3.1.9.6* (Heinrich-Heine University, Düsseldorf, Germany). Assuming a medium effect size (f=0.25), significance level α=0.05, and statistical power (1−β)=0.80, the minimum required total sample size was estimated at 160 references (≈ 40 per chatbot). The assumption of a medium effect size was based on conventional benchmarks proposed by Cohen and was consistent with a previous study evaluating the performance of LLM-based chatbots in dentistry.[16] Given the absence of established benchmarking data specific to AI-based guideline adherence in dental traumatology, a medium effect size was considered a methodologically appropriate and conservative estimate to detect practically meaningful differences between chatbot models. Since the present study analyzed 400 references (4 chatbots × 10 prompts × 10

references), the achieved power exceeded 0.95, confirming adequate sample size for reliable inter-model comparisons.

Prompt Development

Ten standardized clinical prompts were developed to represent the most common types of TDIs as classified by the IADT guidelines.[3–5] Each prompt simulated a realistic case scenario including patient age, injury type, and clinical question. The prompts requested both a short management summary and a list of ten supporting literature references in Vancouver style.

Each scenario prompt consisted of (i) a clinical vignette and (ii) a standardized instruction block applied identically across all chatbots. This block required the model to (a) begin by searching PubMed; (b) select 10 recent and relevant articles; and (c) report, for each article, the title, authors, journal, publication date, citation count, DOI, web link, and PubMed link, formatted consistently. The full prompt templates are provided in Appendix 1.

Each chatbot received the same ten prompts in identical order and formatting. All interactions were carried out by a single experienced investigator (EYM) to ensure standardization and consistency. All responses were exported as plain text and anonymized for evaluation.

Reference Verification

All references generated by the chatbots were manually verified using PubMed, Scopus, and Google Scholar. Each reference was checked for existence, bibliographic accuracy, and relevance to the prompted topic. If a reference could not be found in any of the databases or exhibited falsified information (e.g., fabricated title, incorrect author list, or non-existent DOI), it was classified as hallucinated. Minor discrepancies, such as incorrect publication year or typographical errors, were considered partial hallucinations.

Scoring Criteria

The RHS was applied to quantify hallucination severity for each citation across seven bibliographic identifiers:

1. Title
2. Authors' names
3. Journal name
4. Publication year
5. Digital object identifier (DOI)
6. Web link (URL)
7. Relevance to the trauma topic.

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). All references were independently evaluated by two calibrated reviewers with expertise in dental traumatology. The reviewers were blinded to the chatbot identity during scoring. Each citation was assessed according to the predefined RHS criteria. In cases of disagreement, the reference was re-evaluated through discussion, and consensus was reached. Inter-rater reliability for RHS scoring was assessed using the intraclass correlation coefficient (ICC), demonstrating excellent agreement (ICC=0.93).

A mean RHS value was calculated for each chatbot and for each trauma category.

Higher scores indicated greater factual inaccuracy. All references were assessed independently of whether the corresponding textual answer was clinically accurate or inaccurate. Thus, hallucination scoring reflected citation validity rather than clinical reasoning quality.

Statistical Analysis

Statistical analyses were performed using IBM Statistical Package for the Social Sciences Statistics v29 (IBM Corp., Armonk, NY, USA). Data normality was assessed using the Kolmogorov–Smirnov test, and continuous variables were expressed as mean±standard deviation. Inter-model differences in RHS values were examined using the Kruskal–Wallis test, followed by Bonferroni-adjusted pairwise post hoc comparisons where appropriate. Differences in RHS values across trauma categories were analyzed using the Mann–Whitney U test or Kruskal–Wallis test, depending on the number of groups being compared. To identify independent predictors of hallucination severity, a Generalized Linear Model (gamma distribution, log-link function) was constructed, with chatbot type and trauma category entered as fixed factors. Internal consistency across RHS components was evaluated using Cronbach's α coefficient. A $p<0.05$ was considered statistically significant.

Results

Table 1 shows that there were marked differences in RHS scores between the evaluated chatbots. A statistically significant difference in RHS values among the four chatbots was confirmed using the Kruskal–Wallis test ($p<0.001$).

Table 1. Reference the hallucination score across the evaluated artificial intelligence-based chatbot models

Chatbot	Mean±SD	Median (IQR)	Min–Max
ChatGPT	3.92±2.63	4.0 (2.0–6.0)	0–10
Perplexity AI	4.65±3.04	4.0 (2.75–6.0)	0–10
Gemini	6.21±3.62	6.5 (4.0–10.0)	0–10
DeepSeek	7.18±3.79	10.0 (5.0–10.0)	0–10
p	<0.001		

AI: Artificial intelligence; SD: Standard deviation; IQR: Interquartile range; Min: Minimum; Max: Maximum. Kruskal–Wallis test: χ^2=51.9

Pairwise comparisons showed that both ChatGPT and Perplexity AI produced significantly lower RHS values than Gemini 3 Flash and DeepSeek (ChatGPT vs. Gemini: p<0.001; ChatGPT vs. DeepSeek: p<0.001; Perplexity AI vs. DeepSeek: p<0.001; Perplexity AI vs. Gemini: p=0.010). However, only a borderline difference was observed between ChatGPT and Perplexity AI (p=0.05), and Gemini and DeepSeek also did not differ significantly from one another (p>0.05).

RHS scores were also examined according to trauma category, summarized in Table 2. Although small numerical variations were observed, interquartile ranges overlapped considerably, and there were no statistically significant differences in RHS values across trauma categories (p>0.05) (Fig. 1).

To further assess predictors of hallucination severity, a generalized linear model with gamma distribution and log-link function was constructed including chatbot type and trauma category as fixed effects. As summarized in Table 3, chatbot type remained an independent predictor of RHS, with DeepSeek and Gemini demonstrating significantly higher hallucination severity than ChatGPT (p<0.001).

Internal consistency analysis of the seven RHS components demonstrated acceptable to excellent reliability. As shown in Table 4, Cronbach's α was 0.82 for the total dataset. Model-specific α values were 0.68 for ChatGPT, 0.76 for Perplexity AI, 0.82 for Gemini, and 0.88 for DeepSeek.

Table 2. Reference the hallucination score across dental trauma categories

Trauma category	Mean±SD	Median (IQR)
Avulsion	6.28±3.93	7.0 (2.0–10.0)
Intrusion	5.87±3.66	6.0 (4.0–8.5)
Extrusion	5.78±3.51	6.0 (4.0–8.5)
Lateral luxation	5.41±3.47	6.0 (3.0–8.0)
Subluxation	5.00±3.64	5.0 (3.0–7.0)
Uncomplicated crown fracture	5.33±3.29	6.0 (3.0–7.0)
Complicated crown fracture	5.54±3.45	6.0 (3.0–8.0)
Root fracture	5.72±3.68	6.0 (3.0–8.5)
Alveolar fracture	5.83±3.74	6.0 (3.0–9.0)
Post-traumatic pulp necrosis	5.69±3.59	6.0 (3.0–8.0)
p	0.876	

SD: Standard deviation; IQR: Interquartile range; Kruskal–Wallis test: χ^2=4.49

Discussion

This study evaluated the bibliographic reliability of AI-based chatbots when generating scientific references for standardized dental trauma scenarios. Significant inter-model differences in RHS were observed, whereas trauma category did not significantly influence hallucination severity. The null hypothesis was therefore partially rejected. While statistically significant differences were identified between ChatGPT and other systems – particularly Gemini and DeepSeek – no significant differences were observed among Perplexity AI, Gemini, and DeepSeek. These findings indicate that hallucination severity varies across individual LLMs, although such differences are not uniformly distributed across all chatbot architectures.

The generalized linear model further confirmed chatbot type as an independent predictor of RHS values, reinforcing the robustness of the inter-model comparison. In contrast, the trauma category did not emerge as a significant factor, suggesting that reference instability reflects model-specific characteristics rather than clinical scenario complexity. Whether the prompt concerned avulsion, luxation, fracture, or post-traumatic pulp necrosis, hallucination severity remained relatively stable.

Table 3. Generalized linear model (gamma distribution, log-link) assessing predictors of reference hallucination score

Predictor comparison	β (SE)	Exp(β)	95% CI for Exp(β)	z	p
Perplexity AI versus ChatGPT	0.171 (0.087)	1.19	1.00–1.41	1.96	0.050
Gemini versus ChatGPT	0.463 (0.087)	1.59	1.34–1.89	5.31	<0.001
DeepSeek versus ChatGPT	0.608 (0.087)	1.84	1.55–2.18	6.97	<0.001

β=regression coefficient; SE: Standard error; Exp(β): Rate ratio; CI: Confidence interval. Reference: ChatGPT-5.2.

Table 4. Internal consistency of reference hallucination score components across chatbots

Chatbot	Cronbach's α
ChatGPT	0.68
Perplexity AI	0.76
Gemini	0.82
DeepSeek	0.88
Overall	0.82

Cronbach's α was used to assess the internal consistency of reference hallucination score components across chatbots. ≥0.90: excellent, 0.80–0.89: good, 0.70–0.79: acceptable, <0.70: questionable/poor

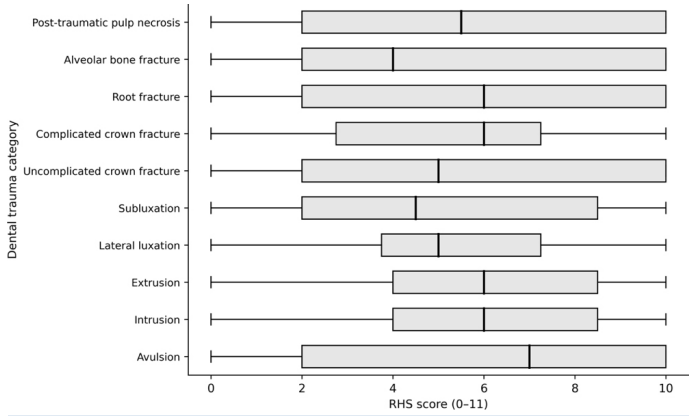


Figure 1. Boxplot distribution of reference hallucination scores across dental trauma categories. The median, interquartile range, and minimum–maximum values are shown for each trauma type.

These findings support emerging evidence that reference hallucination is a model-dependent phenomenon, primarily shaped by system architecture and retrieval mechanisms rather than by content domain.[8,9,13–17] Consistent with prior biomedical AI research, retrieval-augmented systems generated references with fewer inaccuracies than models relying predominantly on generative language prediction. In the present analysis, ChatGPT and Perplexity AI demonstrated significantly lower RHS values compared with Gemini and DeepSeek, suggesting comparatively stronger grounding in indexed databases. These observations align with previous reports indicating that integration of external search functions reduces the likelihood of fabricated bibliographic content.[6,10–12,18–20] Conversely, generative-only models continued to produce plausible yet unverifiable citations, reinforcing concerns raised in earlier medical and dental AI literature.[10–12,21–24]

The study also demonstrated acceptable to good internal consistency of the RHS framework across chatbot outputs, supporting its utility as a structured metric for benchmarking citation reliability. These findings are consistent with the original validation of RHS in medical AI settings.[13] The reproducibility of scoring across systems strengthens the methodological validity of the present comparisons.

Importantly, the present investigation focused exclusively on bibliographic accuracy rather than the clinical correctness of chatbot-generated management recommendations. Therefore, the findings should be interpreted as evidence of variability in citation reliability rather than as an assessment of clinical reasoning quality. In time-sensitive disciplines such as dental traumatology, unreliable or unverifiable citations may complicate rapid evidence verification processes, underscoring the importance of independent source validation. However, the study did not evaluate guideline concordance or treatment accuracy, and such dimensions require dedicated clinical evaluation frameworks.[25–27]

From a broader academic perspective, the results contribute to ongoing discussions regarding the responsible integration of AI technologies into healthcare research and education. Although retrieval-augmented systems performed comparatively better, none achieved complete bibliographic accuracy. This indicates that manual verification remains essential when AI-generated references are used for academic writing or clinical support. Established guideline documents, such as those issued by the IADT,[1–3] continue to represent the authoritative standard for trauma management.

The present study has several strengths, including standardized trauma-based prompts aligned with internationally recognized classifications, a substantial reference sample, and multi-database verification. Nevertheless, limitations should be acknowledged. First, only English-language, general-purpose chatbots were evaluated at a single time point; ongoing model updates may influence performance. Second, the analysis was confined to bibliographic verification and did not examine the clinical validity of AI-generated recommendations. Third, findings cannot be extrapolated to specialized AI systems trained on curated medical datasets.

Future research should incorporate longitudinal designs to assess the stability of reference accuracy across model updates. Comparative studies involving domain-specific or medically trained AI platforms may clarify whether curated training data reduce hallucination rates. In addition, dedicated investigations into clinical guideline concordance are warranted to determine whether bibliographic hallucination correlates with clinical inaccuracy. Expanding analyses to multilingual settings and real-world clinical prompts would further enhance generalizability and practical relevance.

Conclusion

AI-based chatbots showed marked variability in the accuracy of the references they generated for dental trauma scenarios. Retrieval-augmented systems demonstrated comparatively lower hallucination severity; however, none of the evaluated models achieved full bibliographic reliability. Reference hallucination therefore appears to be a model-dependent rather than a context-dependent phenomenon. Because evidence-based guidance is essential for the prognosis of TDIs, inaccurate or unverifiable AI-generated citations may pose challenges for evidence verification processes. However, the present findings do not extend to clinical decision-making accuracy, which requires dedicated investigation. AI chatbots should therefore be regarded as supportive tools rather than independent bibliographic resources, and all AI-generated references should be independently verified before clinical or academic use.

Ethics Committee Approval: Ethics committee approval was not required because the study did not include human subjects, patient information, or biological materials.

Conflict of Interest: None declared.

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

Use of AI for Writing Assistance: During the preparation of this manuscript, the authors used artificial intelligence-based language models solely to improve language clarity, grammar, and overall readability. The AI tools were not used for data analysis, data interpretation, or the generation of scientific content. All conceptualization, study design, data collection, statistical analysis, and interpretation of findings were performed exclusively by the authors. The authors critically reviewed, edited, and take full responsibility for the final content of the manuscript.

Authorship Contributions: Concept: EYM; Design: EYM, MO; Supervision: MO; Materials: EYM, MO; Data collection and/or processing: EYM; Analysis and/or interpretation: EYM; Literature review: EYM; Writing: EYM, MO; Critical review: EYM, MO.

Peer-review: Double blind peer-reviewed.

References

1. 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]
2. 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]
3. 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. Dent Traumatol 2020;36(4):309-13. [CrossRef]
4. Magno MB, Nadelman P, Leite KLDF, Ferreira DM, Pithon MM, Maia LC. Associations and risk factors for dental trauma: a systematic review of systematic reviews. Community Dent Oral Epidemiol 2020;48(6):447-63. [CrossRef]
5. Gao S, Wang X, Xia Z, Zhang H, Yu J, Yang F. Artificial intelligence in dentistry: a narrative review of diagnostic and therapeutic applications. Med Sci Monit 2025;31:e946676. [CrossRef]
6. Liu TY, Lee KH, Mukundan A, Karmakar R, Dhiman H, Wang HC. AI in dentistry: innovations, ethical considerations, and integration barriers. Bioengineering (Basel) 2025;12(9):928. [CrossRef]
7. Ghaffari M, Zhu Y, Shrestha A. A review of advancements of artificial intelligence in dentistry. Dent Rev (Heidelb) 2024;4(2):100081. [CrossRef]
8. 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;55:1-38. [CrossRef]
9. Rawte V, Sheth A, Das A. A survey of hallucination in large language models. arXiv 2023.
10. Munaye YY, Admass W, Belayneh Y, Molla A, Asmare M. ChatGPT in education: a systematic review on opportunities, challenges, and future directions. Algorithms 2025;18(6):352. [CrossRef]
11. Kotsis KT. Scientific authorship in the age of AI: challenges for editors and institutions. Eur J Manag Econ Bus 2025;2(6):209-16. [CrossRef]
12. Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LTJ, Li YCJ. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. J Biomed Sci 2025;32(1):45. [CrossRef]
13. 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(1):e54345. [CrossRef]
14. Gorelik AJ, Li M, Hahne J, Wang J, Ren Y, Yang L, et al. Ethics of AI in healthcare: a scoping review demonstrating applicability of a foundational framework. Front Digit Health 2025;7:1662642. [CrossRef]
15. 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;141:819-24. [CrossRef]
16. Kandemir M, Sarıbaş EE. Comparative performance of large language models in answering periodontology questions from the Turkish Dental Specialty Examination: a cross-sectional study on accuracy and coverage. BMC Oral Health 2025;25:1804. [CrossRef]
17. Künzle P, Paris S. Performance of large language artificial intelligence models on solving restorative dentistry and endodontics student assessments. Clin Oral Investig 2024;28:575. [CrossRef]

18. Turan Gökduman C, Arılı Öztürk E, Aktaş Ş, Çanakçi BC. Comparison of chatbots' accuracy in endodontics questions in dentistry specialization exam in Türkiye: ChatGPT-4o, Gemini Advanced, Copilot, and Claude. BMC Oral Health 2025;26(1):28. [CrossRef]
19. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health 2023;5:e179-81. [CrossRef]
20. Zybaczynska J, Norris M, Modi S, Brennan J, Jhaveri P, Craig TJ, et al. Artificial intelligence-generated scientific literature: a critical appraisal. J Allergy Clin Immunol Pract 2024;12:106-10. [CrossRef]
21. Kung TH, Cheatham M, Medenilla A, Sillos C, Leon LD, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health 2023;2:e0000198. [CrossRef]
22. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords for radiologists. Radiology 2023;307:e230163. [CrossRef]
23. Alyasiri OM, Salman AM, Akhtom DA, Salisu S. ChatGPT revisited: using ChatGPT-4 for finding references and editing language in medical scientific articles. J Stomatol Oral Maxillofac Surg 2024;125:101842. [CrossRef]
24. Goktas P, Grzybowski A. Assessing the impact of ChatGPT in dermatology: a comprehensive rapid review. J Clin Med 2024;13:5909. [CrossRef]
25. Pham T. Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use. R Soc Open Sci 2025;12(5):241873. [CrossRef]
26. Küçük Keleş Ö, 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 2026;30:26. [CrossRef]
27. Gonzalez-Valenzuela RE, Mettes P, Loos BG, Marquering H, Berkhout E. Accuracy of deep learning-based AI models for early caries lesion detection: the influence of annotation quality and reference choice. Clin Oral Investig 2025;29:598. [CrossRef]

Appendix 1

Prompt 1 :

A 10-year-old patient presents with an avulsed permanent maxillary incisor. According to current evidence-based guidelines, describe the recommended immediate management, splinting protocol, endodontic considerations, and prognosis. Please provide 10 supporting references.

Include the following information for each article:

1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:

1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

Prompt 2:

A 9-year-old patient presents with an intruded permanent maxillary central incisor. According to current evidence-based guidelines, describe the recommended management approach, follow-up protocol, possible complications, and prognosis. Please provide 10 supporting references.

Include the following information for each article:

1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:

1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

Prompt 3:

A patient presents with an extruded permanent tooth following dental trauma. According to current evidence-based guidelines, describe the emergency management, repositioning technique, splinting protocol, and prognosis. Please provide 10 supporting references.

Include the following information for each article:

1. Article title.
2. Author(s).

3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:

1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

Prompt 4:

A patient presents with a laterally luxated permanent tooth with alveolar socket displacement. According to current evidence-based guidelines, describe the recommended repositioning, stabilization, endodontic considerations, and prognosis. Please provide 10 supporting references.

Include the following information for each article:

1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:

1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

Prompt 5:

A patient presents with subluxation of a permanent tooth following trauma. According to current evidence-based guidelines, describe the recommended clinical management, monitoring protocol, possible complications, and prognosis. Please provide 10 supporting references.

Include the following information for each article:

1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:

1. Begin by searching PubMed.

2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

Prompt 6:
A patient presents with an uncomplicated enamel–dentin crown fracture without pulp exposure. According to current evidence-based guidelines, describe the recommended treatment approach, restorative options, follow-up protocol, and prognosis. Please provide 10 supporting references.
Include the following information for each article:
1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:
1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

Prompt 7:
A patient presents with a complicated crown fracture with pulp exposure. According to current evidence-based guidelines, describe the recommended vital pulp therapy or endodontic treatment approach, restorative procedures, follow-up, and prognosis. Please provide 10 supporting references.
Include the following information for each article:
1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:
1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

Prompt 8:
A patient presents with a horizontal root fracture of a permanent tooth. According to current evidence-based

guidelines, describe the diagnostic procedures, repositioning and splinting protocol, endodontic considerations, and prognosis. Please provide 10 supporting references.
Include the following information for each article:
1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:
1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

Prompt 9:
A patient presents with an alveolar bone fracture associated with dental trauma. According to current evidence-based guidelines, describe the stabilization procedure, splinting duration, follow-up protocol, possible complications, and prognosis. Please provide 10 supporting references.
Include the following information for each article:
1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:
1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

Prompt 10:
A patient develops post-traumatic pulp necrosis following dental trauma. According to current evidence-based guidelines, describe the indications for endodontic intervention, treatment protocol, follow-up recommendations, and prognosis. Please provide 10 supporting references.
Include the following information for each article:
1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.

6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:

1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.