Caries Research

Research Article

Caries Res
DOI: 10.1159/000553557

Received: November 18, 2025
Accepted: July 5, 2026
Published online: July ▪▪▪, 2026

Performance of Advanced Large Language Models in Caries Risk Assessment and Preventive Decision-Making: A Multidimensional Evaluation of Five Chatbots

Berkant Sezer Tuğba Aydoğdu

Department of Pediatric Dentistry, School of Dentistry, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye

Keywords
Artificial intelligence · Chatbot · ChatGPT · Gemini · LLaMA · Mistral

Abstract
Introduction: Large language models (LLMs) have recently been integrated into dental practice to support clinical reasoning and preventive decision-making. This study compared the performance of 5 advanced chatbots – ChatGPT-5, Claude 4.5 Sonnet, Gemini 2.5 Pro, LLaMA 3.1, and Mistral 7B – in providing evidence-based responses for caries risk assessment and preventive management in pediatric cases. ***Methods:*** Twenty-five validated, case-based questions were developed in accordance with internationally recognized pediatric and preventive dentistry guidelines. Responses were evaluated by 6 pediatric dentistry experts for accuracy, completeness, relevance, clarity, and usefulness using Likert-type scales. Response time, word count, and linguistic readability characteristics (Flesch Reading Ease Score and Flesch-Kincaid Grade Level) were additionally analyzed to compare textual complexity across chatbot-generated responses. Data normality was assessed using the Shapiro-Wilk test; parametric tests (ANOVA with Bonferroni correction) or nonparametric tests (Kruskal-Wallis with Dunn's post hoc) were applied as appropriate. ***Results:*** Statistically significant differences were observed across all qualitative criteria, including accuracy, completeness, relevance, clarity, and usefulness ($p < 0.001$). ChatGPT-5 consistently ranked among the top-performing models, showing balanced and high-quality responses across domains, while Claude 4.5 Sonnet achieved the highest accuracy and completeness scores. Gemini 2.5 Pro produced the fastest responses ($p < 0.001$), whereas Claude 4.5 Sonnet generated the longest and most linguistically complex outputs. Readability metrics also differed significantly among models ($p < 0.001$), with Mistral 7B and LLaMA 3.1 showing the highest readability. ***Conclusions:*** All evaluated chatbots generated generally relevant responses for caries risk assessment and preventive counseling; however, substantial inter-model differences were observed in qualitative performance, linguistic complexity, and response characteristics. Occasional inconsistencies and outdated content highlight the need for cautious interpretation and further externally validated evaluation before broader clinical implementation.

© 2026 S. Karger AG, Basel

Introduction

Artificial intelligence (AI) comprises computational systems designed to emulate human cognitive functions such as reasoning and pattern recognition [1]. Recent advances in deep learning have led to the development of large language models (LLMs), which generate context-

karger@karger.com
www.karger.com/cre

© 2026 S. Karger AG, Basel

Correspondence to:
Berkant Sezer, dt.berkantsezer@gmail.com; berkant.sezer@comu.edu.tr

Karger

aware text using natural language processing [2]. Their integration into healthcare has expanded rapidly, supporting diagnostics, documentation, and communication [3]. Recent models, including ChatGPT-5, Claude 4.5 Sonnet, and Gemini 2.5 Pro, as well as open-source frameworks such as LLaMA 3.1 and Mistral 7B, reflect diverse architectures that may influence reasoning performance and reliability. Consequently, evaluating their performance in specialized domains such as clinical decision support in dentistry has become an emerging research focus [4].

LLM-based systems have gained significant attention in dentistry for their capacity to improve diagnostic reasoning, clinical education, and professional communication [4, 5]. Their applications span multiple fields, including restorative dentistry, orthodontics, periodontology, oral radiology, and pediatric and special care dentistry, supporting tasks such as caries detection, radiographic interpretation, treatment planning, and diagnostic support [6–10]. By synthesizing biomedical literature and guidelines, LLMs can convert complex information into coherent, evidence-based outputs that may support professional interpretation and guideline-based clinical reasoning. Recent systematic evidence suggests that advanced LLMs can generate coherent and evidence-informed responses across diverse dental applications; however, their performance remains variable depending on model architecture, evaluation methodology, and clinical context, emphasizing the need for professional oversight and guideline-based interpretation [4].

Clinical decision-making in dentistry requires integrating patient-specific factors, diagnostic findings, and scientific evidence. However, clinician experience, cognitive bias, and limited access to updated information may compromise accuracy and consistency [11]. Chatbots offer decision-support capabilities by consolidating validated sources into coherent, context-specific outputs [4]. This can reduce diagnostic uncertainty and promote consistency in preventive and therapeutic recommendations. Particularly in pediatric dentistry, where care must balance efficacy with behavioral and developmental considerations, AI-supported guidance may enhance both precision and individualization of treatment [12, 13].

This study aims to perform a comparative evaluation of 5 LLM-based chatbots – ChatGPT-5, Claude 4.5 Sonnet, Gemini 2.5 Pro, LLaMA 3.1, and Mistral 7B – in the context of caries risk assessment and preventive counseling in children and adolescents. The evaluation was designed to reflect guideline-based clinical scenarios relevant to dental professionals. The primary objective was to assess the accuracy, completeness, relevance, clarity, and usefulness of their responses to case-based clinical scenarios reflecting diverse caries risk profiles and preventive management approaches. Secondary outcomes included comparisons of response time, word count, and readability across models. Accordingly, the null hypothesis (H_0) of this study states that there is no statistically significant difference among the chatbots in terms of accuracy, completeness, relevance, clarity, usefulness, word count, response time, or readability scores.

Methods

Study Design and Ethical Considerations

This study was designed as a cross-sectional comparative analysis to evaluate the performance of LLM-based chatbots in providing evidence-based responses related to caries risk assessment and preventive management strategies. This study was conducted and reported in accordance with the DECIDE-AI (Developmental and Exploratory Clinical Investigation of Decision-support systems driven by Artificial Intelligence) reporting guideline [14]. Because the present study represented a preclinical benchmarking evaluation rather than a real-world clinical implementation study, the guideline was adapted accordingly where applicable. Data collection was performed between September and October 2025, during which 5 publicly available chatbots were systematically evaluated: ChatGPT-5 (OpenAI, San Francisco, CA, USA), Claude 4.5 Sonnet (Anthropic, San Francisco, CA, USA), Gemini 2.5 Pro (Google DeepMind, London, UK), LLaMA 3.1 8B Instruct (Meta Platforms, Menlo Park, CA, USA), and Mistral 7B Instruct v0.3 (Mistral AI, Paris, France). ChatGPT-5, Claude 4.5 Sonnet, and Gemini 2.5 Pro were evaluated through their official web interfaces, as these proprietary models cannot be executed locally. In contrast, the open-source models LLaMA 3.1 and Mistral 7B were run in a Python-based environment using their standard pre-trained, non-fine-tuned versions to benefit from the transparency, reproducibility, and local execution advantages inherent to open-source systems.

Question Development and Content

The questions used in this study were developed with reference to the official educational and clinical resources of the American Academy of Pediatric Dentistry (AAPD) [15–19], the International Association of

Paediatric Dentistry (IAPD) [20, 21], the European Academy of Paediatric Dentistry (EAPD) [22–24], and the European Organisation for Caries Research (ORCA) [25–28]. These organizations provide widely accepted recommendations and position statements concerning caries risk assessment, preventive counseling, and management strategies in pediatric populations.

The initial set of case-based questions was developed by the study authors, both academic pediatric dentists involved in clinical care and research in preventive dentistry. An initial pool of 33 case-based questions was developed to reflect scenarios commonly encountered in pediatric dental practice. The cases represented children with diverse caries risk profiles and incorporated multiple influencing factors, including previous caries experience, dietary habits, parental knowledge, oral hygiene support, toothpaste formulations, fluoride exposure, developmental anomalies, and orthodontic conditions. Preventive and educational components were also included to address child- and parent-focused behavioral guidance.

The questions were reviewed by 4 pediatric dentists with over 10 years of clinical and academic experience, all of whom were formally trained pediatric dentistry specialists with recognized postgraduate qualifications and were working in university-affiliated academic and clinical settings. Redundant or overlapping items were removed, resulting in 27 questions. Each item was evaluated for clarity, clinical relevance, and evidence-based validity, and content validity index (CVI) values were calculated. Two items with I-CVI values below 0.78 were excluded, yielding a final validated set of 25 questions with an S-CVI of 0.93. These questions formed the basis for the chatbot evaluation, and full details are provided in the Electronic online supplementary Tables S1–S5 (for all online suppl. material, see https://doi.org/10.1159/000553557).

Querying Procedure

All evaluations were conducted using the latest publicly available versions of the 5 chatbots. All 25 questions were submitted sequentially on the same day, each entered individually and in English, with original phrasing preserved to maintain consistency and avoid contextual bias. To prevent memory effects, every query was submitted in a separate session, without rephrasing or prompt optimization. The prompts consisted only of the case scenarios presented in this study, without additional instructions or prompt engineering. All chatbot interactions were conducted from a single geographic location, and the computer IP address used for prompt submission was located in Türkiye. For the open-source models (LLaMA 3.1 and Mistral 7B), responses were generated using the following execution parameters: temperature = 0.2, top_p = 0.9, and a maximum output length of 512 tokens, applied uniformly across all open-source models to ensure comparability.

To assess consistency, all questions were re-entered 2 weeks later under identical conditions, producing 2 independent response sets. The strong correlation between sets (r = 0.94, p < 0.001) confirmed high reproducibility. All queries were submitted by a single investigator, and response time (seconds) and word count were recorded systematically. Responses were anonymized and coded before expert evaluation to minimize bias.

Expert Evaluation and Rating Process

The chatbot responses were independently evaluated by 6 pediatric dentists with at least 10 years of academic and clinical experience. The evaluators were provided with the same evidence-based references used during question development [15–28]. During assessment, evaluators independently compared each chatbot response against these evidence-based recommendations and assigned scores according to the degree of alignment with current preventive pediatric dentistry guidance. Prior to the formal evaluation process, all evaluators received standardized scoring instructions and were calibrated regarding the interpretation of the evaluation domains and scoring criteria to support consistency in evidence-based evaluations. Each response was assessed using a modified version of the Global Quality Scale, adapted for evaluating dental information quality [29–31]. The assessment included 5 qualitative domains – accuracy, completeness, relevance, clarity, and usefulness – each representing distinct yet complementary dimensions of response quality. All domains were evaluated for each of the 25 case-based questions generated by the chatbots.

Accuracy, relevance, clarity, and usefulness were evaluated using five-point Likert scales, reflecting increasing levels of factual precision, contextual appropriateness, linguistic coherence, and clinical/educational value, where 1 indicated very poor performance and 5 indicated excellent performance [30, 31]. Completeness was scored using a three-point Likert scale, assessing the degree to which responses addressed essential components of the question, ranging from superficial coverage (1) to comprehensive inclusion of key information (3) [30, 31]. Together, these metrics provided a structured and

multidimensional evaluation of the quality and applicability of chatbot-generated content.

To ensure rating consistency, all assessments were conducted under blinded conditions, and the chatbot identities were concealed from the evaluators. The same group of experts repeated the evaluation process 2 weeks after the initial assessment, with randomized response order. The 2 rounds of evaluations demonstrated excellent intra-rater reliability (ICC = 0.91, $p < 0.001$).

In parallel, the readability of all chatbot responses was evaluated using the Flesch Reading Ease Score (FRES) and the Flesch-Kincaid Grade Level (FKGL) indices, which quantify textual complexity and linguistic structure. The FRES assigns a score from 0 to 100, where higher values indicate less linguistically complex text [30, 31]. The FKGL estimates the US school grade level required to comprehend a text, with higher scores reflecting greater linguistic complexity [30, 31]. All calculations were performed using an automated online tool (Readable.com; Added Bytes Ltd., Brighton, UK).

Statistical Analysis

All data analyzed using IBM SPSS Statistics (version 26; IBM Corp., Chicago, IL, USA). Correlation heatmaps were generated using GraphPad Prism (version 10; GraphPad Software, San Diego, CA, USA). Each chatbot response to the 25 case-based questions was independently evaluated by 6 experts, and the individual ratings were subsequently aggregated across evaluators to obtain domain-level scores used for statistical analysis. Descriptive statistics for continuous variables were expressed as mean ± standard deviation for normally distributed data and as median (Q1–Q3) for non-normally distributed data. The Shapiro-Wilk test was applied to assess the normality of each variable. Based on these results, the expert-rated parameters and readability indices did not follow a normal distribution, whereas response time and word count exhibited normality. Accordingly, Kruskal-Wallis tests were used for non-normally distributed data, followed by Bonferroni-adjusted pairwise Wilcoxon tests, whereas one-way ANOVA with Bonferroni post hoc correction was used for normally distributed variables. Spearman's rank correlation coefficient (ρ) was employed to examine relationships among all qualitative and quantitative variables. Kendall's coefficient of concordance (W) and the ICC using a two-way random-effects model for absolute agreement were calculated to assess inter-rater reliability. ICC values were interpreted as follows: 0.40–0.59 = moderate, 0.60–0.79 = good, and ≥0.80 = excellent agreement. All analyses were two-tailed, and p values of <0.05 were considered statistically significant.

Results

The inter-rater reliability analysis demonstrated consistent agreement among the 6 pediatric dentistry specialists across all evaluated parameters. Using a two-way random-effects model for absolute agreement, the coefficients indicated good agreement for accuracy (ICC = 0.73), completeness (ICC = 0.68), clarity (ICC = 0.65), and usefulness (ICC = 0.72). For the relevance domain, evaluators consistently assigned identical scores across responses, resulting in negligible variance. Because the absence of between-subject variance renders the calculation of the ICC mathematically undefined, inter-rater agreement for this domain was instead reported using the percentage of absolute agreement, which reached 100%. Kendall's coefficient of concordance ($W = 0.87$, $p < 0.001$) further supported a strong overall concordance across all 5 domains, confirming that the evaluators' assessments were stable, reproducible, and statistically reliable.

Comparative analysis of expert ratings revealed statistically significant differences among the chatbots across all qualitative parameters ($p < 0.001$ for all). Pairwise post hoc comparisons identified distinct performance trends across evaluation domains. For accuracy, ChatGPT-5, Claude 4.5 Sonnet, and Gemini 2.5 Pro achieved higher median scores compared to Mistral 7B and LLaMA 3.1. In terms of completeness, Claude 4.5 Sonnet received the highest ratings, while Mistral 7B consistently obtained the lowest scores. For relevance, ChatGPT-5, Claude 4.5 Sonnet, LLaMA 3.1, and Gemini 2.5 Pro outperformed Mistral 7B. In clarity, Mistral 7B produced the most fluent and well-organized language, while the remaining chatbots performed comparably. Finally, for usefulness, Claude 4.5 Sonnet received the highest ratings, while ChatGPT-5 and Gemini 2.5 Pro followed closely and formed the next significant group. Mistral 7B and LLaMA 3.1 showed significantly lower usefulness scores. Although Claude 4.5 Sonnet excelled in completeness and usefulness, ChatGPT-5 consistently ranked highly across accuracy, relevance, clarity, and usefulness, reflecting strong overall performance. In contrast, Mistral 7B provided relatively clear text but demonstrated lower clinical adequacy. Median (Q1–Q3) values and significance groupings are presented in Table 1.

Table 1. Median (Q1–Q3) scores and post hoc groupings for each criterion

Criterion	Claude 4.5 Sonnet	Mistral 7B	ChatGPT-5	LLaMA 3.1	Gemini 2.5	*p* value
Accuracy	5 (4–5)[a]	4 (3–5)[b]	5 (4–5)[a]	4 (4–5)[b]	5 (4–5)[a]	<0.001
Completeness	3 (3–3)[a]	1 (1–2)[c]	2 (2–3)[b]	2 (2–2)[b]	2 (2–3)[b]	<0.001
Relevance	5 (5–5)[a]	4 (4–5)[b]	5 (5–5)[a]	5 (5–5)[a]	5 (5–5)[a]	<0.001
Clarity	4 (4–5)[b]	5 (3–5)[a]	4 (4–5)[b]	4 (3–5)[b]	4 (4–5)[b]	<0.001
Usefulness	5 (4–5)[a]	3 (2–4)[c]	4 (4–5)[b]	3 (3–4)[c]	4 (4–5)[b]	<0.001

Accuracy, relevance, clarity, and usefulness were evaluated using five-point Likert scales (1 = very poor and 5 = excellent), whereas completeness was assessed using a three-point scale (1 = superficial coverage and 3 = comprehensive coverage). Overall differences among chatbots were assessed using the Friedman test. When significant differences were detected, pairwise comparisons were performed using Wilcoxon signed-rank tests with Bonferroni correction. Different superscript letters indicate statistically significant pairwise differences.

For the response time, one-way ANOVA showed a significant variation across models ($F = 416.08$, $p < 0.001$). Mistral 7B and LLaMA 3.1 produced the longest response durations, whereas Gemini 2.5 Pro generated the fastest outputs. ChatGPT-5 and Claude 4.5 Sonnet displayed intermediate performance. Significant variation was also observed in word count ($F = 65.03$, $p < 0.001$). Claude 4.5 Sonnet generated the longest outputs, with Gemini 2.5 Pro and ChatGPT-5 following at intermediate lengths. Mistral 7B and LLaMA 3.1 produced the shortest responses, reflecting a more concise style. The readability indices showed significant inter-model variation ($H = 74.82$ and $H = 74.93$, both $p < 0.001$). Claude 4.5 Sonnet displayed the lowest FRES scores and highest FKGL values, indicating the greatest linguistic complexity among the evaluated models. In contrast, Mistral 7B and LLaMA 3.1 generated responses characterized by lower linguistic complexity, reflected by higher FRES and lower FKGL scores. Full descriptive statistics and significance groupings are provided in Table 2.

The Spearman correlation analysis revealed strong positive correlations among the qualitative domains ($\rho = 0.74$–0.96, $p < 0.001$), indicating that responses rated as more accurate were also perceived as more complete, relevant, and useful. The word count showed a high positive correlation with all expert-rated parameters ($\rho = 0.82$–0.94, $p < 0.001$). In contrast, response time exhibited moderate negative correlations with most quality scores ($\rho = -0.52$ to -0.71, $p < 0.01$). The readability metrics demonstrated an inverse relationship with expert ratings: the FRES correlated negatively with all qualitative criteria ($\rho = -0.69$ to -0.88, $p < 0.001$), whereas the FKGL correlated positively ($\rho = 0.73$–0.90, $p < 0.001$). The overall correlation structure is illustrated in Figure 1.

Discussion

This study comparatively evaluated the performance of 5 advanced LLM-based chatbots in generating responses related to caries risk assessment and preventive management in children and adolescents. The evaluation was conducted within a clinician-oriented framework using guideline-based clinical scenarios reflecting real preventive decision-making situations in pediatric dentistry. The evaluation incorporated multiple qualitative and quantitative parameters, including accuracy, completeness, relevance, clarity, usefulness, response time, word count, and readability. Statistically significant differences were observed across most dimensions, leading to the rejection of the null hypothesis. These results suggest that although all evaluated chatbots generated informative and clinically relevant preventive pediatric dentistry responses, their overall performance varied substantially, with each model exhibiting distinct strengths and limitations across evaluative domains.

Clinical decision-making is central to modern dental practice, requiring the integration of patient-specific variables, clinical findings, and current scientific evidence to guide optimal preventive and therapeutic strategies [11]. Accurate caries risk assessment forms the foundation of preventive dentistry, guiding individualized interventions such as fluoride application, fissure sealants, and dietary or behavioral counseling. Within

Table 2. Comparison of response time, word count, and readability scores among chatbots

Measure	Claude 4.5 Sonnet	Mistral 7B	ChatGPT-5	LLaMA 3.1	Gemini 2.5	Test (F/H)	*p* value
Response time, s	45.83±15.64[c]	244.49±49.26[d]	15.54±2.77[b]	266.95±45.04[d]	11.07±1.36[a]	F = 416.08	<0.001
Word count	1345.52±446.24[d]	405.4±100.2[a]	712.84±137.03[b]	469.64±69.45[a]	928.04±189.78[c]	F = 65.03	<0.001
FRES (readability)	−9.3 (−22.9 to −3.2)[a]	35.1 (29.9–40.2)[c]	22.6 (19.4–28.8)[b]	31.5 (27.8–40)[b]	24.9 (21–31.9)[b]	H = 74.82	<0.001
FKGL (readability)	25.50 (22.8–29.3)[c]	12.20 (11–13.4)[a]	13.5 (12.6–14.3)[b]	11.9 (10.8–12.5)[a]	13.9 (12.9–14.5)[b]	H = 74.93	<0.001

Values are expressed as mean ± standard deviation (SD) for normally distributed variables and as median (Q1–Q3) for non-normally distributed variables. Response time and word count were analyzed using one-way ANOVA followed by Bonferroni post hoc tests, as they met normality assumptions. Readability indices were analyzed using Kruskal-Wallis tests followed by Bonferroni-adjusted pairwise comparisons. Different superscript letters indicate statistically significant differences between chatbots. s, seconds; FRES, Flesch Readability Ease Score; FKGL, Flesch Kincaid Grade Level; F, ANOVA F statistic; H, Kruskal-Wallis H statistic.

this framework, advanced LLMs can serve as supportive tools by synthesizing extensive clinical data and guidelines into structured, evidence-based insights that promote timely and consistent decision-making [4].

In this study, ChatGPT-5 and Claude 4.5 Sonnet consistently demonstrated the strongest qualitative performance overall, with both models achieving high scores in accuracy, completeness, relevance, clarity, and usefulness. These findings are generally consistent with recent systematic evidence suggesting that advanced LLMs can generate clinically relevant responses across various dental applications, although performance varies according to task complexity, evaluation criteria, and model architecture [4, 32, 33]. While ChatGPT-5 delivered the most comprehensive and context-rich responses, Claude 4.5 produced highly accurate and clear outputs with superior usefulness ratings. Both models, however, generated linguistically more complex and information-dense content, reinforcing the need for careful professional interpretation of AI-generated recommendations. Open-source models showed comparatively lower performance across several evaluative domains.

Gemini 2.5 Pro exhibited intermediate performance, aligning with previous studies reporting balanced response characteristics among advanced proprietary LLMs [7]. In contrast, open-source models, while advantageous for accessibility and transparency, tended to generate shorter and less detailed outputs, possibly reflecting differences in training scale and domain-specific optimization [34]. Collectively, these findings suggest that model architecture and training strategies may influence qualitative performance across preventive pediatric dentistry tasks.

Across dental disciplines, comparative studies have similarly reported stronger overall performance and greater response consistency for advanced proprietary LLMs such as ChatGPT and Claude models compared with smaller or open-source systems [35, 36]. Studies using large datasets of dental multiple choice questions and open-ended clinical questions have reported accuracy rates exceeding 80% for ChatGPT and Claude, whereas Gemini and Mistral generally perform in the 75–79% range [36].

From a comparative perspective, performance variations among models reflected differences in response structure, linguistic complexity, and qualitative scoring patterns. Models such as ChatGPT-5 and Claude 4.5 Sonnet generally achieved higher scores in accuracy and completeness, whereas LLaMA 3.1 and Mistral 7B produced shorter responses with lower linguistic complexity that occasionally lacked clinically relevant detail. These findings suggest that higher readability and simpler linguistic structure do not necessarily correspond to greater informational depth or comprehensiveness in guideline-related pediatric dentistry responses. This interpretation may also help explain why Mistral 7B achieved the highest clarity ratings despite demonstrating lower completeness and usefulness scores. The substantially simpler linguistic structure and higher readability of its responses may have positively influenced expert perceptions of clarity, potentially reflecting a fluency-related evaluation effect even when the generated content contained less comprehensive clinical

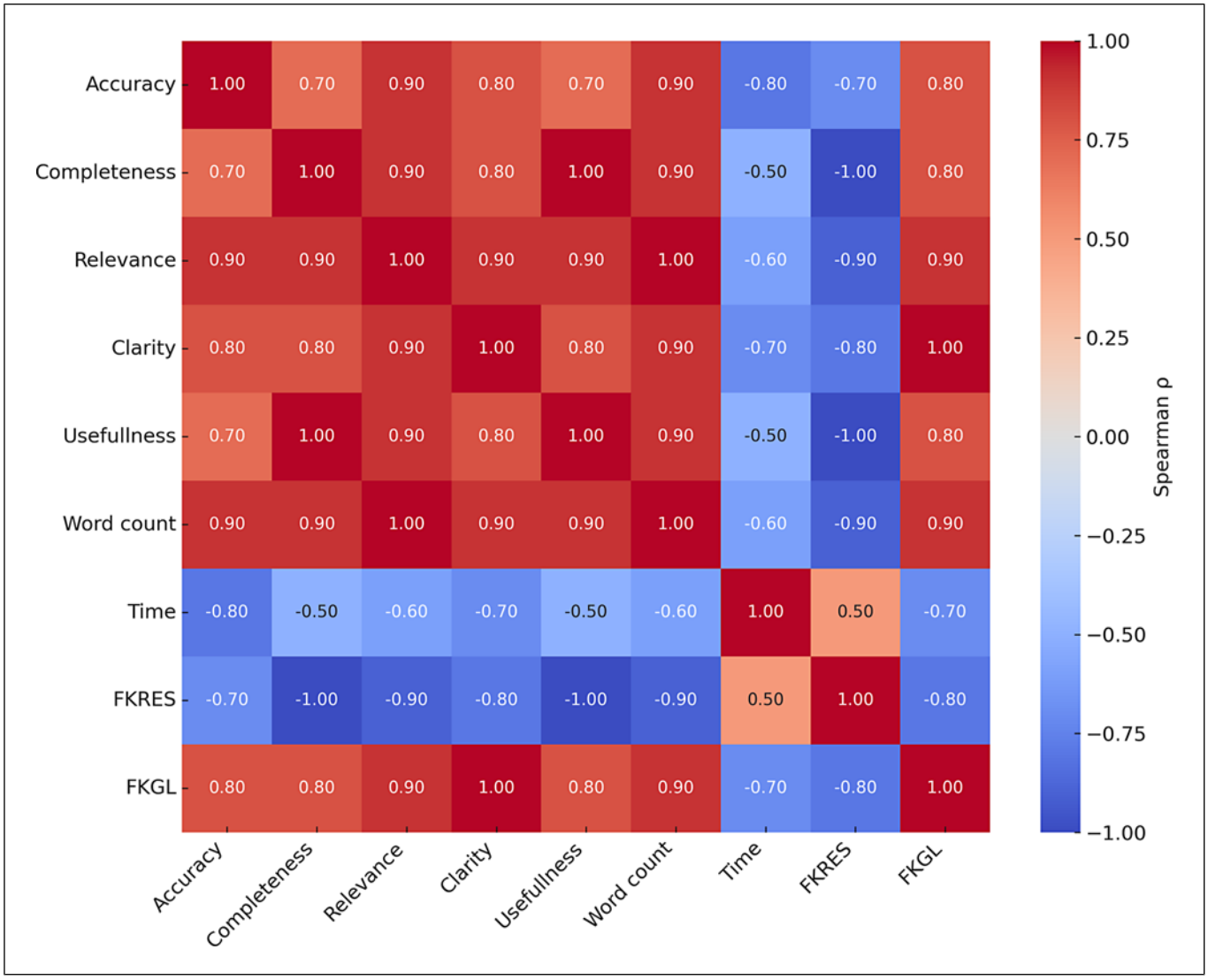


Fig. 1. Comprehensive heatmap demonstrating the comparative performance of 5 LLM-based chatbots across qualitative and quantitative evaluation criteria. Color gradients represent relative performance intensity, with darker shades indicating higher values.

detail. Accordingly, within clinician-oriented evaluations, perceived clarity should not be interpreted as a direct indicator of greater clinical precision or informational depth.

The observed correlations among word count, response time, readability, accuracy, and completeness provide insight into the structural characteristics of LLM-generated clinical responses and the multidimensional nature of expert-based healthcare benchmarking studies [37]. Positive associations between accuracy, completeness, relevance, usefulness, and word count suggest that responses containing broader contextual and preventive information generally received higher qualitative scores, consistent with previous literature emphasizing informational richness and contextual integration in LLM-generated healthcare responses [37, 38]. Conversely, the negative correlations between readability indices and completeness- and usefulness-related parameters indicate that more detailed responses were often associated with greater linguistic complexity and lower readability. Previous readability-focused evaluations of chatbot-generated medical and dental texts have similarly shown that more linguistically accessible responses do not necessarily correspond to

greater informational comprehensiveness in clinically complex contexts [39]. In contrast, response time demonstrated moderate-to-strong negative correlations with most qualitative parameters, suggesting that longer response generation times did not necessarily correspond to higher-quality outputs in the evaluated benchmarking setting. Accordingly, within the present benchmarking framework, response latency alone should not be interpreted as a marker of deeper reasoning or superior clinical performance among LLM-based systems.

Because the aim of this study was a comparative evaluation of LLM performance, no predefined cutoff threshold was established to determine whether chatbot outputs should be considered sufficient for independent clinical use. Overall, the evaluated chatbots demonstrated performance levels corresponding to scores of 4–5 for accuracy, relevance, clarity, and usefulness, and 2–3 for completeness, indicating that the generated responses were generally correct and useful despite variable informational depth. These findings suggest that LLMs may provide useful informational support for clinical reasoning and knowledge synthesis, particularly when rapid theoretical guidance or preliminary overviews are required. Nevertheless, expert oversight remains essential, as occasional inconsistencies and factual inaccuracies persist across models. Some responses reflected outdated clinical recommendations – particularly regarding fluoride concentrations, caries risk classification, dietary advice, and recall intervals – highlighting the importance of verifying AI-generated outputs against current evidence-based guidelines.

This study presents several important strengths. It is among the first comprehensive, multidimensional evaluations of 5 advanced LLM-based chatbots within a preventive dentistry framework. By jointly assessing accuracy, completeness, relevance, clarity, usefulness, readability, word count, and response time, it provides an integrated understanding of cognitive and linguistic performance among LLMs. The use of case-based questions covering diverse caries risk profiles enhanced the ecological validity of the results, reflecting real diagnostic and counseling conditions in pediatric dentistry. However, several limitations should be acknowledged. Because LLMs evolve rapidly, future model updates may affect the reproducibility and temporal generalizability of these findings. The evaluation was restricted to English-language, text-based prompts and did not include multimodal clinical inputs such as radiographs, which are important components of comprehensive caries assessment. In addition, the case scenarios were designed to reflect commonly encountered situations in pediatric dentistry; therefore, variations across healthcare systems, cultural settings, and clinical environments may limit the broader transferability of the findings. Furthermore, differences in training data transparency, model architecture, and fine-tuning strategies may have influenced how information was retrieved, prioritized, and reflected in the generated responses. The expert panel consisted exclusively of academically experienced, university-based pediatric dentists working within similar professional environments. Although this approach supported consistency in guideline-based evaluation and reduced variability in expert scoring, it may limit the transferability of the findings to clinicians practicing in different geographical, educational, or nonacademic clinical settings. As only single-session responses were analyzed, potential variability across repeated trials or software versions remains untested. Moreover, real-world pediatric dental decision-making often involves shared decision-making with caregivers; however, this family-centered dimension was beyond the scope of the present guideline-based evaluation of LLM responses. Despite these constraints, the study's broad model scope, multi-parameter design, and expert-validated scoring provide a robust foundation for future benchmarking and validation studies on AI-generated preventive dental recommendations.

Conclusions

This study comparatively evaluated 5 advanced LLM-based chatbots – ChatGPT-5, Claude 4.5 Sonnet, Gemini 2.5 Pro, LLaMA 3.1, and Mistral 7B – for caries risk assessment and preventive counseling. Significant inter-model differences were observed across accuracy, completeness, relevance, clarity, usefulness, readability, and response efficiency. ChatGPT-5 and Claude 4.5 Sonnet received the highest evaluator-assigned scores across multiple assessed domains, whereas open-source models such as LLaMA 3.1 and Mistral 7B demonstrated lower performance in several evaluated areas. These findings indicate that LLM-generated preventive pediatric dentistry responses may vary substantially according to model architecture and underlying design characteristics. Further externally validated and clinically grounded evaluation studies are required to better understand the reliability and consistency of LLM-generated dental recommendations.

Statement of Ethics

An ethics statement was not required for this study type; no human or animal subjects or materials were used.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

This study was not supported by any sponsor or funder.

Author Contributions

Conceptualization, methodology, and writing – review and editing: B.S. and T.A. Formal analysis, writing – original draft preparation, and supervision and integrity of the data and analysis: B.S. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data that support the findings of this study are not publicly available due to confidentiality restrictions but are available from the corresponding author upon reasonable request.

References

1 Dong Y, Hou J, Zhang N, Zhang M. Research on how human intelligence, consciousness, and cognitive computing affect the development of artificial intelligence. Complexity. 2020;2020:1–10. https://doi.org/10.1155/2020/1680845

2 Zubiaga A. Natural language processing in the era of large language models. Front Artif Intell. 2023;6:1350306. https://doi.org/10.3389/frai.2023.1350306

3 Lin C, Kuo CF. Roles and potential of Large language models in healthcare: a comprehensive review. Biomed J. 2025;48(5):100868. https://doi.org/10.1016/j.bj.2025.100868

4 Kaplan TT, Akyol GE. The role of large language models in dental diagnosis, decision-making, and communication: a systematic review. Jpn Dent Sci Rev. 2026;62:57–67. https://doi.org/10.1016/j.jdsr.2026.01.002

5 Alhazmi N, Alshehri A, BaHammam F, Philip M, Nadeem M, Khanagar S. Can large language models serve as reliable tools for information in dentistry? A systematic review. Int Dent J. 2025;75(4):100835. https://doi.org/10.1016/j.identj.2025.04.015

6 Sezer B, Okutan AE. Evaluation of ChatGPT-4's performance on pediatric dentistry questions: accuracy and completeness analysis. BMC Oral Health. 2025;25(1):1427. https://doi.org/10.1186/s12903-025-06791-9

7 Shirani M, Emami M. Performance comparison of large language models in treatment planning for the restoration of endodontically treated teeth over time. J Dent. 2025;161:105998. https://doi.org/10.1016/j.jdent.2025.105998

8 Farhadi Nia M, Ahmadi M, Irankhah E. Transforming dental diagnostics with artificial intelligence: advanced integration of ChatGPT and large language models for patient care. Front Dent Med. 2024;5:1456208. https://doi.org/10.3389/fdmed.2024.1456208

9 Buldur M, Sezer B. Evaluating the accuracy of chat generative pre-trained transformer version 4 (ChatGPT-4) responses to United States food and drug administration (FDA) frequently asked questions about dental amalgam. BMC Oral Health. 2024;24(1):605. https://doi.org/10.1186/s12903-024-04358-8

10 Rokhshad R, Fadul M, Zhai G, Carr K, Jackson JG, Zhang P. A comparative analysis of responses of artificial intelligence chatbots in special needs dentistry. Pediatr Dent. 2024;46(5):337–44.

11 Croskerry P. From mindless to mindful practice--cognitive bias and clinical decision making. N Engl J Med. 2013;368(26):2445–8. https://doi.org/10.1056/NEJMp1303712

12 Assiry AA, Alrehaili RS, Mahnashi A, Alkam H, Mahdi R, Hakami R, et al. How is artificial intelligence transforming the intersection of pediatric and special care dentistry? A scoping review of current applications and ethical considerations. Prosthesis. 2025;7(5):119. https://doi.org/10.3390/prosthesis7050119

13 Hatipoğlu Palaz Z, Akın Y, Karademir Ü, Çege EE, Bani M. Artificial intelligence applications in pediatric dentistry: a comprehensive review. Biomimetics. 2026;11(5):339. https://doi.org/10.3390/biomimetics11050339

14 Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: decide-ai. Nat Med. 2022;28(5):924–33. https://doi.org/10.1038/s41591-022-01772-9

15 American Academy on Pediatric Dentistry; American Academy of Pediatrics. Policy on Early Childhood Caries (ECC): classifications, consequences, and preventive strategies. Pediatr Dent. 2008;30(7 Suppl):40–3.

16 Policy on the role of dental prophylaxis in pediatric dentistry. Pediatr Dent. 2016;38(6):43–44.

17 American Academy on Pediatric Dentistry Liaison with Other Groups Committee; American Academy on Pediatric Dentistry Council on Clinical Affairs. Policy on use of Fluoride. Pediatr Dent. 2008;30(7 Suppl):34–5.

18 American Academy on Pediatric Dentistry Clinical Affairs Committee; American Academy on Pediatric Dentistry Council on Clinical Affairs. Policy on dietary recommendations for infants, children, and adolescents. Pediatr Dent. 2008;30(7 Suppl):47–8.

19 Caries-risk assessment and management for infants, children, and adolescents Pediatr Dent. 2017;39(6):197–204.

20 Early Childhood Caries: IAPD Bangkok declaration. Int J Paediatr Dent. 2019;29(3):384–6. https://doi.org/10.1111/ipd.12490

21 Tinanoff N, Banerjee A, Buzalaf MAR, Chen JW, Dhar V, Ekstrand KR, et al. Principles and care pathways for caries management in children: IAPD Rome forum. Int J Paediatr Dent. 2024;34(5):692–9. https://doi.org/10.1111/ipd.13192

22 Toumba KJ, Twetman S, Splieth C, Parnell C, van Loveren C, Lygidakis NA. Guidelines on the use of fluoride for caries prevention in children: an updated EAPD policy document. Eur Arch Paediatr Dent. 2019;20(6):507–16. https://doi.org/10.1007/s40368-019-00464-2

23 van Loveren C, van Palenstein Helderman W. EAPD interim seminar and workshop in Brussels May 9 2015: non-invasive caries treatment. Eur Arch Paediatr Dent. 2016;17(1):33–44. https://doi.org/10.1007/s40368-015-0219-3

24 Kühnisch J, Ekstrand KR, Pretty I, Twetman S, van Loveren C, Gizani S, et al. Best clinical practice guidance for management of early caries lesions in children and young adults: an EAPD policy document. Eur Arch Paediatr Dent. 2016;17(1):3–12. https://doi.org/10.1007/s40368-015-0218-4

25 Machiulskiene V, Campus G, Carvalho JC, Dige I, Ekstrand KR, Jablonski-Momeni A, et al. Terminology of dental caries and dental caries management: consensus report of a workshop organized by ORCA and cariology research group of IADR. Caries Res. 2020;54(1):7–14. https://doi.org/10.1159/000503309

26 Splieth CH, Christiansen J, Foster Page LA. Caries epidemiology and community dentistry: chances for future improvements in caries risk groups. Outcomes of the ORCA Saturday afternoon symposium, Greifswald, 2014. Part 1. Caries Res. 2016;50(1):9–16. https://doi.org/10.1159/000442713
27 Splieth CH, Banerjee A, Bottenberg P, Breschi L, Campus G, Ekstrand KR, et al. How to intervene in the caries process in children: a joint ORCA and EFCD expert delphi consensus statement. Caries Res. 2020;54(4):297–305. https://doi.org/10.1159/000507692
28 Huysmans MC, Fontana M, Lussi A, Jablonski-Momeni A, Banerjee A, Ricketts D, et al. European organisation for caries research and the European federation of conservative dentistry consensus report on clinical recommendations for caries diagnosis: paper III - caries diagnosis at the individual level. Caries Res. 2024;58(5):521–32. https://doi.org/10.1159/000539427
29 Martina S, Cannatà D, Paduano T, Schettino V, Giordano F, Galdi M. Reliability of large language model-based chatbots versus clinicians as sources of information on orthodontics: a comparative analysis. Dent J. 2025;13(8):343. https://doi.org/10.3390/dj13080343
30 Sezer B, Aydoğdu T. Performance of advanced artificial intelligence models in traumatic dental injuries in primary dentition: a comparative evaluation of ChatGPT-4 omni, DeepSeek, Gemini advanced, and Claude 3.7 in terms of accuracy, completeness, response time, and readability. Appl Sci. 2025;15(14):7778. https://doi.org/10.3390/app15147778
31 Sezer B, Aydoğdu T. Performance of advanced artificial intelligence models in pulp therapy for immature permanent teeth: a comparison of ChatGPT-4 omni, DeepSeek, and Gemini advanced in accuracy, completeness, response time, and readability. J Endod. 2025;51(11):1675–84. https://doi.org/10.1016/j.joen.2025.08.011
32 Albalawi F, Khanagar SB, Iyer K, Alhazmi N, Alayyash A, Alhazmi AS, et al. Evaluating the performance of artificial intelligence-based large language models in orthodontics—a systematic review and meta-analysis. Appl Sci. 2025;15(2):893. https://doi.org/10.3390/app15020893
33 Liu Z, Nalley A, Hao J, Ai QYH, Yeung AWK, Tanaka R, et al. The performance of large language models in dentomaxillofacial radiology: a systematic review. Dentomaxillofac Radiol. 2025;54(8):613–31. https://doi.org/10.1093/dmfr/twaf060
34 Zhihao L, Wei M, Tao L, Yaowen Z, Jingquan G, Jun W, et al. Open source AI-based SE tools: opportunities and challenges of collaborative software learning. ACM Trans Softw Eng Methodol. 2025;34(5):126. https://doi.org/10.1145/3708529
35 Viet Anh N, Trang NT. Assessment of the large language models in creating dental board-style questions: a prospective cross-sectional study. Eur J Dent Educ. 2026;30(2):673–84. https://doi.org/10.1111/eje.70015
36 Nguyen HC, Dang HP, Nguyen TL, Hoang V, Nguyen VA. Accuracy of latest large language models in answering multiple choice questions in dentistry: a comparative study. PLoS One. 2025;20(1):e0317423. https://doi.org/10.1371/journal.pone.0317423
37 Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med. 2024;7(1):258. https://doi.org/10.1038/s41746-024-01258-7
38 Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80. https://doi.org/10.1038/s41586-023-06291-2
39 Olszewski R, Brzeziński J, Watros K, Rysz J. Quantifying readability in chatbot-generated medical texts using classical linguistic indices: a review. Appl Sci. 2026;16(3):1423. https://doi.org/10.3390/app16031423