RESEARCH PAPER

Human Versus Artificial Intelligence in Pediatric Dental Pain Assessment: A Comparative Evaluation of Diagnostic Accuracy, Sensitivity and Specificity

[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 assessment of dental pain in children is essential for appropriate diagnosis, treatment planning, and determination of treatment urgency. However, pediatric pain assessment is inherently challenging because children's ability to communicate and express pain is influenced by developmental, behavioral, emotional, and cognitive factors. Generative artificial intelligence (AI) systems have increasingly demonstrated potential in dental diagnosis and clinical decision support, but their performance in pediatric dental pain assessment remains insufficiently explored.

Aim

To comparatively evaluate the accuracy of contemporary generative AI chatbots—ChatGPT, Gemini, DeepSeek, and Claude—against a human dental evaluator in standardized pediatric dental pain assessment scenarios.

Materials and Methods

A cross-sectional comparative observational study was conducted using 30 standardized pediatric dental pain scenarios. A human dental evaluator and four generative AI models—ChatGPT, Gemini, DeepSeek, and Claude—independently assessed the same scenarios. Responses were coded dichotomously as correct (1) or incorrect (0) according to the predetermined reference response. Overall diagnostic accuracy was calculated as the proportion of correctly answered scenarios. Comparative analysis was performed to evaluate differences in performance among the human evaluator and AI models.

Results

Marked variation in performance was observed among the evaluated systems. Claude demonstrated the highest overall accuracy, correctly answering 23 of 30 scenarios (76.67%), followed by the human dental evaluator and Gemini, each with 20 correct responses (66.67%). ChatGPT correctly answered 15 scenarios (50.00%), whereas DeepSeek demonstrated the lowest accuracy, with 11 correct responses (36.67%). The findings demonstrated that AI performance was heterogeneous, with individual models showing substantial differences in their ability to interpret standardized pediatric dental pain scenarios. Scenario-level analysis further demonstrated variability in performance, with some scenarios being correctly answered by all evaluators while others were answered incorrectly across all evaluated systems.

Conclusion

Contemporary generative AI models demonstrated variable performance in standardized pediatric dental pain assessment. Claude showed the highest observed accuracy in the present dataset, while Gemini demonstrated performance comparable to the human dental evaluator. However, these findings should not be interpreted as evidence that AI can replace pediatric dental expertise, as standardized text-based scenarios do not reproduce the complexity of real clinical assessment, including direct examination, behavioral observation, radiographic interpretation, and contextual clinical judgment. Generative AI may have potential as a clinical decision-support and educational tool,

but prospective validation using real pediatric patients and multimodal clinical information is required before routine clinical implementation.

Keywords
How to cite this article: Swarnima Bhatnagar, Dinesh Kumar. Human Versus Artificial Intelligence in Pediatric Dental Pain Assessment: A Comparative Evaluation of Diagnostic Accuracy, Sensitivity and Specificity. Int J Drug Deliv Technol. 2026;16(78s):380-387. DOI: 10.25258/ijddt.16.78s.46.

Introduction

Evaluating the utility of generative artificial intelligence as a clinical decision support tool in managing pediatric dental pain demands a shift from measuring diagnostic accuracy toward assessing the consistency and reproducibility of model responses in alignment with established clinical expert consensus. While current literature primarily emphasizes diagnostic quality, there remains a critical gap in understanding whether these models exhibit sufficient inter-model stability to be considered reliable adjuncts in standardized clinical environments (1,2). This study aims to address this deficit by quantifying the reliability and agreement of ChatGPT, Gemini, Claude, and DeepSeek against the diagnostic thresholds established by pediatric dentists (3,4). By calculating Cohen's kappa coefficients, this research evaluates whether these platforms demonstrate the high level of inter-rater agreement required to harmonize AI-generated recommendations with evidence-based pediatric practice (5,6). Specifically, this investigation prioritizes the stability of clinical decision-making over static performance metrics, identifying potential discrepancies in pharmacological and procedural recommendations (7). By shifting the analytical focus from isolated accuracy percentages to systematic reliability, this study interrogates the variability inherent in large language models when processing complex, symptom-based pediatric scenarios (8,9). Consequently, determining the extent to which these systems maintain consistent therapeutic trajectories is vital, as even minor deviations in clinical logic could compromise safety in pediatric populations (10).

Materials and Methodology

Study Design

A cross-sectional comparative observational study was conducted to evaluate and compare the performance of pediatric dentists and generative artificial intelligence (AI)-based chatbots in the assessment of pediatric dental pain. The study was designed to determine the diagnostic performance, reliability, and level of agreement of AI-generated responses in comparison with clinical assessments made by pediatric dentists.

Study Participants and AI Models

A total of 25 pediatric dentists participated in the study. In parallel, four generative AI-based chatbot models were evaluated: ChatGPT-5, Gemini, Claude, and DeepSeek. Each pediatric dentist and AI model assessed the same standardized pediatric dental pain scenarios to ensure uniformity of the clinical information used for comparison.

Development and Administration of the Assessment Tool

Assessment was performed using a validated questionnaire developed in Google Forms comprising 30 standardized pediatric dental pain scenarios. Each clinical scenario was structured to represent a pediatric dental pain-related situation requiring diagnostic interpretation and clinical decision-making.

For each scenario, responses were obtained for four predefined domains:

1. Probable diagnosis
2. Pain severity
3. Urgency of treatment
4. Recommended management

These domains were selected to provide a comprehensive assessment of both diagnostic interpretation and subsequent clinical decision-making in pediatric dental pain scenarios.

Assessment by Pediatric Dentists

The standardized questionnaire containing the 30 pediatric dental pain scenarios was administered to the participating pediatric dentists. Each participant independently evaluated the scenarios and provided responses regarding the probable diagnosis, perceived severity of pain, urgency of treatment, and recommended clinical management.

Responses obtained from the pediatric dentists were recorded and subsequently used for comparison with the responses generated by the AI chatbot models.

Assessment by Artificial Intelligence Chatbots

The same 30 standardized pediatric dental pain scenarios were independently evaluated using ChatGPT-5, Gemini, Claude, and DeepSeek. Each AI model was required to provide responses corresponding to the same four clinical domains evaluated by the pediatric dentists: probable diagnosis, pain severity, urgency of treatment, and recommended management.

The responses generated by each AI chatbot were recorded systematically for subsequent assessment of diagnostic performance and agreement.

Assessment of AI Response Reproducibility

To determine the reproducibility and consistency of AI-generated responses, the AI models were assessed over two rounds. Responses generated during the two assessment rounds were compared for each AI model to evaluate the stability of its responses when assessing standardized pediatric dental pain scenarios.

The agreement between the two rounds was used as an indicator of AI response reproducibility and test–retest reliability. The poster reports that the AI models demonstrated almost perfect agreement across the two rounds.

Outcome Measures

The performance of the pediatric dentists and AI chatbot models was evaluated using the following outcome measures:

Diagnostic accuracy: The proportion of correctly identified responses among the total responses evaluated.

Sensitivity: The ability of the evaluated participant or AI model to correctly identify the target clinical condition.

Specificity: The ability to correctly identify scenarios in which the target clinical condition was absent.

Agreement: Agreement between responses was assessed using Cohen's kappa coefficient (κ).

In addition, the consistency of responses generated by the AI models across the two assessment rounds was evaluated to determine intra-model reliability. Accuracy, sensitivity, specificity, Cohen's kappa, and intra-examiner/AI reliability are identified as outcome measures in the original study.

Statistical Analysis

The collected data were compiled and analyzed to compare the performance of pediatric dentists with that of the four AI chatbot models. Diagnostic performance was expressed in terms of accuracy, sensitivity, and specificity. Agreement between responses was assessed using Cohen's kappa coefficient (κ). The reproducibility of responses generated by each AI model across the two assessment rounds was also evaluated.

Comparative analyses were performed between pediatric dentists and the individual AI chatbot models and among the AI models to determine differences in their performance in pediatric dental pain assessment.

Results

A total of 30 standardized pediatric dental pain scenarios were evaluated across the human dentist and four generative artificial intelligence models: ChatGPT, Gemini, DeepSeek, and Claude. The number and proportion of correctly answered scenarios varied considerably among the evaluated systems.

Claude demonstrated the highest overall performance, correctly answering 23 of 30 scenarios (76.67%). The human dentist and Gemini each correctly answered 20 of 30 scenarios (66.67%). ChatGPT correctly answered 15 of 30 scenarios (50.00%), while DeepSeek demonstrated the lowest performance, with 11 of 30 correct responses (36.67%). Thus, Claude showed the highest descriptive accuracy among all evaluated systems, whereas Gemini demonstrated an accuracy identical to that of the human dentist. The complete question-wise response pattern is provided in the source dataset.

Evaluator/model	Correct responses, n	Incorrect responses, n	Accuracy (%)	95% CI*
Human dentist	20	10	66.67	48.78–80.77
ChatGPT	15	15	50.00	33.15–66.85
Gemini	20	10	66.67	48.78–80.77

DeepSeek	11	19	36.67	21.87–54.49
Claude	23	7	76.67	59.07–88.21

Table 1 : shows the number and percentage of correctly and incorrectly answered pediatric dental pain scenarios across the human dentist and four generative AI models. Claude demonstrated the highest observed accuracy (76.67%), followed by the human dentist and Gemini (66.67% each), ChatGPT (50.00%), and DeepSeek (36.67%). CI, confidence interval. *95% confidence intervals calculated using the Wilson method.*

Because the same 30 scenarios were evaluated by all five systems, the observations were paired. An overall comparison demonstrated a statistically significant difference in performance across the five evaluators/models (Friedman $\chi^2 = 19.74$, df = 4, $p < 0.001$).

This finding indicates that the probability of obtaining a correct response differed significantly across the human dentist and evaluated AI models

Statistical test	Test statistic	df	p-value	Interpretation
Friedman test	$\chi^2 = 19.74$	4	<0.001	Statistically significant

Table 2 : presents the overall comparison of correct-response performance across the human dentist, ChatGPT, Gemini, DeepSeek, and Claude. The Friedman test demonstrated a statistically significant difference among the five evaluators/models ($p < 0.001$).

Pairwise analysis using the exact McNemar test was performed to compare the performance of the human dentist with each AI model because responses were paired at the individual scenario level.

The human dentist demonstrated higher observed accuracy than ChatGPT (66.67% vs 50.00%); however, this difference did not reach conventional statistical significance (p = 0.063). Human and Gemini demonstrated identical overall accuracy (66.67% each; p = 1.000).

DeepSeek showed significantly lower performance than the human dentist (36.67% vs 66.67%; p = 0.035).

Although Claude demonstrated a numerically higher accuracy than the human dentist (76.67% vs 66.67%), the difference was not statistically significant (p = 0.250).

Comparison	Human accuracy (%)	AI accuracy (%)	Difference (percentage points)	p-value †
Human vs ChatGPT	66.67	50.00	+16.67	0.063
Human vs Gemini	66.67	66.67	0.00	1.000
Human vs DeepSeek	66.67	36.67	+30.00	0.035
Human vs Claude	66.67	76.67	−10.00	0.250

Table 3 compares the proportion of correct responses obtained by the human dentist with each generative AI model across the same 30 standardized scenarios. Pairwise comparisons were performed using the exact McNemar test. A p-value <0.05 was considered statistically significant. †Unadjusted pairwise p-values.

Pairwise comparison among the AI models demonstrated variable performance. Claude significantly outperformed ChatGPT (76.67% vs 50.00%; p = 0.008) and DeepSeek (76.67% vs 36.67%; $p < 0.001$). Gemini also demonstrated significantly higher accuracy than DeepSeek (66.67% vs 36.67%; p = 0.004).

No statistically significant difference was observed between ChatGPT and Gemini (p = 0.227), ChatGPT and DeepSeek (p = 0.503), or Gemini and Claude (p = 0.250) in the unadjusted pairwise analyses.

AI model comparison	Accuracy (%)	Accuracy (%)	p-value†
ChatGPT vs Gemini	50.00	66.67	0.227
ChatGPT vs DeepSeek	50.00	36.67	0.503
ChatGPT vs Claude	50.00	76.67	0.008
Gemini vs DeepSeek	66.67	36.67	0.004
Gemini vs Claude	66.67	76.67	0.250
DeepSeek vs Claude	36.67	76.67	<0.001

Table 4 presents pairwise comparisons of correct-response performance among ChatGPT, Gemini, DeepSeek, and Claude using the exact McNemar test.

Bold values indicate $p < 0.05$ in the unadjusted analyses. †Pairwise p-values are unadjusted for multiple comparisons.

Considerable variation was observed in the question-wise performance of the evaluators. Several scenarios were answered correctly by all or most evaluators, whereas others resulted in incorrect responses across all five systems. For example, Questions 4, 8, 14, 18, 24, 28, and 30 were answered incorrectly by all five evaluators/models in the final dataset.

Conversely, Questions 5, 15, and 25 were answered correctly by all five evaluators/models.

These findings suggest that performance was not uniform across scenarios and that certain pediatric dental pain cases presented difficulty to both the human evaluator and AI systems.

Number of evaluators/models answering correctly	No. of scenarios	Interpretation
5/5	3	Correctly answered by all
4/5	10	High overall agreement on correct response
3/5	3	Moderate variation
2/5	5	Considerable difficulty/disagreement
1/5	2	Very low correct-response rate
0/5	7	Incorrectly answered by all

Table 5 illustrates the distribution of pediatric dental pain scenarios according to the number of evaluators/models providing a correct response. Three scenarios were correctly answered by all five evaluators, whereas seven scenarios were incorrectly answered by every evaluator/model, demonstrating substantial variation in scenario difficulty.

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

	ntra-class coefficient
Test-retest reliability	.0

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

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

The present study compared the accuracy of a single pediatric dental evaluator with that of four generative AI chatbots—ChatGPT, Gemini, DeepSeek, and Claude—using 30 standardized pediatric dental pain scenarios. No evaluator consistently outperformed the others, and performance varied substantially across AI models. Claude achieved the highest accuracy (23/30), followed by Gemini and the human dentist, who each scored 20/30. ChatGPT scored 15/30, whereas DeepSeek had the lowest score (11/30) (1,11). These findings are clinically relevant for two reasons. First, one AI model exceeded the human comparator's accuracy in this text-based scenario set. Second, another model performed considerably worse than the same comparator. Thus, "AI" should not be treated as a uniform category, because clinically important differences exist between models. Similar heterogeneity has been reported in previous studies (12,13). However, the accuracy observed in this text-based assessment should not be interpreted as evidence that AI is clinically superior to pediatric dentists. The dataset measured only binary correctness against a single expert answer key and did not assess the broader competencies required in real-world pediatric dental practice (14,15).

Evidence from clinical medicine shows a similar pattern. A recent systematic review and meta-analysis found that LLMs had approximately 89% of the diagnostic accuracy of healthcare professionals when only the top diagnosis was considered. Accuracy increased when broader differential diagnoses were assessed, but substantial differences remained between models (16). Another meta-analysis found that LLM accuracy was still lower than that of clinical professionals under rigorous evaluation (15). The present findings are consistent with this model-dependent variation: one model outperformed the human comparator, whereas another performed substantially worse.

Claude's highest accuracy (76.67%) should be interpreted cautiously because the present dataset cannot determine the reasons for this performance. Possible explanations include stronger interpretation of clinical vignettes, more conservative responses to ambiguous cases, better integration of multiple clinical clues, and differences in training data and alignment procedures (12,13,17). Recent comparative benchmark studies have also reported favorable performance by Claude models on some clinical reasoning tasks, although other studies have identified safety-related weaknesses in specific contexts (13,18). However, none of these possible explanations was directly assessed in the present study.

The finding that Gemini and the human dentist both achieved 20/30 agrees with previous research indicating that Gemini can provide clinically relevant responses when used with structured prompts (1,11,19). However, comparable accuracy in standardized, text-based scenarios does not demonstrate equivalent real-world competence in pediatric dentistry. The human evaluator could use clinical experience, behavioral assessment, examination findings, and contextual judgment—information not included in the written prompts. Moreover, the study included only one human comparator, and the binary accuracy measure did not assess the reasoning or safety of the responses from either the dentist or Gemini.

ChatGPT and DeepSeek showed lower descriptive accuracy than the human comparator, Claude, and Gemini. Similar differences between AI models have been reported in comparative medical studies (12,13,15,18,19). Possible explanations—although not directly demonstrated in this study—include greater sensitivity to prompt wording, weaker interpretation of ambiguous pediatric scenarios, difficulty distinguishing clinically similar conditions without physical examination, and the absence of tactile, radiographic, and behavioral information available to clinicians during real-world assessments (14,20). Hager et al. found that current LLMs perform significantly worse than physicians in clinical decision-making, follow guidelines inconsistently, and are sensitive to the order and amount of information provided (14). The present findings are consistent with these broader limitations.

Pediatric dental pain assessment is challenging because it involves the child's developmental age, communication skills, cognitive maturity, anxiety, fear, previous dental experiences, parental interpretation of symptoms, and nonverbal behavioral cues, including facial expression, posture, guarding, and crying (21–24). Validated instruments such as the FLACC, Wong-Baker FACES, and FPS-R scales have been developed because self-report may be unreliable in younger children, making behavioral assessment by a trained clinician an important alternative (23,25,26). Text-based scenarios cannot capture these clinical and behavioral factors. Therefore, pediatric dentists can use information that is not available in the prompts, limiting the conclusions that can be drawn from AI performance in such simplified settings (27).

The present study found substantial variation in performance across scenarios. Several cases were answered correctly by all evaluators, whereas others were answered incorrectly by all five. This pattern suggests that some scenarios were clear, while others required contextual information that the prompts did not provide. Similar findings have been reported in clinical LLM benchmarks, where performance varies with case complexity and the amount of clinical information available (12,14,19). Because the clinical details of each scenario were not analyzed, it is not possible to determine whether particular cases were inherently complex or straightforward. Nevertheless, these results support previous evidence that case difficulty strongly influences AI performance.

The framing of "human versus machine" is misleading in this context. Pediatric dentists use direct patient interaction, behavioral assessment, physical examination, contextual reasoning, knowledge of the individual patient, and real-time treatment adjustment. In contrast, LLMs process text and cannot currently examine a child, observe behavior, interpret radiographs, or adapt management to changing clinical findings (14,20,28). Generative AI is therefore better viewed as an adjunctive decision-support tool, with optimal pediatric dental care combining clinical expertise and appropriately supervised AI assistance

(28–30). The present study does not validate this combined approach, which requires prospective evaluation.

Potential uses of generative AI in pediatric dentistry include preliminary pain triage, differential diagnosis generation, clinical reasoning support, patient and parent education, and dental education and postgraduate training (28–30). These applications should remain supportive rather than autonomous, and a qualified clinician should review all AI-generated recommendations before clinical use.

Strengths of the study

The present study has several strengths: it used identical standardized scenarios for human and AI assessment, included four contemporary AI models and a human clinical comparator, evaluated performance at both the scenario and aggregate levels, and applied a predefined answer key to enable direct comparisons across evaluators.

Limitations

Important limitations include the small number of standardized scenarios, the absence of real-patient assessment, and the text-only format, which excluded physical examination, radiographs, and behavioral observation. Other limitations were the use of a single human pediatric dentist as comparator, the possible influence of prompt wording on AI responses, changes in AI model behavior over time and across versions, and limited generalizability to other pediatric dental conditions (14,18–20). In addition, the study produced one binary correctness score for each model and scenario. This measure showed only whether a response matched the answer key; it did not assess reasoning quality, safety, clinical appropriateness, or response consistency, all of which are important for future clinical use (13,18).

Future research

Future studies should examine larger and more diverse pediatric dental scenarios, involve multiple practicing pediatric dentists, and include real clinical cases with radiographs and photographs. They should evaluate AI performance across different data types and over time, with attention to accuracy, safety, explainability, reproducibility, and inappropriate recommendations (14,20,28,30). Before implementation in pediatric dental practice, prospective trials should assess AI as a decision-support tool under clear human supervision.

In summary, contemporary generative AI chatbots showed variable accuracy on standardized pediatric dental pain scenarios, with one model performing comparatively well. However, these results do not demonstrate clinical competence. Generative AI may support pediatric dental pain assessment, but it should remain an adjunct to professional expertise until validated in real-world clinical settings.

Conclusion

While observed inter-model reliability establishes a baseline for algorithmic consistency, it remains distinct from clinical validity, as high agreement does not guarantee the safety or appropriateness of generated advice in complex pediatric scenarios (31).

References

1.Rengarajan N, Ravindran V. Multimodel Analysis of Generative AI for Diagnostic and Therapeutic Decision Making in Pediatric Dentistry. Journal of international oral health [Internet]. 2026 Jan 1 [cited 2026 Feb];18(1):60–7. Available from: https://doi.org/10.4103/jioh.jioh_233_25

2.Kusaka S, Akitomo T, Hamada M, Asao Y, Iwamoto Y, Tachikake M, et al. Usefulness of Generative Artificial Intelligence (AI) Tools in Pediatric Dentistry. Diagnostics [Internet]. 2024 Dec 14 [cited 2026 Mar];14(24):2818–2818. Available from: https://doi.org/10.3390/diagnostics14242818

3.Raj M, Ravindran V. A Comparative Study on Generative Artificial Intelligence by Evaluating Multiple Large Language Models for Guidance to Parents Toward Pediatric Dentistry: A Multimodal Comparative LLM Study. Journal of international oral health [Internet]. 2025 Sept 1 [cited 2025 Nov];17(5):378–87. Available from: https://doi.org/10.4103/jioh.jioh_143_25

4.Kaplan TT. A comparative evaluation of two large language models in pediatric dentistry. BMC Oral Health [Internet]. 2026 Feb 5 [cited 2026 Mar];26(1). Available from: https://doi.org/10.1186/s12903-026-07810-z

5.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 [Internet]. 2024 Aug 27 [cited 2026 Mar];51(3):299–309. Available from: https://doi.org/10.5933/jkapd.2024.51.3.299

6.Katebzadeh S, Pickett-Nairne K, Nguyen PT, Puranik CP. Comparative Accuracy of Generative Artificial Intelligence Platforms on Predoctoral Pediatric Dentistry Examination. PubMed [Internet]. 2025 Mar 15 [cited 2025 Nov];47(2):79–84. Available from: https://pubmed.ncbi.nlm.nih.gov/40296265

7.Tosun B. Performance of five large language models in managing acute dental pain: A comprehensive

analysis. Turkish Endodontic Journal [Internet]. 2025 Jan 1 [cited 2026 Mar];39–49. Available from: https://doi.org/10.14744/tej.2025.27147
8.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 [Internet]. 2026 Jan 28 [cited 2026 Feb];76(2):109401–109401. Available from: https://doi.org/10.1016/j.identj.2025.109401
9.Bhadila G, Alhomied M, Mahmoud A, Farsi NJ. Accuracy of Artificial Intelligence in Making Diagnoses and Treatment Decisions in Pediatric Dentistry. PubMed [Internet]. 2025 Mar 15 [cited 2025 Nov];47(2):73–8. Available from: https://pubmed.ncbi.nlm.nih.gov/40296263
10.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 [Internet]. 2025 Feb 22 [cited 2026 Jan];26(3):527–35. Available from: https://doi.org/10.1007/s40368-025-01012-x
11.Bardakci E, Aydin G, Dogan MS, Dogan MS. AI-based decision support in pediatric dentistry: a comparative study of ChatGPT-5 and gemini advanced. BMC Oral Health [Internet]. 2026 Feb 14 [cited 2026 Mar];26(1). Available from: https://doi.org/10.1186/s12903-026-07897-4
12.Urdă-Cîmpean AE, Leucuta D, Drugan C, Duţu A, Călinici T, Drugan T. Assessing the Accuracy of Diagnostic Capabilities of Large Language Models. Diagnostics [Internet]. 2025 June 29 [cited 2026 Mar];15(13):1657–1657. Available from: https://doi.org/10.3390/diagnostics15131657
13.Al-Risheq AN. Comparative Benchmarking of Five Contemporary Language Models on Clinical Reasoning. medRxiv [Internet]. 2025 Dec 29 [cited 2026 Mar]; Available from: https://doi.org/10.64898/2025.12.29.25343145
14.Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature Medicine [Internet]. 2024 July 4 [cited 2026 Mar];30(9):2613–22. Available from: https://doi.org/10.1038/s41591-024-03097-1
15.Shan G, Chen X, Wang C, Liu L, Gu Y, Jiang H, et al. Comparing Diagnostic Accuracy of Clinical Professionals and Large Language Models: Systematic Review and Meta-Analysis. JMIR Medical Informatics [Internet]. 2025 Apr 25 [cited 2026 Mar];13. Available from: https://doi.org/10.2196/64963
16.Chen M, Wu Y, Ma J, Jia X, Gao C, Zhao F, et al. Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage. npj Digital Medicine [Internet]. 2026 Feb 6 [cited 2026 Mar];9(1). Available from: https://doi.org/10.1038/s41746-026-02409-8
17.Singhal KK, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nature Medicine [Internet]. 2025 Jan 8 [cited 2026 Mar];31(3):943–50. Available from: https://doi.org/10.1038/s41591-024-03423-7
18.Hoyt RE, Bajwa M. Measuring the Accuracy and Reproducibility of DeepSeek R1, Claude 3.5 Sonnet, and GPT-4.1 on Complex Clinical Scenarios. Applied Clinical Informatics [Internet]. 2026 Jan 1 [cited 2026 Feb];17(1):64–72. Available from: https://doi.org/10.1055/a-2807-4256
19.Recinella G, Altini C, Cupardo M, Cricelli I, Maestri L. Comparative accuracy of ChatGPT-o1, DeepSeek R1, and Gemini 2.0 in answering general primary care questions [Internet]. medRxiv. 2025 [cited 2026 Mar]. Available from: https://doi.org/10.1101/2025.04.15.25325518
20.Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. npj Digital Medicine [Internet]. 2024 Jan 24 [cited 2026 Jan];7(1):20–20. Available from: https://doi.org/10.1038/s41746-024-01010-1
21.Alkhouli M, Al-Nerabieah Z, Dashash M. Analyzing facial action units in children to differentiate genuine and fake pain during inferior alveolar nerve block: a cross-sectional study. Scientific Reports [Internet]. 2023 Sept 20 [cited 2026 Jan];13(1):15564–15564. Available from: https://doi.org/10.1038/s41598-023-42982-6
22.de'Angelis GL, Vincenzi F, Fornaroli F, Buonvicino D, Chiarugi A. New perspectives for optimizing fever and pain management in pediatrics: evidence supporting therapeutic awareness in clinical practice. The Italian Journal of Pediatrics/Italian journal of pediatrics [Internet]. 2025 Aug 15 [cited 2026 Mar];51(1):255–255. Available from: https://doi.org/10.1186/s13052-025-02107-3
23.Birnie KA, Hundert A, Lalloo C, Nguyen C, Stinson J. Recommendations for selection of self-report pain intensity measures in children and adolescents: a systematic review and quality assessment of measurement properties. Pain [Internet]. 2018 Aug 22 [cited 2025 Nov];160(1):5–18. Available from: https://doi.org/10.1097/j.pain.0000000000001377
24.Mohanasundari SK, Ravikoti S, Vashistha N, Singh R, Radhakrishnan DM, Manulata M, et al. Methods of Pain Assessment in Children. International Journal of Nursing Education and Research [Internet]. 2025 Oct

27 [cited 2025 Nov];285–93. Available from: https://doi.org/10.52711/2454-2660.2025.00057
25.Crellin D, Harrison D, Santamaria N, Babl FE. Systematic review of the Face, Legs, Activity, Cry and Consolability scale for assessing pain in infants and children. Pain [Internet]. 2015 July 24 [cited 2026 Jan];156(11):2132–51. Available from: https://doi.org/10.1097/j.pain.0000000000000305
26.Mayerhoffer H, Friganović A. Pain Assessment in Pediatric Patients – A Literature Review. Croatian Nursing Journal [Internet]. 2023 Aug 21 [cited 2026 Jan];7(1):87–96. Available from: https://doi.org/10.24141/2/7/1/7
27.Jacox LA, Strauman TJ, Nanney E, Rapolla A, Hodges EA, Divaris K, et al. A software-based observational coding approach for evaluating paediatric dental pain, anxiety, and fear. UNC Libraries [Internet]. 2024 Sept 14 [cited 2025 Nov]; Available from: https://doi.org/10.17615/arm3-vx66
28.Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. International Journal of Oral Science [Internet]. 2023 July 28 [cited 2026 Jan];15(1):29–29. Available from: https://doi.org/10.1038/s41368-023-00239-y
29.Pawar VV, Vanjare C. AI-powered chatbots reshaping dentistry: Opportunities, challenges, and future directions. Oral Oncology Reports [Internet]. 2024 Jan 11 [cited 2025 Nov];9:100156–100156. Available from: https://doi.org/10.1016/j.oor.2024.100156
30.Bommanavar S, Varma SR, Karobari MI. Comprehensive Scoping Review on Discrepancy in Accuracy of ChatGPT in Dental Health Practice. Human Behavior and Emerging Technologies [Internet]. 2025 Jan 1 [cited 2026 Mar];2025(1). Available from: https://doi.org/10.1155/hbe2/1269498
31.Karamüftüoğlu N, Üçpunar BY, BİRBEN İ, Altundağ AE, MULLAOĞLU KÖ, BAL C. Artificial Intelligence in Pediatric Dentistry: A Systematic Review and Meta-Analysis. Children [Internet]. 2026 Jan 21 [cited 2026 Mar];13(1):152–152. Available from: https://doi.org/10.3390/children13010152

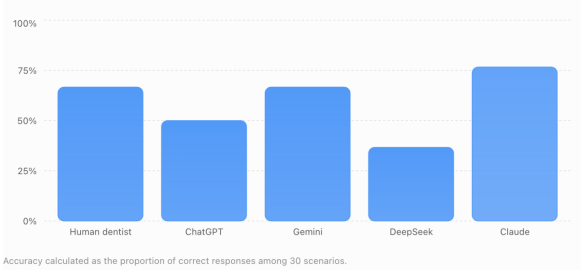

Figure 1. *Overall diagnostic accuracy of the human dentist and generative artificial intelligence models across 30 standardized pediatric dental pain scenarios. Claude demonstrated the highest accuracy (76.67%), followed by the human dentist and Gemini (66.67% each), ChatGPT (50.00%), and DeepSeek (36.67%).*

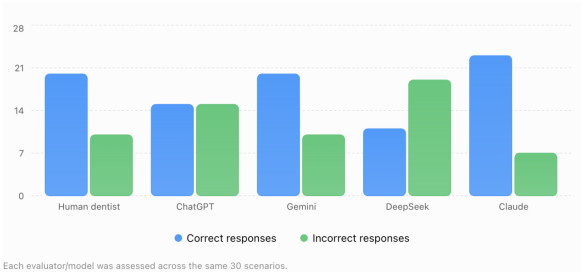

Figure 2. *Distribution of correct and incorrect responses among the human dentist and four generative AI models across 30 standardized pediatric dental pain scenarios. Claude produced the greatest number of correct responses (n=23), whereas DeepSeek produced the greatest number of incorrect responses (n=19).*

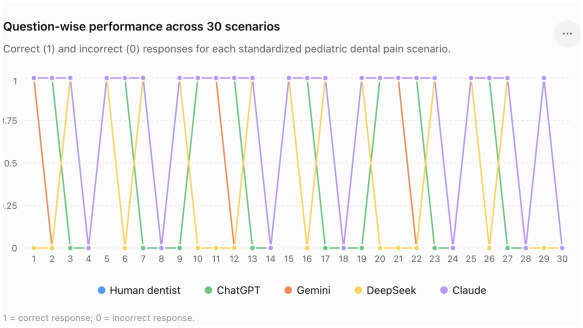

Figure 3. *Question-wise response pattern of the human dentist and four generative AI models across 30 standardized pediatric dental pain scenarios. A value of 1 represents a correct response and 0 represents an incorrect response. The figure demonstrates variation in performance across individual clinical scenarios and identifies scenarios in which all evaluators provided concordant correct or incorrect responses.*