ACTA SCIENTIFIC DENTAL SCIENCES

Volume 9 Issue 10 October 2025

Research Article

Comparative Evaluation of Artificial Intelligence Models for Traumatic Dental Injuries Based on Clinical Guideline Adherence

Okan Turgut, H Melike Bayram* and Emre Bayram

Associate Professor, Tokat Gaziosmanpasa University, Faculty of Dentistry, Department of Endodontics, Tokat, Turkiye

***Corresponding Author:** Huda Melike Bayram, Associate Professor, Tokat Gaziosmanpasa University, Faculty of Dentistry, Department of Endodontics, Tokat, Turkiye.

Received: August 25, 2025
Published: September 10, 2025

Abstract

Objective: This study aimed to evaluate the performance of three large language models (LLMs)-Grok, ChatGPT, and DeepSeek-in managing traumatic dental injuries (TDIs) based on their alignment with the International Association of Dental Traumatology (IADT) 2020 clinical guidelines.

Materials and Methods: Twenty open-ended prompts were constructed to reflect real-life TDI scenarios, aligned with the 2020 IADT guidelines. Each model was queried once per prompt with no re-prompting or interaction refinement. Responses were evaluated by a trained rater using a five-criteria rubric: scientific accuracy, reliability of information, comprehensibility, level of detail, and clinical applicability. Scoring was performed using a 3-point ordinal scale. One-way ANOVA and post-hoc comparisons were applied for statistical analysis.

Results: Grok outperformed both ChatGPT and DeepSeek in scientific accuracy, detail level, and information reliability ($p < 0.001$). ChatGPT and DeepSeek showed relatively higher scores in comprehensibility ($p = 0.007$). For clinical applicability, only the Grok–DeepSeek comparison was statistically significant ($p = 0.016$). Total score comparisons were substantial across all model pairs ($p < 0.001$).

Conclusion: Large language models exhibit distinct strengths across clinical performance metrics. Grok appears more suitable for guideline-based clinical decision support in TDI management, whereas ChatGPT and DeepSeek may be better suited for educational and communicative purposes. Purpose-driven model selection and continuous performance monitoring are recommended for safe and effective clinical integration.

Keywords: Artificial Intelligence; Large Language Models; Traumatic Dental Injuries; Clinical Decision Support; Guideline Adherence; IADT Guidelines

Abbreviations

AI: Artificial Intelligence; LLM: Large Language Models; IADT: International Association of Dental Traumatology; TDIs: Traumatic Dental Injuries TDIs); SD: Standard Deviation

Introduction

Artificial intelligence (AI) applications have rapidly proliferated across dentistry, spanning clinical care and education, with a marked acceleration after 2020 as interdisciplinary use cases expanded [1,2]. This growth reflects not only algorithmic advances but also the maturation of data ecosystems and increased academic and institutional awareness[1],[2]. Generative AI and large language models (LLMs) are reshaping text-centric processes such as patient communication, guideline explanation, and scholarly writing, necessitating planned and measurable integration into dental curricula [1]. Curricular updates should encompass foundational concepts, ethics and governance, performance metrics, and alignment of outputs with principles of evidence-based dentistry [1].

Systematic reviews indicate that AI can deliver meaningful accuracy and decision support for diagnosis, risk prediction, and treatment planning across multiple dental disciplines [2].

However, much of the literature is image-centric, and rigorous prospective validation is required to ensure generalizability across data types and clinical contexts [2]. For clinical deployment, the selection and purchasing of AI systems should proceed within an institutional governance frame that includes independent performance verification, workflow fit, information-technology requirements, total ownership cost, and patient safety and quality monitoring [3]. Such governance includes continuous performance surveillance, assessment of potential clinical impact of errors, and mechanisms to restrict use or implement improvements when needed [3].

AI's impact in dentistry extends beyond LLMs; for example, stereolithography-based customized surgical guides have been evaluated with artificial neural networks and have achieved high classification accuracy in dimensional-fit assessments [4].

These exemplars show that when design, validation, and monitoring are appropriate, AI can yield clinically robust outputs and offer methodological benchmarks for text-based decision support [4].

Management of traumatic dental injuries (TDIs) is structured by standardized protocols and decision trees, with the International Association of Dental Traumatology (IADT) 2020 guidelines recognized as a principal reference [5,6]. Because domains such as avulsion involve time-critical decisions and follow-up algorithms, producing information faithful to these guidelines is crucial to clinical outcomes [5,6]. Comparative, guideline-anchored evaluations of LLMs remain limited, and the literature emphasizes the need for pre-deployment validation and institutional oversight [1,3]. Accordingly, multi-criteria assessments that evaluate scientific accuracy, reliability of information, comprehensibility, level of detail, and clinical applicability are needed to characterize model performance [1,3]. The 2020 IADT guidelines provide standardized, evidence-based decision pathways that form a clinically meaningful reference standard for benchmarking LLM outputs [5,6].

Measuring the degree to which LLM outputs align with these pathways can improve clinical decision support safety and validate dental education instructional content [5,6].

Evaluation should not be confined to aggregate accuracy or output quality; consistent with purchasing and quality-assurance guidance, it should adopt context-appropriate multidimensional metrics embedded within governance structures and supported by ongoing monitoring [3]. This approach establishes an evidence base aligned with responsible AI use in curricula and service delivery at the institutional level [1,3].

This study aimed to compare ChatGPT, Grok, and DeepSeek in generating TDI management information against the IADT 2020 guidelines using a structured content-analysis framework across scientific accuracy, reliability of information, comprehensibility, level of detail, and clinical applicability. There are no differences among Grok, ChatGPT, and DeepSeek in mean scores across the five criteria and the total score.

Materials and Methods

This study is a cross-sectional, comparative, protocol-driven content analysis evaluating the performance of three large language models (LLMs)-ChatGPT, Grok, and DeepSeek-in generating knowledge on managing traumatic dental injuries (TDIs). This study did not involve human participants or patient data and was deemed exempt from IRB review according to institutional policy. Before data collection, the study scope, question set, scoring criteria, and statistical plan were specified in writing. The construction of prompts and the adjudication of correctness were based on the International Association of Dental Traumatology (IADT) 2020 guidelines; the papers on management of fractures and luxations of permanent teeth and management of avulsion of permanent teeth were adopted as the normative reference [5].

The IADT 2020 scope was systematically reviewed, and clinically critical subtopics were selected: avulsion, types of luxation (concussion, subluxation, extrusive/lateral/intrusive), complicated/uncomplicated crown fractures, root fractures, alveolar fractures, and post-trauma follow-up protocols. For each subtopic, open-ended, guideline-aligned prompts reflecting real-world scenarios were drafted. Topic distribution was balanced to yield a final bank of 20 prompts. Prompts combined a brief case vignette with an open question; a pilot readability check was performed to minimize ambiguity and double meanings.

Each of the twenty prompts was administered once per model in separate sessions. No re-prompts, guiding sub-prompts, plugins, or browser tools were used. All interactions were text-only; no images or files were provided. The model was captured and exported verbatim for every prompt and entered into the evaluation workflow.

Responses were scored on a pre-specified five-criterion, 1–3 point ordinal scale (higher scores indicate better performance):

- **Scientific accuracy:** Consistency with IADT 2020; absence of critical errors or clinically material omissions.
- **Reliability of information:** Appropriate citation or guideline referencing, presence of warnings/precautions, and communication of uncertainty.
- **Comprehensibility:** Fluency, terminological correctness, and clinical traceability of the narrative.
- **Level of detail:** Completeness of algorithms/steps; adequacy of timing, duration/dose, and follow-up specifications.

Clinical applicability: Capacity to guide decision-making; practicable, safe, action-oriented recommendations.

Anchors: 1 = inadequate/misaligned (contradicts the guideline or omits critical steps); 2 = partially adequate (core facts present but key details/alerts missing); 3 = fully aligned and detailed (clearly mirrors the guideline algorithm and includes necessary cautions). For each prompt, scores across the five criteria were summed to a total score. Descriptive reporting included mean, standard deviation (SD), and min–max for each criterion and the total.

Scoring was performed by one trained rater familiar with the study protocol and rubric. To support decision consistency, the rater consulted the IADT texts during scoring using bookmarks/keywords for rapid access. Blinding (masking model identity) was not implemented; on-the-spot verification against the guidelines was conducted at critical decision points to mitigate potential bias. This study did not calculate inter-rater reliability (e.g., Kappa/ICC); a second independent rater is planned for reliability confirmation in future replications.

For each prompt × model cell (20 × 3), a five-element score vector was recorded in a spreadsheet (60 cells; 300 criterion scores total). The de-identified prompt list, rubric, and raw scoring matrix are available in the online repository/upon reasonable request. No post-hoc edits or redactions were applied to responses or scoring records.

To reduce selective reporting, the prompt set was locked before data collection, a single-response rule was enforced for each prompt–model interaction, and model outputs were not edited. The scoring rubric and anchors were finalized before rating commenced and used as the operative reference.

Statistical analysis

Between-model comparisons of mean criterion scores (comprehensibility, reliability of information, scientific accuracy, level of detail, clinical applicability) and total score were conducted using one-way ANOVA. Normality was assessed using the Shapiro–Wilk test, and homogeneity of variances was assessed using the Levene test. When the omnibus test was significant at $\alpha = 0.05$, Tukey's HSD was used for pairwise comparisons under homogeneity and Tamhane's T2 when the assumption was violated. All statistical analyses were performed in IBM SPSS Statistics, Version 26 (IBM Corp., Armonk, NY, USA), and a p-value < 0.05 was considered statistically significant.

Results and Discussion

Table 1 presents the mean ± SD and minimum–maximum values for five evaluation criteria and the total score across the three large language models (Grok, ChatGPT, and DeepSeek). One-way ANOVA revealed statistically significant differences among the models for all outcomes: comprehensibility ($p = 0.007$), reliability of information ($p < 0.001$), scientific accuracy ($p < 0.001$), level of detail ($p < 0.001$), clinical applicability ($p = 0.016$), and total score ($p < 0.001$).

Criterion	Grok (Mean ± SD)	ChatGPT (Mean ± SD)	DeepSeek (Mean ± SD)	p-value	Post-hoc Comparison
Comprehensibility	2.15 ± 0.49	2.60 ± 0.50	2.60 ± 0.50	0.007*	1-2;1-3
Reliability of Information	2.75 ± 0.44	2.30 ± 0.47	1.80 ± 0.70	<0.001*	1-2;1-3;2-3
Scientific Accuracy	2.90 ± 0.31	2.40 ± 0.60	2.00 ± 0.46	<0.001*	1-2;1-3
Level of Detail	2.90 ± 0.31	1.50 ± 0.51	1.60 ± 0.60	<0.001*	1-2;1-3
Clinical Applicability	1.95 ± 0.39	1.55 ± 0.76	1.40 ± 0.60	0.016*	1-3
Total Score	12.65 ± 1.04	10.35 ± 1.76	8.10 ± 1.29	<0.001*	1-2;1-3;2-3

Table 1: Mean ± SD; p-values based on one-way ANOVA. Post-hoc tests (Tukey's HSD or Tamhane's T2) applied where appropriate. $p < 0.05$ considered statistically significant.

Overall, Grok consistently achieved the highest scores in scientific accuracy, level of detail, and reliability of information. In contrast, ChatGPT and DeepSeek demonstrated relative advantages in comprehensibility. Grok again received the highest score for clinical applicability among the three models. Post-hoc analyses further clarified the differences between models. Regarding comprehensibility, Grok scored significantly lower than ChatGPT and DeepSeek, while the difference between ChatGPT and DeepSeek was not statistically significant. All pairwise comparisons reached statistical significance for information reliability, with the performance ranking as Grok > ChatGPT > DeepSeek. Grok also significantly outperformed both models in scientific accuracy and level of detail, whereas no significant difference was found between ChatGPT and DeepSeek for these criteria. Only the comparison between Grok and DeepSeek was statistically substantial for clinical applicability. All pairwise comparisons for the total score were statistically significant, confirming the consistent superiority of Grok in multidimensional performance evaluation.

This study compares three LLMs-Grok, ChatGPT, and DeepSeek-using a structured multi-criteria framework based on the IADT 2020 guidelines for TDI management. The results reveal significant differences across all evaluated dimensions, with Grok demonstrating superior performance in scientific accuracy, level of detail, and reliability of information. At the same time, ChatGPT and DeepSeek showed relative advantages in comprehensibility. The high performance of Grok may be attributed to its enhanced capacity for clinical context alignment and medical content retrieval. These findings align with recent systematic reviews of LLMs in healthcare settings, which highlight the variation in performance across clinical tasks depending on language model architecture and prompt specificity [7,8].This suggests that Grok may employ a more specialized training architecture or retrieval mechanism, contributing to its consistency in protocol adherence and data accuracy. Grok's superiority in accuracy and detail supports prior results from Hartman et al., who found that domain-specific LLMs yielded safer and more complete emergency handoff notes compared to general-purpose models [9].Similarly, Shool et al. and Wang et al. reported that LLM performance in clinical queries is heavily influenced by training specificity and guideline integration [10,11].

The relatively higher comprehensibility scores of ChatGPT and DeepSeek point toward their superior fluency and language generation capabilities, making them potentially more useful in patient education, explanatory communication, and non-specialist training. However, this linguistic advantage should be balanced against the need for guideline-conformant and clinically actionable content, especially in time-sensitive domains like dental trauma management. For instance, incorrect timing or storage recommendations in avulsion cases can result in irreversible pulp necrosis [5,6]. Conversely, ChatGPT and DeepSeek showed stronger performance in comprehensibility, consistent with findings by Sivaramakrishnan et al., who demonstrated that LLMs generated more accessible patient education materials than conventional sources [12]. Ozdemir., *et al.* corroborated this by showing that AI-driven chatbot responses in dentistry were more readable, though occasionally lacked depth and source attribution [13].

In clinical decision support, Grok's higher detail level and alignment with IADT algorithms reinforce its potential utility in generating safe and structured recommendations. The critical steps of luxation diagnosis, replantation timing, and splint duration require exact replication of algorithmic pathways, where Grok outperformed its counterparts in providing unambiguous, guideline-compliant responses. The variation in performance across models underscores the need for context-specific model deployment in clinical settings. For guideline-adherent decision support in acute dental trauma, LLMs like Grok may offer greater fidelity, while ChatGPT or DeepSeek may be more suitable for educational or patient-facing applications. This duality has been highlighted in broader reviews of medical LLMs, where model purpose alignment is considered critical for safe integration [14,15].

These results align with existing literature that distinguishes the capabilities of general-purpose and medically optimized LLMs. In systematic reviews, domain-specific models tend to produce higher factual accuracy but require continuous validation [11]. The current findings reinforce this by demonstrating significant inter-model variance even within a single clinical domain, underlining the importance of model selection tailored to clinical goals. From a deployment standpoint, structured governance protocols should guide practical integration of LLMs into clinical workflows. Filice et al. propose a multidimensional checklist that includes independent performance validation, workflow integration, IT compatibility, and bias monitoring [3].Our findings support the relevance of this approach: Grok's performance may justify more direct clinical application, while other models might be confined to patient-facing or low-risk educational contexts.

As noted earlier, clinical integration of LLMs requires not only attention to output quality but also adherence to governance frameworks that ensure safety, reliability, and ongoing performance monitoring. Tam et al. proposed the QUEST framework, which incorporates quality, safety, and trust domains to evaluate LLM outputs in real-time practice [14]. Asgari., *et al.* further emphasized the importance of hallucination monitoring, especially in scenarios where decision-making risks are high [16]. From an educational and ethical standpoint, models that prioritize readability must also maintain guideline fidelity. This balance was discussed by Haltaufderheide & Ranisch, who cautioned that generative AI, while beneficial for learner engagement, could introduce epistemic drift if not continuously aligned with validated medical standards [15].

Furthermore, item-level analyses showed internal consistency with the aggregated performance pattern: Grok frequently received higher scores in accuracy, detail, and reliability, while ChatGPT and DeepSeek were more prominent in comprehensibility. This suggests that model-specific strengths are not scenario-bound but inherent to their design and optimization.

It is important to emphasize that while Grok excelled in clinical metrics, its lower comprehensibility could limit its standalone use in patient communication. This supports a hybrid deployment model where high-accuracy models support diagnostic processes and fluent models assist in communication and education.

The current study contributes uniquely to dental literature by being one of the first to systematically benchmark LLMs against a recognized clinical standard (IADT 2020). Unlike prior evaluations that lacked structured rubrics or focused on general prompts, this study applies a replicable scoring methodology across evidence-based clinical domains.

Nevertheless, several limitations must be acknowledged. Limitations of this study include reliance on single responses per model and evaluation by a single rater. Although this design ensured consistency, it may underrepresent inter-rater variability and prompt sensitivity. Future studies should incorporate blinded, multi-rater designs and explore task complexity and prompt phrasing as potential moderators.

Conclusion

This study demonstrates that LLMs display domain-specific strengths in managing dental trauma cases. Grok's superior guideline alignment suggests utility in decision support, while ChatGPT and DeepSeek may be best suited for patient education. For safe deployment, model selection should be purpose-driven and supported by institutional oversight, real-time monitoring, and continuous revalidation protocols.

Acknowledgements

None.

Conflict of Interest

The authors declare no conflict of interest.

Bibliography

1. Thurzo A., *et al.* "Impact of artificial intelligence on dental education: a review and guide for curriculum update". *Education Science* 13.1 (2023): 150.

2. Ahmed N., *et al*. "Artificial intelligence techniques: analysis, application, and outcome in dentistry-a systematic review". *BioMed Research International* 1.3 (2021): 9751564.

3. Filice RW., *et al.* "Evaluating artificial intelligence systems to guide purchasing decisions". *Journal of the American College of Radiology* 17.11 (2020): 1405-1409.

4. Türker H., *et al.* "Fabrication of Customized dental guide by stereolithography method and evaluation of dimensional accuracy with artificial neural networks". *Journal of the Mechanical Behavior of Biomedical Materials* 126.3 (2022): 105071.

5. Bourguignon C., *et al.* "International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: 1. Fractures and luxations". *Dental Traumatology* 36.4 (2020): 314-330.

6. Day PF., *et al.* "International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: 3. Injuries in the primary dentition". *Dental Traumatology* 36.4 (2020): 343-359.

7. Meng X., *et al*. "The application of large language models in medicine: A scoping review". *iScience* 27.5 (2024): 109713.

8. Alkalbani AM., *et al.* "A Systematic Review of Large Language Models in Medical Specialties: Applications, Challenges and Future Directions". (2025).

9. Hartman V., *et al.* "Developing and evaluating large language model–generated emergency medicine handoff notes". *JAMA Network Open* 7.12 (2024): e2448723-e2448723.

10. Shool S., *et al*. "A systematic review of large language model (LLM) evaluations in clinical medicine". *BMC Medical Informatics and Decision Making* 25.1 (2025): 117.

11. Wang L., *et al.* "Accuracy of large language models when answering clinical research questions: Systematic review and network meta-analysis". *Journal of Medical Internet Research* 27.2 (2025): e64486.

12. Sivaramakrishnan G., *et al.* "Assessing the power of AI: a comparative evaluation of large language models in generating patient education materials in dentistry". *BDJ Open* 11.1 (2025): 59.

13. Ozdemir ZM., *et al.* "Evaluating the Accuracy, Reliability, Consistency, and Readability of Different Large Language Models in Restorative Dentistry". *Journal of Esthetic and Restorative Dentistry* (2025).

14. Tam TYC., *et al.* "A framework for human evaluation of large language models in healthcare derived from literature review". *NPJ Digital Medicine* 7.1 (2024): 258.

15. Haltaufderheide J., *et al*. "The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs)". *NPJ Digital Medicine* 7.1 (2024): 183.

16. Asgari E., *et al.* "A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation". *NPJ Digital Medicine* 8.1 (2025): 274.