Can AI Make the Right Call? Large Language Models in Pediatric Dental Trauma Management

Sadaf Taheri, DDS ; Rata Rokhshad, DDS ; Sana Baghizadeh, DDS; Nika Azari; Arad Rokhshad; Samah Omar, DDS, MSD
Loma Linda School of Dentistry, Department of Pediatric Dentistry

LOMA LINDA UNIVERSITY
School of Dentistry

Background & Objectives

Traumatic dental injuries are common in children and require timely, accurate management to prevent long-term complications.[1,2] The International Association of Dental Traumatology (IADT) provides evidence-based guidelines for their diagnosis and treatment. [2]

With the rise of large language models (LLMs) in clinical decision support, their accuracy in dental trauma assessment and guideline-based management remains unclear.[3]

The objectives of this study are to:

- Evaluate and compare the performance of five large language models in responding to dental trauma case scenarios.
- Assess their accuracy in diagnosis, treatment planning, and prognostic evaluation based on IADT guidelines.

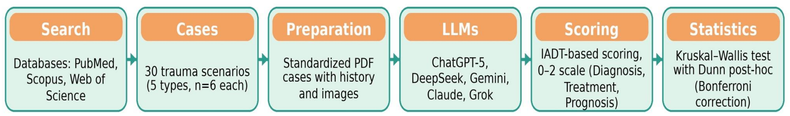

Figure 1: Study Design Overview

Methods

- This cross-sectional prospective study evaluated the response quality of five large language models (LLMs) in dental trauma assessment. The study was approved by LLU IRB (#5250319).
- Thirty clinical dental trauma scenarios were developed from published case reports and radiographs extracted from patient's records. The cases included five trauma categories: avulsion, extrusion, intrusion, lateral luxation, and crown/root fractures (n = 6 each).
- Each case included patient history and corresponding clinical or radiographic images and was compiled into a standardized PDF case file.
- Five LLMs were evaluated: ChatGPT-5, DeepSeek, Gemini, Claude, and Grok. Each model received identical prompts and case files; chat history and browser cache were cleared prior to each evaluation to minimize contextual bias.

  Model responses were assessed across three clinical domains:
  - Diagnosis
  - Treatment planning
  - Prognosis and follow-up recommendations.
- After creating a key for all cases, responses were graded using a 0–2 scoring scale based on IADT guidelines by a pediatric dentistry resident and a board-certified pediatric dentist.
- Statistical comparisons between models were performed using Kruskal–Wallis tests with Dunn's post-hoc analysis and Bonferroni correction and the significance level was set at $P < 0.05$. (Figure 1)

Results

- Overall comparison showed significant differences between AI models in their performance(Kruskal–Wallis H = 11.24, P = .024). ChatGPT-5 achieved the highest overall accuracy (75.0%), closely followed by DeepSeek (74.4%). Other models demonstrated lower performance: Gemini (66.7%), Claude (65.6%), and Grok (64.4%). (Table 1)
- No significant differences were observed among the model's diagnosis accuracy (P = .314) or treatment planning (P = .158) (Table.1). Significant differences emerged in prognostic evaluation (P< .001), where ChatGPT-5 provided more comprehensive guideline-based responses.
- Trauma-type analysis revealed significant differences only in intrusion injuries P = .012). (Figure 3), while the lowest accuracy was observed in lateral luxation cases, although not significant statistically. (Figure 3)

Model	Diagnosis	Treatment	Prognosis	Overall
ChatGPT-5	68.3%	70.0%	86.7%	75.0%
DeepSeek	80.0%	68.3%	75.0%	74.4%
Gemini	73.5%	58.3%	68.3%	66.7%
Claude	78.5%	53.3%	65.0%	65.6%
Grok	76.5%	61.7%	55.0%	64.4%
P-value	*0.314*	*0.158*	*<0.001**	

Table 1. Performance of Large Language Models Across Three Clinical Domains
* Statistically significant results

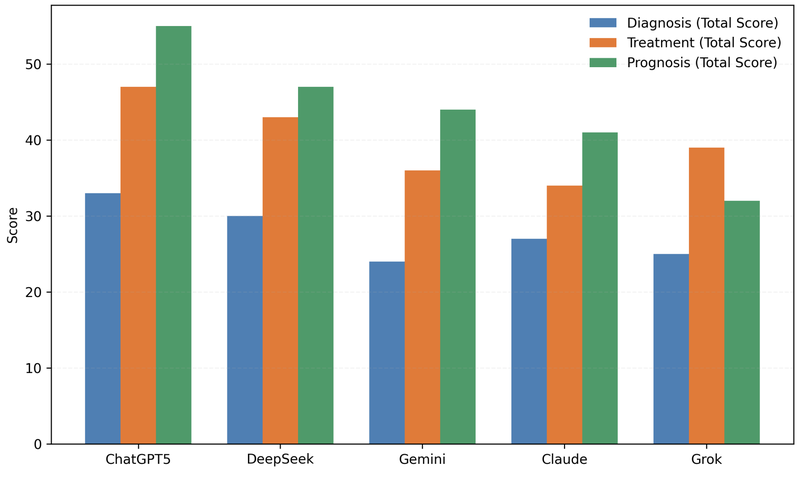

Figure 2. Performance of LLMs across clinical domains.

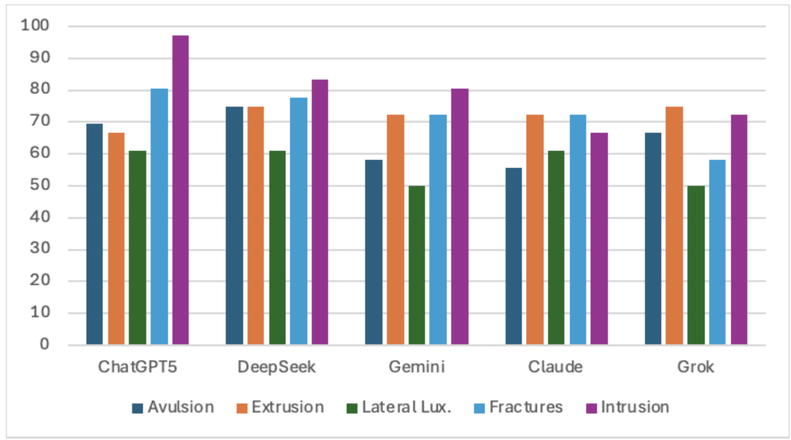

Figure 3. Accuracy of LLMs across dental trauma types.

Discussion & Conclusion

- LLMs performed similarly in diagnosis and treatment but differed in prognostic reasoning, with ChatGPT-5 showing the most consistent guideline-concordant performance. Differences were most evident in intrusion injuries, suggesting variability in handling complex scenarios.
- Compared to Bubna et al., where a purpose-built dental trauma model achieved near-perfect accuracy, our findings suggest that general-purpose LLMs lack the structured, rule-based framework required for consistent guideline-driven decision-making. [4] In contrast to Johnson et al., who reported Claude as the top-performing model, our results suggest that performance depends on task type, with complex, case-based reasoning and radiographic interpretation favoring different models than knowledge-based assessments.[5]
- We conclude that multimodal capabilities enhance clinical interpretation, but limitations remain in complex decision-making. These findings highlight the need for cautious integration of LLMs into pediatric clinical practice, where performance is task-dependent, especially in complex prognostic scenarios.

References

1. Levin L, Day PF, Hicks L, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: General introduction. Dent Traumatol. 2020;36(4):309–13.
2. Day PF, Flores MT, O'Connell AC, 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.
3. Rokhshad R, Fadul M, Zhai G, et al. A Comparative Analysis of Responses of Artificial Intelligence Chatbots in Special Needs Dentistry. Pediatr. Dent. 2024;46(5):337–44.
4. Bubna DP, Felipe de Jesus Freitas P, Ferraz AX, et al. Dental Trauma Evo – Development of an Artificial Intelligence-powered Chatbot to Support Professional Management of Dental Trauma. J Endod. 2025;51(9):1229–34.
5. Johnson AJ, Singh TK, Gupta A, et al. Evaluation of validity and reliability of Chatbots as public sources of information on dental trauma. Dent Traumatol. 2025;41(2):187–93.