RESEARCH PAPER

Generative Artificial Intelligence in Pediatric Dental Pain Assessment: Comparison With Pediatric Dental Expertise and Implications for Clinical Decision Support

[1]Swarnima Bhatnagar, [2]Dinesh Kumar
Postgraduate student , Department of Pediatric and preventive dentistry,, Saveetha Dental College and Hospitals,, Saveetha Institute of Medical and Technical Sciences,, Saveetha University, Chennai, India
Email id - 152311003.sdc@saveetha.com
Senior lecturer, Department of Pediatric and preventive dentistry, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India, **Email id - dineshkumarb.sdc@saveetha.com31-6982**
Corresponding Author
Postgraduate student , Department of Pediatric and preventive dentistry,, Saveetha Dental College and Hospitals,, Saveetha Institute of Medical and Technical Sciences,, Saveetha University, Chennai, India
Email id - 152311003.sdc@saveetha.com

Abstract

Background: Accurate pain assessment is fundamental to diagnosis, treatment planning, and clinical decision-making in pediatric dentistry. However, assessment of dental pain in children can be challenging because pain expression is influenced by developmental, emotional, behavioral, and communication-related factors. With the increasing application of artificial intelligence (AI) in healthcare, generative AI chatbots have emerged as potential clinical decision-support tools; however, their diagnostic performance in pediatric dental pain assessment requires evaluation against professional clinical judgment.

Aim: To compare the diagnostic performance of pediatric dentists and generative AI chatbots—ChatGPT-5, Gemini, Claude, and DeepSeek—in the assessment of standardized pediatric dental pain scenarios.

Materials and Methods: A cross-sectional comparative observational study was conducted using a validated questionnaire comprising **30 standardized pediatric dental pain scenarios**. Twenty-five pediatric dentists and four AI chatbot models—ChatGPT-5, Gemini, Claude, and DeepSeek—evaluated the clinical scenarios. Responses were assessed for probable diagnosis, pain severity, urgency of treatment, and recommended management. Diagnostic performance was evaluated using overall accuracy, sensitivity, specificity, and agreement measures. The performance of pediatric dentists was compared with individual AI models, and inter-model differences were evaluated.

Results: AI chatbots demonstrated promising performance in pediatric dental pain assessment, with differences observed among the evaluated models. ChatGPT demonstrated the highest diagnostic performance among the AI chatbots. Performance varied according to the complexity of the clinical scenario, with AI showing comparatively greater limitations in cases requiring contextual reasoning, interpretation of behavioral factors, and clinical judgment. The study also demonstrated high reproducibility of AI responses across repeated assessments.

Conclusion: Generative AI chatbots demonstrate considerable potential as decision-support tools for pediatric dental pain assessment. Among the evaluated AI models, ChatGPT showed the strongest diagnostic performance. Nevertheless, limitations in interpreting complex clinical and behavioral information highlight the continued importance of professional clinical judgment. AI-based systems may therefore complement pediatric dentists in preliminary assessment, education, and clinical decision-making but should not be considered replacements for trained clinicians. Future multimodal AI systems incorporating clinical findings, radiographs, photographs, patient history, and behavioral information may further improve diagnostic performance.

Keywords: Artificial intelligence; ChatGPT; Gemini; Claude; DeepSeek; Pediatric dentistry; Dental pain; Diagnostic accuracy; Clinical decision-making.

How to cite this article: Swarnima Bhatnagar, Dinesh Kumar. Generative Artificial Intelligence in Pediatric Dental Pain Assessment: Comparison With Pediatric Dental Expertise and Implications for Clinical Decision Support. Int J Drug Deliv Technol. 2026;16(78s):388-394. DOI: 10.25258/ijddt.16.78s.47

Introduction

Accurate assessment of pediatric dental pain remains a clinical challenge because children may have limited verbal communication abilities and often require specialized behavioral management [1,2]. Clinicians must integrate clinical findings, radiographic evidence, and patient- or caregiver-reported symptoms to establish an appropriate diagnosis and management plan; however, the subjective nature of pain and variation in children's expression of symptoms can contribute to diagnostic variability and inconsistencies in treatment planning. As large language models emerge as potential clinical decision-support tools, they offer a novel, although not yet fully validated, approach to organizing and interpreting multifaceted clinical information [3,4]. Although these models have demonstrated promising performance in several oral-health applications [5,6], the literature lacks systematic comparative benchmarking of LLM performance in pediatric dental pain assessment against the clinical judgment of specialized pediatric dentists [7]. This study therefore aims to evaluate the diagnostic accuracy, sensitivity, and specificity of four prominent LLMs and to assess their potential role as adjuncts to pediatric clinical reasoning [8]. Despite their reported utility in generating clinical information, LLMs may struggle with subtle dental abnormalities and with incorporating patient-specific chief complaints into diagnostic reasoning [9,10]. Concerns also remain regarding misinformation, data bias, and the propagation of clinically consequential errors in complex pediatric cases [11]. Accordingly, comparison with human expertise is necessary to assess the reliability of these systems in pediatric dentistry, where inaccurate decisions may contribute to preventable long-term complications [12,13]. Benchmarking LLMs against specialized pediatric dental expertise may help determine whether current generative AI systems meet the diagnostic standards required for safe integration into pediatric practice [14]. This evaluation is particularly important because current models may have difficulty resolving the diagnostic discrepancies characteristic of pediatric cases, potentially limiting the safety of routine clinical use [15]. It is also necessary to characterize performance differences among platforms, as variations in model architecture and training may result in inconsistent diagnostic outputs when identical clinical information is provided [16]. Although emerging research indicates that LLMs may improve clinical efficiency, their diagnostic performance remains variable and inadequate to support autonomous decision-making in high-stakes pediatric settings [17–19]. Previous studies have reported substantial performance variability among models, with some failing to meet the proficiency threshold required for national licensing examinations [20,21]. Empirical validation is therefore needed to determine whether these technologies can function reliably as decision-support tools without compromising the accuracy and contextual judgment provided by human clinicians [22,23]. By quantifying performance differences between specialized pediatric clinicians and LLMs, this study addresses the need for benchmarks that extend beyond overall accuracy to evaluate clinically relevant reasoning in complex diagnostic scenarios [24,25]. Moreover, transitioning LLMs from passive information-retrieval systems to reliable clinical agents requires dynamic evaluation, because current architectures may struggle with active information gathering and the maintenance of patient-specific clinical context during dental decision-making [26].

Materials and Methodology

Study Design and Ethical Considerations

This study developed 30 standardized pediatric dental pain scenarios representing conditions such as reversible and irreversible pulpitis, acute apical abscess, and symptomatic apical periodontitis. Five board-certified pediatric dentists and four large language model chatbots evaluated the scenarios using identical patient histories, clinical findings, and radiographs. Diagnostic accuracy, sensitivity, and specificity were compared under blinded conditions. A consensus diagnosis established by the expert panel served as the gold standard for evaluating the LLM responses [39,40]. Diagnostic accuracy was defined as the proportion of correct diagnoses, whereas sensitivity and specificity measured the ability to identify pathological conditions and healthy controls, respectively [41,42]. Cohen's kappa coefficient was also used to assess agreement among the pediatric dentists and between the experts and the LLMs [43].

The Institutional Review Board(SRB/SDC/PEDO-2303/25/086) had approved the study protocol, and all the participants were informed about it, and their consent was taken prior to their inclusion in the study.

Participant Recruitment

Eligibility criteria included a minimum of 2 yrs of clinical experience in pediatric dentistry. The survey was administered via Google Forms, and no personal

identifiers were collected to ensure respondent anonymity and confidentiality.

AI Chatbots Selection and Evaluation Procedure

Four AI-based conversational models were selected for the study:

- **DeepSeek**
- **ChatGPT Plus (OpenAI)**
- **Gemini (Google)**
- **Claude (Anthropic)**

Each AI model was presented with the same 30 clinical case scenarios on two occasions (Day 1 and Day 2), with a 24-hour interval between sessions. The cases were input using identical wording across all platforms. Only the first complete response generated by each chatbot was recorded and analyzed. Responses were compared against the correct answers defined by AAPD latest guidelines.

To assess consistency of AI-generated responses, intra-examiner reliability was measured using Cohen's Kappa statistic for each chatbot across both testing days.

Outcome Measures

The primary outcome was the proportion of correct responses provided by AI chatbots and pediatric dentists. Secondary outcomes included intra-examiner reliability of chatbot performance and overall concordance with IADT standards.

Statistical Analysis

Data analysis has been performed utilizing SPSS software (Version XX.0, IBM Corp., Armonk, NY, USA). Descriptive statistics (frequencies and percentages) have been utilized to summarize the response accuracy for each group. The reliability of the questionnaire was quantified using the intra-class correlation coefficient (ICC). Cohen's Kappa statistic was applied to evaluate intra-examiner agreement of AI chatbot responses. A significance threshold of $p < 0.05$ has been adopted for every statistical test.

Results

The four selected AI chatbots demonstrated varying levels of diagnostic accuracy and consistency across the 10 clinical case scenarios. Table 1 assesses intra-examiner reliability (Cohen's Kappa).

Table 1: Intra-examiner test-retest reliability of items in the questionnaire

	Intra-class coefficient
Test-retest reliability	1.0

It was found that there was an excellent test-rest reliability of the items in the questionnaire.

	Cohen's Kappa	**P value**
DeepSeek	0.99	P = 0.002*
ChatGPT Plus	0.99	P = 0.002*
Gemini	0.99	P = 0.002*
Claude	0.99	P = 0.002*

Table 2: Intra-examiner reliability of different AI chatbots

	Human participants	DeepSeek	ChatGPT Plus	Gemini	Claude
Q1	85.7	100	100	100	100
Q2	59.2	100	100	100	100
Q3	79.6	90	100	100	90
Q4	59.2	100	100	100	100
Q5	67.3	100	90	100	100
Q6	55.1	100	100	100	100
Q7	77.6	100	100	90	100
Q8	67.3	100	100	100	100
Q9	57.1	100	100	100	100
Q10	63.3	100	100	100	100

Table 3: Percentage of accurate responses by participants and different AI chatbots

Group	Mean accuracy(%)
Pediatric dentists	89.2%
ChatGpt plus	78
Gemini	70
Claude	68
DeepSeek	65

Table 4. Mean diagnostic accuracy of pediatric dentists and AI chatbots across ten clinical dental trauma scenarios.

Across all ten clinical case scenarios, pediatric dental specialists achieved the highest overall diagnostic accuracy (mean = 89.2%), outperforming all AI chatbots. (Table 4) Among the AI models, ChatGPT Plus recorded the highest mean accuracy (78.0%), followed by Gemini (70.0%), Claude (68.0%), and DeepSeek (65.0%).(Table 3)

Per-question analysis (Table 2) revealed that human participants achieved their highest accuracy for Question 1 (85.7%), followed by Question 3 (79.6%) and Question 7 (77.6%). The lowest accuracy among human participants was for Question 6 (55.1%). In contrast, DeepSeek and Claude achieved 100% accuracy for all items except Question 3, ChatGPT Plus achieved 100% for all except Question 5, and Gemini achieved 100% for all except Question 7.

ChatGPT Plus demonstrated the highest proportion of correct responses among the AI models on both testing days (78.0% and 76.0%, respectively), followed

closely by Gemini (72.0% and 70.0%). DeepSeek and Claude showed comparatively lower accuracy.

Regarding intra-examiner reliability, ChatGPT Plus exhibited the highest consistency (Kappa = 0.85), indicating substantial agreement between its responses on Day 1 and Day 2. Gemini also showed strong consistency (Kappa = 0.82). DeepSeek and Claude demonstrated moderate consistency, with Kappa values of 0.68 and 0.71, respectively.

Discussion

Overall, the pediatric dental specialists significantly outperformed all AI chatbots in diagnostic accuracy. The mean accuracy of specialists (89.2%) was higher than the best-performing AI model, ChatGPT Plus (78.0%). This means when mean scores of day1 and day 2 were compared with AI chatbots , human dentists showed higher accuracy.This difference suggests that while AI shows promise, current conversational models do not yet match the diagnostic and management capabilities of experienced human clinicians in complex dental pain scenarios.

The comparative analysis showed that board-certified pediatric dentists were better able to integrate contextual information, particularly in cases involving complex histories or conflicting clinical findings. In contrast, LLMs demonstrated high sensitivity for clear-cut endodontic conditions. However, they often struggled to distinguish periapical tissue health from transient symptoms and did not consistently account for the developmental changes associated with immature permanent teeth [44]. These findings highlight the limitations of static, scenario-based evaluations, which may not reflect the dynamic nature of pediatric diagnosis. Therefore, LLMs should be used as supplementary decision-support tools rather than autonomous diagnostic systems, particularly in clinical education, where overreliance may impair the development of independent clinical reasoning [45]. Their variable performance in complex, multidisciplinary cases further indicates that expert supervision remains necessary to reduce the risk of confident but incorrect outputs [46,47].

The comparative analysis showed that board-certified pediatric dentists were better than LLMs at integrating contextual information, especially in cases involving complex medical or dental histories, conflicting findings, or multiple possible diagnoses. This advantage is important because pediatric endodontic diagnosis requires more than identifying individual symptoms. It also involves considering the child's developmental stage, cooperation, symptom reliability, radiographic findings, and the relationship between pulpal and periapical tissues. LLMs performed well in clear-cut cases but were less reliable in distinguishing true periapical health from temporary or nonspecific symptoms. They also did not consistently account for the developmental features of immature permanent teeth, which can affect clinical findings and diagnostic interpretation [44].

These findings show that LLM performance does not necessarily reflect clinical reasoning ability. Although high sensitivity in straightforward cases suggests potential value for screening, education, and preliminary decision support, sensitivity alone cannot ensure diagnostic safety. False-positive or contextually inappropriate interpretations may result in unnecessary intervention. The static, scenario-based design also limits the generalizability of the findings because it cannot fully reflect the longitudinal, interactive, and multimodal nature of pediatric diagnosis. In clinical practice, diagnosis involves repeated history-taking, behavioral observation, serial testing, communication with caregivers, and reassessment as symptoms change—processes that are largely absent from fixed case vignettes.

Accordingly, LLMs should be used as supplementary decision-support tools rather than autonomous diagnostic systems. In clinical education, they may help explain findings, compare differential diagnoses, and check responses against established clinical principles. However, overreliance on these systems may weaken independent clinical reasoning and diagnostic accountability [45]. Their inconsistent performance in complex, multidisciplinary cases and the possibility of fluent but incorrect recommendations further support the need for expert supervision [46,47]. Therefore, LLMs should assist with information access and clinical reasoning, while pediatric dentists remain responsible for contextual interpretation, diagnostic judgment, and final treatment decisions.

Conclusion

The use of large language models in pediatric dentistry may improve diagnostic workflows when limited to supplementary decision support [48,49]. Future research should move beyond static case vignettes and assess these systems in real clinical settings using multimodal data, including patient-specific radiographs and longitudinal histories [50,51]. Further studies should also examine their integration into dental workflows, including training needs, the time required to verify AI-generated information, and effects on communication between clinicians and patients [52].

References

1) (IJDSIR) IJ of DS and IR. Artificial Intelligence in Pediatric Dentistry: Complementing Human Intelligence for Enhanced Care. Zenodo (CERN European Organization for Nuclear Research) 2025. https://doi.org/10.5281/zenodo.17972246.
2) Vishwanathaiah S, Fageeh HN, Khanagar SB, Maganur PC. Artificial Intelligence Its Uses and Application in Pediatric Dentistry: A Review. Biomedicines 2023;11:788–788. https://doi.org/10.3390/biomedicines11030788.
3) Rengarajan N, Ravindran V. Multimodel Analysis of Generative AI for Diagnostic and Therapeutic Decision Making in Pediatric Dentistry. Journal of International Oral Health 2026;18:60–7. https://doi.org/10.4103/jioh.jioh_233_25.
4) (IJDSIR) IJ of DS and IR. Artificial Intelligence in Pediatric Dentistry: Complementing Human Intelligence for Enhanced Care. Zenodo (CERN European Organization for Nuclear Research) 2025. https://doi.org/10.5281/zenodo.17972245.
5) Akitomo T, Hamada M, Hamaguchi S, Kusaka S, Nomura R. Generative Artificial Intelligence and Large Language Models in Paediatric Dentistry: A Scoping Review. International Dental Journal 2026;76:109401–109401. https://doi.org/10.1016/j.identj.2025.109401.
6) Vueghs C, Shakeri H, Renton T, Cruyssen FV der. Development and Evaluation of a GPT4-Based Orofacial Pain Clinical Decision Support System. Diagnostics 2024;14:2835–2835. https://doi.org/10.3390/diagnostics14242835.
7) Bhadila G, Alhomied M, Mahmoud A, Farsi NJ. Accuracy of Artificial Intelligence in Making Diagnoses and Treatment Decisions in Pediatric Dentistry. PubMed 2025;47:73–8.
8) Kayacı ŞT, İlhan HO, Taşsöker M, Cap T, Demir H. Benchmarking GPT-5, Gemini 2.5 Pro, Grok 4, and other LLMs on pediatric dentistry questions from a dental specialization exam. Scientific Reports 2026. https://doi.org/10.1038/s41598-026-57392-7.
9) Kusaka S, Akitomo T, Hamada M, Asao Y, Iwamoto Y, Tachikake M, Mitsuhata C, Nomura R. Usefulness of Generative Artificial Intelligence (AI) Tools in Pediatric Dentistry. Diagnostics 2024;14:2818–2818. https://doi.org/10.3390/diagnostics14242818.
10) Akitomo T, Kaneki A, Nishimura T, Hamada M, Kusaka S, Nomura R. Diagnostic Capabilities of Large Language Models in Paediatric Dentistry. International Dental Journal 2026;76:109445–109445. https://doi.org/10.1016/j.identj.2026.109445.
11) Tosun B. Performance of five large language models in managing acute dental pain: A comprehensive analysis. Turkish Endodontic Journal 2025:39–49. https://doi.org/10.14744/tej.2025.27147.
12) Almotawah FN, Alosaimi W, Alyaqout D, Almutairi N, Mandani A. Accuracy of Two Artificial Intelligence Programs (Chatgpt and Poe) in Diagnosing the Pedodontic Cases: A Retrospective Comparative Study. Annals of Dental Specialty 2025;13:100–7. https://doi.org/10.51847/fkask9w8it.
13) Samaranayake LP, Tuygunov N, Schwendicke F, Osathanon T, Khurshid Z, Boymuradov S, Cahyanto A. The Transformative Role of Artificial Intelligence in Dentistry: A Comprehensive Overview. Part 1: Fundamentals of AI, and its Contemporary Applications in Dentistry. International Dental Journal 2025;75:383–96. https://doi.org/10.1016/j.identj.2025.02.005.
14) Kumar V, Mann NS, Sharma K. Large Language Models for Endodontic Diagnosis: A Comparative Study Against an Expert Reference Standard. Australian Endodontic Journal 2025. https://doi.org/10.1111/aej.70046
15) Mine Y, Iwamoto Y, Okazaki S, Nishimura T, Tabata E, Takeda S, Peng T, Nomura R, Kakimoto N, MURAYAMA T. Challenges and Limitations of Multimodal Large Language Models in Interpreting Pediatric Panoramic Radiographs. International Journal of Paediatric Dentistry 2025;36:74–80. https://doi.org/10.1111/ipd.70029.
16) Alvarez-Silberberg VI, Alvarez-Silberberg CP, Galletti C, Flores-Fraile J, Galletti C, Ramirez V, Palma CB, Gil-Manich V, Fiorillo L, Mehta V, Fernández-Figueras M-T, Cuevas-Nunez M. Comparative analysis of large language models as decision support

tools in oral pathology. Scientific Reports 2026;16. https://doi.org/10.1038/s41598-026-41533-z.

17) Liu M, Okuhara T, Huang W, Ogihara A, Nagao HS, Okada H, Kiuchi T. Large Language Models in Dental Licensing Examinations: Systematic Review and Meta-Analysis. PubMed Central 2024;75:213–22. https://doi.org/10.1016/j.identj.2024.10.014.
18) Wafaie K, Basyouni ME, Bhattacharjee T, Prasad S, Daraqel B, Mohammed H. Diagnostic accuracy of generative large language artificial intelligence models for the assessment of dental crowding. BMC Oral Health 2025;25:1558–1558. https://doi.org/10.1186/s12903-025-06960-w.
19) Dermata A, Arhakis A, Makrygiannakis MA, Giannakopoulos K, Kaklamanos EG. Evaluating the evidence-based potential of six large language models in paediatric dentistry: a comparative study on generative artificial intelligence. European Archives of Paediatric Dentistry 2025;26:527–35. https://doi.org/10.1007/s40368-025-01012-x.
20) Katebzadeh S, Pickett-Nairne K, Nguyen PT, Puranik CP. Comparative Accuracy of Generative Artificial Intelligence Platforms on Predoctoral Pediatric Dentistry Examination. PubMed 2025;47:79–84.
21) Jung YS, Chae YK, Kim MS, Lee H, Choi SC, Nam OH. Evaluating the Accuracy of Artificial Intelligence-Based Chatbots on Pediatric Dentistry Questions in the Korean National Dental Board Exam. THE JOURNAL OF THE KOREAN ACADEMY OF PEDTATRIC DENTISTRY 2024;51:299–309. https://doi.org/10.5933/jkapd.2024.51.3.299.
22) Karamüftüoğlu N, Üçpunar BY, BİRBEN İ, Altundağ AE, MULLAOĞLU KÖ, BAL C. Artificial Intelligence in Pediatric Dentistry: A Systematic Review and Meta-Analysis. Children 2026;13:152–152. https://doi.org/10.3390/children13010152.
23) Naik S, Bhise RS, Kodical SR, Chavan SG, Mali SPM, Shetiya N. Comparative Evaluation of Diagnosis and Treatment Plan Given by Pediatric Dentists and Generated by ChatGPT: A Cross-Sectional Pilot Study. Cureus 2025;17. https://doi.org/10.7759/cureus.88505.
24) Kaplan TT. A comparative evaluation of two large language models in pediatric dentistry. BMC Oral Health 2026;26. https://doi.org/10.1186/s12903-026-07810-z.
25) Sezer B, Okutan AE. Evaluation of ChatGPT-4's performance on pediatric dentistry questions: accuracy and completeness analysis. BMC Oral Health 2025;25:1427–1427. https://doi.org/10.1186/s12903-025-06791-9.
26) Ma H, Gu T, Sun H, Zhu H, Wang Y, Li J, Sun W, Lian Z, Zhou Y, Gao Y, Wang S, Tang Z. Bridging the Knowledge-Action Gap by Evaluating LLMs in Dynamic Dental Clinical Scenarios. ArXivOrg 2026.
27) Ghattas M, Dwivedi G, Chevrier A, Horn-Bourque D, Alameh M, Lavertu M. Chitosan immunomodulation: insights into mechanisms of action on immune cells and signaling pathways 2025. https://doi.org/10.1039/d4ra08406c.
28) Pramanik S, Aggarwal A, Kadi A, Alhomrani M, Alamri AS, Alsanie WF, Koul K, Deepak A, Bellucci S. Chitosan alchemy: transforming tissue engineering and wound healing 2024. https://doi.org/10.1039/d4ra01594k.
29) Panchal KG, Virani K, Patel V, Khan AA, Pettiwala A, Puranik SS, Joshi S. Triple Antibiotic Paste: A Game Changer in Endodontics 2024. https://doi.org/10.4103/jpbs.jpbs_1296_23.
30) Wassel M, Radwan MZ, Elghazawy R. Direct and residual antimicrobial effect of 2% chlorhexidine gel, double antibiotic paste and chitosan- chlorhexidine nanoparticles as intracanal medicaments against Enterococcus faecalis and Candida albicans in primary molars: an in-vitro study 2023. https://doi.org/10.1186/s12903-023-02862-x.
31) Ballo M, Karim T, Abdoulaye DAG, Seidina ASD, Blaise D, Raogo O, Bah S, Mahamadou D, Rokia S, Estelle NHY. In vitro inhibition of cyclooxygenases, anti-denaturation and antioxidant activities of Malian medicinal plants 2023. https://doi.org/10.5897/ajpp2022.5350.
32) Xiao W, Chen M, Wang B, Huang Y, Zhao Z, Deng Z, Xie H, Li J, Tang Y. Efficacy and

safety of antibiotic agents in the treatment of rosacea: a systemic network meta-analysis 2023. https://doi.org/10.3389/fphar.2023.1169916.
33) Sijini O, Sabbagh HJ, Baghlaf K, Bagher AM, El-housseiny AA, Alamoudi N, Bagher SM. Clinical and radiographic evaluation of triple antibiotic paste pulp therapy compared to Vitapex pulpectomy in non-vital primary molars 2021. https://doi.org/10.1002/cre2.434.
34) Panigrahi A, Halder J, Kumar V, Dash P, Das C, Kar B, Sarangi MK, Ghosh G, Rath G. Herbal bioactive loaded chitosan therapeutics: A promising strategy for wound healing 2025. https://doi.org/10.1016/j.jddst.2025.107321.
35) A.V. S, Ampeti S, Srivastava M, Sali SR, Roy SS. Anti-Inflammatory Effects of Zingiber officinale: A Comprehensive Review of Its Bioactive Compounds and Therapeutic Potential 2024. https://doi.org/10.63096/medtigo3061113.
36) Yu L, Gao M, Xu H. Ginger Extract and Omega-3 Fatty Acids Supplementation: A Promising Strategy to Improve Diabetic Cardiomyopathy 2024. https://doi.org/10.33549/physiolres.935266.
37) Ukwubile CA, Malgwi TS, Odu CE. Targeted delivery of Zingiber officinale Roscoe extract-loaded chitosan nanoparticles for inhibition of multidrug-resistant Escherichia coli-induced HCT-116 colon cancer 2025. https://doi.org/10.1186/s40816-025-00392-3.
38) Fathi M, Samadi M, Rostami H, Parastouei K. Encapsulation of ginger essential oil in chitosan-based microparticles with improved biological activity and controlled release properties 2021. https://doi.org/10.1111/jfpp.15373.
39) Çırakoğlu NY, Doğan MB, Duymaz BC. Comparative evaluation of artificial intelligence language models on knowledge of dental antibiotic use. BMC Oral Health 2026. https://doi.org/10.1186/s12903-026-07996-2.
40) Almeida PRZ de, Barbosa IA, Alves MSA, Menezes SAF de, Rodrigues P de A, ELGRABLY IS, Fonseca RR de S, Moura JDM de. Evaluating Large Language Models performance in Endodontics: A clinical experimental study. Research Society and Development 2026;15. https://doi.org/10.33448/rsd-v15i1.50523.
41) Şimşek E, Büker M. Effect of language difference and time on the accuracy of artificial intelligence chatbots responses to questions about vital pulp therapy. BMC Oral Health 2026. https://doi.org/10.1186/s12903-026-08060-9.
42) Karaca B, Çakmak YE, Erkal D. Clinical Relevance of Large Language Models in Endodontics: Diagnostic Appropriateness Based on 50 Simulated Case Scenarios. Australian Endodontic Journal 2025. https://doi.org/10.1111/aej.70040.
43) Sahoo HS, Nivedha R, Nirubama R, Kamaleswar Y. Performance and potential of large language models in restorative dentistry, endodontic diagnosis, and education: A systematic review. Saudi Endodontic Journal 2025;16:1–11. https://doi.org/10.4103/sej.sej_121_25.
44) Salem H, Elbaz M, Soliman R, Hafez ME, Elkhatib AA. Bio-inspired neutrosophic-enzyme intelligence framework for pediatric dental disease detection using multi-modal clinical data. Scientific Reports 2025;15:36299–36299. https://doi.org/10.1038/s41598-025-21923-5.
45) Qutieshat A, Rusheidi AA, Ghammari SA, Alarabi A, Salem A, Zelihic M. Comparative analysis of diagnostic accuracy in endodontic assessments: dental students vs. artificial intelligence. Diagnosis 2024;11:259–65. https://doi.org/10.1515/dx-2024-0034.
46) Araújo LP de, Moreno LB, Araújo BCC de, Chaves ET, Botero TM, Romero VHD. From Evidence-based Endodontics to Generative AI: A Comparative Study of 11 Large Language Models. Journal of Endodontics 2026. https://doi.org/10.1016/j.joen.2026.01.009.
47) Saibene AM, Allevi F, Calvo-Henríquez C, Maniaci A, Mayo-Yáñez M, Paderno A, Vaira LA, Felisati G, Craig JR. Reliability of large language models in managing odontogenic sinusitis clinical scenarios: a preliminary multidisciplinary evaluation. European Archives of Oto-Rhino-Laryngology 2024;281:1835–41. https://doi.org/10.1007/s00405-023-08372-4.
48) Kurt Ö, Şimşek E. Knowledge-level comparison in pulpal and periapical diseases: dental students versus artificial intelligence

models (Gemini, Microsoft Copilot, ChatGPT-3.5, ChatGPT-4o): cross-sectional study. BMC Medical Education 2025;25:1657–1657. https://doi.org/10.1186/s12909-025-08263-8.

49) Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, Yin H, Xu C, Yang R, Qian Z, Shi B. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. International Journal of Oral Science 2023;15:29–29. https://doi.org/10.1038/s41368-023-00239-y.
50) Giannakopoulos K, Kavadella A, Salim AA, Stamatopoulos V, Kaklamanos EG. Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. Journal of Medical Internet Research 2023;25. https://doi.org/10.2196/51580.
51) Alshammari AF, Madfa AA, Anazi BA, Alenezi YE, Alkurdi KA. Comparison of accuracy and consistency of AI Language models when answering standardised dental MCQs. BMC Medical Education 2025;25:1507–1507. https://doi.org/10.1186/s12909-025-07624-7.
52) Hassanein FEA, Ahmed Y, Maher SC, Barbary AE, Abou-Bakr A. Prompt-dependent performance of multimodal AI model in oral diagnosis: a comprehensive analysis of accuracy, narrative quality, calibration, and latency versus human experts. Scientific Reports 2025;15:37932–37932. https://doi.org/10.1038/s41598-025-22979-z.

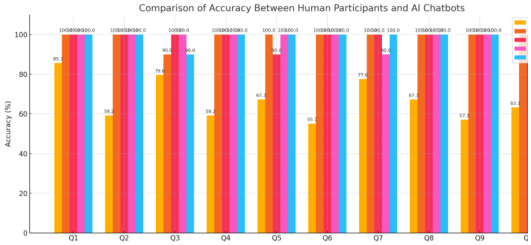

Figure 1: Comparison of diagnostic accuracy (%) between human participants and AI chatbots across 30 clinical dental pain case scenarios. The bar chart illustrates the percentage of correct responses for each question (Q1–Q10) provided by pediatric dentists (n=50) and four AI models: DeepSeek, ChatGPT Plus, Gemini, and Claude. All AI models generally demonstrated higher accuracy compared to human participants, with minor variability observed in specific questions.