Submitted: 21 August, 2025 Accepted: 23 September, 2025 Published: 03 March, 2026 DOI:10.22514/jocpd.2026.042

ORIGINAL RESEARCH

Journal of Clinical Pediatric Dentistry

Evaluating the efficacy of large language models in providing information for parental inquiries regarding primary care of pediatric oral and dental health

Ceren Sağlam[1], Aslı Aşık[2,*], Elif Kuru[3], Handan Çelik[4], Nazan Ersin[1], Arzu Aykut Yetkiner[1], Dilşah Çoğulu[1]

[1]Department of Pediatric Dentistry, Faculty of Dentistry, Ege University, 35040 Izmir, Turkey
[2]Department of Pediatric Dentistry, Faculty of Dentistry, Izmir Tınaztepe University, 35400 Izmir, Turkey
[3]Department of Pediatric Dentistry, Faculty of Dentistry, Uşak University, 64200 Uşak, Turkey
[4]Department of Pediatric Dentistry, Faculty of Dentistry, Izmir Demokrasi University, 35140 Izmir, Turkey

***Correspondence**
asli.asik@tinaztepe.edu.tr
(Aslı Aşık)

Abstract

Background: The use of artificial intelligence (AI)-based large language models (LLMs) for accessing health information is rapidly increasing; however, limited research has evaluated their efficacy in pediatric dentistry, particularly regarding the accuracy of oral health information. This study aimed to assess the accuracy, readability, and similarity of four AI-based LLMs; ChatGPT 4.0, Google Gemini, Microsoft Copilot, and DeepSeek-R1 in responding to parental questions about children's oral and dental health. **Methods**: Twenty frequently asked questions, developed by experienced pediatric dentists, were presented to each LLM over a seven-day period, from 27 January 2025 to 03 February 2025. Responses were independently assessed for accuracy, readability, and similarity. Statistical analyses used IBM SPSS 27.0. Normality was checked by Kolmogorov-Smirnov. *T*-tests/Analysis of Variance (ANOVA) with Tukey's Honestly Significant Difference (HSD) were applied for normal data; Mann-Whitney U/Kruskal-Wallis with Dunn's *post hoc* and Bonferroni correction for non-normal data. Fisher's Exact test examined categorical variables; binary accuracy after Likert dichotomization. Inter/intra-rater reliability employed two-way random effects Intraclass Correlation Coefficients (ICC). $p < 0.05$ indicated significance. **Results**: The study demonstrated that ChatGPT 4.0 and DeepSeek-R1 achieved significantly higher accuracy scores compared to Google Gemini and Microsoft Copilot ($p < 0.001$). ChatGPT 4.0 also produced the most readable content, reflected by the lowest Average Reading Level Consensus (ARLC) score of 8.05, whereas DeepSeek-R1 generated the shortest responses ($p < 0.001$). Regarding originality, ChatGPT 4.0 and Google Gemini exhibited the lowest similarity indices, indicating a greater degree of response diversity ($p = 0.02$). **Conclusions**: These findings underscore the potential role of AI-based LLMs in facilitating parental access to evidence-based pediatric dental health information. Based on accuracy, readability, and similarity results, ChatGPT 4.0 appears the most reliable platform for delivering oral health information to parents. Furthermore, the newly introduced DeepSeek-R1 showed comparable accuracy and originality to ChatGPT 4.0, highlighting its promise as an efficient tool for user-friendly dental health guidance.

Keywords

Large language models; Artificial intelligence; Oral health; Pediatric dentistry

1. Introduction

Early childhood dental care is widely recognized as a critical determinant for establishing lifelong oral health and preventing future dental diseases. Regular dental visits, optimal oral hygiene practices, and a balanced diet play an important role in ensuring healthy oral development [1, 2].

Despite the well-documented benefits of early dental care, many parents face significant barriers in accessing accurate and timely information regarding their children's oral and dental health. Limited access to dental professionals, fragmented healthcare infrastructure, and financial constraints often drive parents to rely on inconsistent or inaccurate information sources. This lack of reliable guidance impedes informed decision-making, frequently leading to delays in preventive care and worsening oral health outcomes. Addressing this information gap through clear, evidence-based, and accessible educational resources is therefore essential to support parents in safeguarding their children's oral health [3–5].

Parents frequently encounter multiple, overlapping chal-

lenges when seeking reliable oral and dental health information, including uncertainty about appropriate fluoride use, optimal timing of the first dental visit, dietary recommendations, and oral hygiene practices, as well as conflicting advice about non-nutritive habits, such as thumb-sucking or pacifier use. Prior studies have shown that parents often depend on informal or inconsistent sources, which may delay preventive care or result in suboptimal oral and dental health practices [1, 6]. To overcome these real-world challenges, parents require accurate, comprehensible, and evidence-based resources tailored to their needs, ensuring informed decision-making and early preventive interventions.

In recent years, technological advances have transformed the way health information is disseminated, with artificial intelligence (AI) particularly large language models (LLMs) emerging as a promising tool for real-time, interactive health communication. LLMs can overcome geographic, economic, and structural barriers, offering parents accessible guidance on their children's oral and dental health. By integrating AI into healthcare communication, essential medical knowledge can be delivered more broadly, consistently, and efficiently, thereby ensuring that accurate, evidence-based information reaches diverse populations [7, 8].

Nevertheless, despite the increasing adoption of AI-based platforms in various sectors, research on their applications in dentistry remains limited. Current dental implementations primarily focus on administrative tasks, such as appointment scheduling, with minimal exploration into their potential for delivering accurate, evidence-based clinical guidance, particularly in pediatric dentistry [6, 9]. Given the critical role of parental knowledge in preventive pediatric oral and dental health, understanding the ability of LLMs to provide accurate and reliable responses to parental inquiries represents an important yet underexplored area of investigation.

Several preliminary studies have examined LLM applications within dentistry, though with limited scope and focus. While early applications of LLMs in dentistry were predominantly administrative, recent studies have demonstrated their integration across multiple disciplines, including dental public health, oral and maxillofacial surgery, periodontology, orthodontics, endodontics, prosthodontics, preventive dentistry, and dental radiology for tasks such as diagnostic support, post-operative patient queries, radiology report generation, and clinical decision-making [10–12]. Mustuloğlu *et al.* [13] evaluated ChatGPT 4.0 in the emergency management of avulsed teeth, reporting higher accuracy compared to earlier AI models. Similarly, Guven *et al.* [14] assessed ChatGPT and Google Gemini in responding to traumatic dental injury queries, demonstrating variability in both accuracy and readability. Additional investigations have explored LLMs in dental education and special needs dentistry, yielding promising but inconsistent results [15, 16]. However, no study to date has systematically evaluated the ability of LLMs to address parental questions in pediatric dentistry, highlighting a critical gap in the literature.

To the best of our knowledge, this is the first study to rigorously evaluate multiple LLMs, including the recently introduced DeepSeek-R1, for their capacity to generate accurate, reliable, and readable responses to parental inquiries in pediatric dentistry. Using standardized scoring metrics for accuracy, readability, and similarity across twenty expert-validated questions, this research advances beyond administrative or narrowly clinical applications, thereby providing novel insights into how AI-based tools can empower parents with evidence-based oral health information and support early preventive care.

Therefore, this study aimed to systematically compare the accuracy, readability, and similarity of four AI-based LLMs in responding to parental pediatric oral and dental health questions, thereby addressing a critical gap in current research and highlighting the transformative potential of LLMs in preventive pediatric dentistry.

2. Materials and methods

2.1 Study design

The aim of this study was to assess the effectiveness of responses generated by four different AI-based LLMs; ChatGPT 4.0 (OpenAI, San Francisco, CA, USA), Google Gemini (Gemini 1.5, Google, DeepMind, Mountain View, CA, USA), Microsoft Copilot (Copilot powered by GPT-4, Microsoft, Redmond, WA, USA), and DeepSeek-1 (v1.0, DeepSeek, Hangzhou DS AI, Hangzhou, Zhejiang, China) in providing information to address parental inquiries regarding children's oral and dental health. This study did not involve human participants, access to medical records, or identifiable personal data. All analyses were performed exclusively on AI-based responses to pre-defined, expert-validated pediatric dentistry questions. In accordance with institutional and international research ethics standards for studies involving no human subjects, no ethical approval or informed consent was required.

The twenty standardized questions were prepared by three expert pediatric dentists (NE, AAY, DC), each with over 20 years of clinical and academic experience in pediatric dentistry, following the American Academy of Pediatric Dentistry (AAPD) guidelines [17–21]. The responses were independently evaluated by four pediatric dentists who completed their undergraduate dental education at the same dental faculty, undertook five years of doctoral training in Pediatric Dentistry within the same department, and have since served as academic staff members at different university dental faculties. Data collection was conducted over seven consecutive days, from 27 January to 03 February 2025. In total, 20 standardized questions were presented to each of the four LLMs over seven consecutive days by four independent evaluators, yielding a dataset of 2240 responses (20 questions $\times$ 4 LLMs $\times$ 7 days $\times$ 4 examiners), all of which were included in the final analysis.

Each platform used a pre-defined set of twenty questions designed to address common parental concerns about children's dental care (Table 1). For clarity and consistency with the subsequent analysis, the twenty standardized questions were categorized into three thematic domains: (i) Tooth eruption, dental visits, and treatment needs (4 questions); (ii) Oral health and hygiene practices (11 questions); and (iii) Diet and nutrition (5 questions). This classification ensured comprehensive coverage of common parental concerns while

TABLE 1. The questions presented to the large language models.

Questions
1. When will my child's first primary tooth erupt?
2. When should the first dental visit be?
3. When should I start tooth brushing in my child?
4. What type of toothbrush should I use for my child?
5. What type of toothpaste should I use for my child?
6. How often should I replace my child's toothbrush?
7. How long should my child brush his/her teeth for?
8. Is using fluoride toothpaste safe for my child?
9. Until what age should I brush my child's teeth for him/her?
10. Does my child need to use dental floss?
11. At what age should my child begin to use dental floss for himself/herself?
12. At what age should my child begin to use mouthwash for himself/herself?
13. What kind of diet should my child have to protect his/her teeth?
14. Are sugary snacks and drinks harmful to my child's oral health?
15. Is it harmful to my child's teeth if he/she sucks his/her thumb or uses a pacifier?
16. Is breastfeeding while sleeping harmful to my child's teeth?
17. When should my child stop using a baby bottle or pacifier to prevent damage to his/her teeth?
18. Do primary teeth require dental treatment?
19. When will my child's permanent tooth erupt?
20. How often should I take my child to the dentist for routine check-ups?

enabling domain-specific performance comparisons among the LLMs. To ensure objective and standardized evaluation of response accuracy, a set of gold-standard reference responses were developed by expert pediatric dentists, based on pediatric dentistry guideline.

2.2 Accuracy evaluation

Four independent pediatric dentists (CS, AA, EK, HC) initially evaluated each response using a five-point Likert scale (1 = completely inappropriate; 5 = fully appropriate) to capture a full spectrum of accuracy levels. However, preliminary analyses revealed that ratings were highly polarized toward the two extremes, with minimal representation in intermediate categories. To improve statistical robustness, inter-rater reliability, and interpretability, the Likert scores were subsequently collapsed into a binary classification ("Appropriate" *vs*. "Inappropriate"). The full Likert-scale data remain available upon reasonable request for meta-analytic purposes or sensitivity analyses. Specifically, for each LLM response, examiners documented whether any additional explanatory information or supporting references were included beyond the direct response to the question. Both elements were coded as either "present" or "absent", and the proportion of responses containing additional information or references was calculated for each LLM and reported in the results as the percentage of presence. The "delete our chat history" prompt was implemented, and manual deletion of messages was carried out. Browser history and temporary files were also cleared.

2.3 Readability analysis

Each response generated by the LLMs was assessed for its reading difficulty using the Average Reading Level Consensus (ARLC) Calculator. This online tool, accessible at https://readabilityformulas.com/calculator-arlc-formula.php, computes the average of eight widely recognized readability formulas. These formulas include the Automated Readability Index, Flesch-Kincaid Grade Level, Flesch Reading Ease, Gunning Fog Index, Coleman-Liau Readability Index, Simple Measure of Gobbledygook (SMOG) Index, FORCAST Readability Formula, and Linsear Write Readability Index. Based on these calculations, the tool generates a difficulty score that corresponds to grade levels, ranging from very easy (first grade, ages 6–7 years) to extremely difficult (college graduate, age 23+ years) [22]. For each question, the mean number of characters and word counts in the responses generated by the LLMs were recorded and analyzed. Although eight standard readability indices were computed for methodological completeness, only three representative metrics are presented in the results to enhance interpretability.

2.4 Similarity analysis

The similarity index was used to quantitatively assess the degree of similarity between the responses provided by the software and the texts contained in various databases. In order to assess the potential plagiarism rate and the degree of originality of the responses, all results generated by the AI

model were entered into a plagiarism detection programme (Turnitin, Oakland, CA, USA, http://www.turnitin.com). Similarity rates were expressed as percentages and categorized into five different ranges: 0%, 1–24%, 25–49%, 50–74% and 75–100%. Given that standard clinical guidelines frequently employ similar phrasing across multiple sources, similarity percentages were interpreted with caution to avoid overattributing duplication to AI-based content.

2.5 Statistical analysis

All statistical analyses were performed using IBM SPSS Statistics version 27.0 (IBM Corp., Armonk, NY, USA). Descriptive statistics, including measures of central tendency (mean, median) and dispersion (standard deviation, minimum, maximum), were calculated for all variables. Normality of continuous variables was assessed using the Kolmogorov-Smirnov test. For normally distributed data, independent samples *t*-tests and one-way ANOVA with Tukey's Honestly Significant Difference (HSD) *post hoc* tests were applied; for non-normal data, Mann-Whitney U and Kruskal-Wallis tests with Dunn's *post hoc* tests and Bonferroni correction were used. Binary accuracy outcomes were analyzed following dichotomization of the original five-point Likert scores. Categorical variables, including similarity index categories, were compared using Fisher's Exact test. Inter- and intra-rater reliability were assessed using two-way random effects Intraclass Correlation Coefficients (ICC) with absolute agreement. All statistical tests were two-tailed, with p-values < 0.05 considered statistically significant.

3. Results

3.1 Dataset characteristics and reliability

The results of the study evaluating the performance of four AI-based LLMs, ChatGPT 4.0, Google Gemini, Microsoft Copilot, and DeepSeek-R1, in disseminating information regarding children's oral and dental health are presented below. Each of the 20 questions was presented to all 4 LLMs by 4 independent examiners over 7 consecutive days, generating 2240 responses to ensure reliability and reproducibility. The study focused on the accuracy, readability, and similarity of responses provided by these LLMs to a set of twenty frequently asked questions about children's oral health, as well as the evaluation of additional text and cited references. The intra- and inter-examiner reliability scores were given in Table 2. The ICC values were calculated exclusively for the accuracy scores to evaluate intra- and inter-rater reliability.

3.2 Accuracy of responses

Each LLM was assessed for the accuracy of its responses based on pediatric dentistry guidelines provided by four paediatric dentists. A total of 2240 LLM responses were evaluated over seven days and scored on a binary scale (1: Appropriate, 2: Inappropriate). As detailed in the Materials and Methods section, although initial scoring employed a five-point Likert scale, data polarization toward the extremes necessitated dichotomization for statistical analyses. Accuracy percentages for all responses are summarized in **Supplementary Table 1**. For improved visual interpretation, Fig. 1 presents a heatmap representation of these data, highlighting the contrast in accuracy scores among the four LLMs across the three major domains. As illustrated, ChatGPT 4.0 and DeepSeek-R1 achieved consistently higher accuracy compared to Google Gemini and Microsoft Copilot, particularly in the oral health and diet domains ($p < 0.001$). *Post-hoc* analysis of inaccurate responses revealed three main error categories: (i) errors of omission, where essential details or recommendations were missing; (ii) factual contradictions, where information conflicted with AAPD or other pediatric dentistry guidelines; and (iii) unreliable or unreferenced advice, where statements lacked verifiable scientific support. Among all inaccuracies, 52% were omissions, 31% involved unreferenced advice, and only 17% were factual contradictions, suggesting that most errors reflected incomplete rather than clinically misleading information.

3.3 Additional information and references

The presence and accuracy of both additional text and cited references were recorded separately for each LLM and are summarized in Table 3. Statistical analysis revealed significant differences in both the presence and accuracy of additional text ($p = 0.002$ and $p < 0.001$, respectively), with ChatGPT 4.0 and DeepSeek-R1 performing best and Microsoft Copilot performing least.

In contrast, although additional references were frequently provided, no statistically significant differences were observed across LLMs for either the presence or accuracy of these references (both $p > 0.05$). Moreover, the accuracy of cited references was consistently lower than that of the explanatory text, highlighting the persistent limitation of current LLMs in providing reliable source citations despite generating generally accurate explanatory content.

3.4 Readability analysis

ChatGPT 4.0 was identified as the most readable LLM in the study, demonstrating the lowest Average Reading Level Consensus (ARLC) score of 8.05, indicating ease of readability. In contrast, DeepSeek-R1 received the highest ARLC score of 9.69, signifying it as the most challenging LLM to read (Fig. 2).

A Kruskal-Wallis test indicated that there were statistically significant differences in the reading difficulty scores between the four LLMs ($p < 0.05$). Representative readability indices are shown in Table 4 and Fig. 3, while the full set of eight indices is available in **Supplementary Table 2**.

3.5 Response length

Among the LLMs evaluated, DeepSeek-R1 provided the shortest response, with 111.60 ± 34.75 words (554.7 ± 173.98 characters), while Google Gemini produced the longest response, containing 367.9 ± 103.06 words (1897.85 ± 543.31 characters). A statistically significant difference was found between the groups ($p < 0.001$) (Table 5).

TABLE 2. Intra- and inter-examiner reliability (intra-class correlation coefficients).

	Intraclass correlation coefficient (95% CI)	
The classification of the questions	Intra-examiner reliability	Inter-examiner reliability
ChatGPT 4.0	0.76 (0.65–0.82)	0.77 (0.64–0.88)
Google Gemini	0.72 (0.65–0.82)	0.70 (0.67–0.75)
Microsoft Copilot	0.73 (0.61–0.80)	0.71 (0.64–0.84)
DeepSeek-R1	0.85 (0.72–0.97)	0.82 (0.78–0.86)

Intraclass Correlation Coefficient (ICC) (Two-way random, absolute agreement). CI: Confidence Interval.

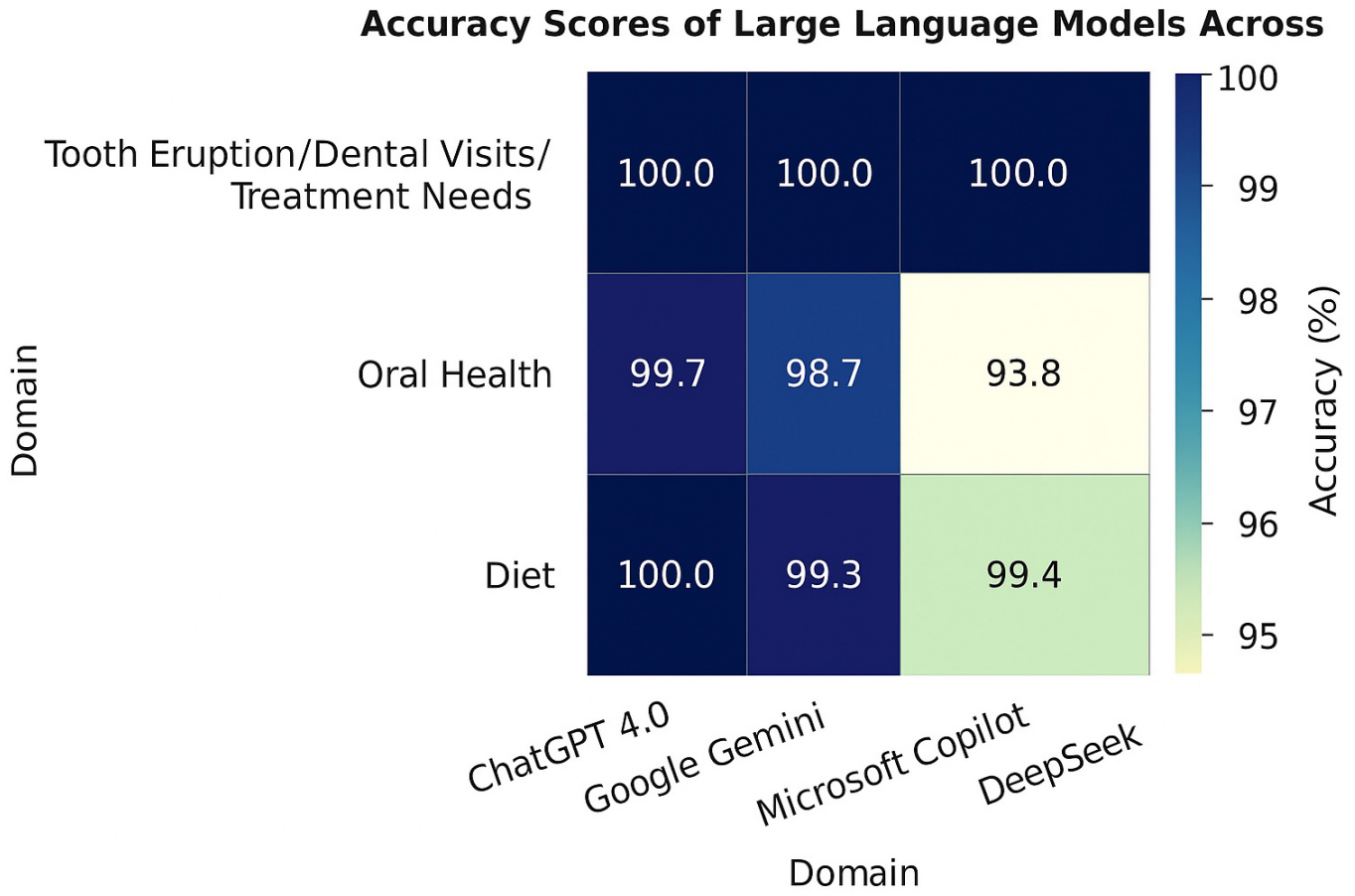

FIGURE 1. Heatmap illustrating the accuracy scores (%) of four LLMs. Kruskal-Wallis test + Dunn's *post hoc* test (Bonferroni correction), (ChatGPT 4.0, Google Gemini, Microsoft Copilot, and DeepSeek-R1) across three major domains: Tooth eruption/Dental visits/Treatment needs (4 questions), Oral health (11 questions), and Diet (5 questions). Darker shades indicate higher accuracy percentages. Full numerical data are provided in **Supplementary Table 1**.

TABLE 3. Presence and accuracy of additional text and cited references for each LLM.

LLM	Presence of Additional Text (%)	Accuracy of Additional Text (%)	Presence of Additional References (%)	Accuracy of Additional References (%)
ChatGPT 4.0 (n = 560)	100	100	90	48
Google Gemini (n = 560)	95	95	80	42
Microsoft Copilot (n = 560)	70	65	80	44
DeepSeek-R1 (n = 560)	100	100	85	35
p-value	0.002*	<0.001*	0.062	0.067

Kruskal-Wallis test + Dunn's post hoc test (Bonferroni correction). LLM: large language models.
**p < 0.05 considered statistically significant.*

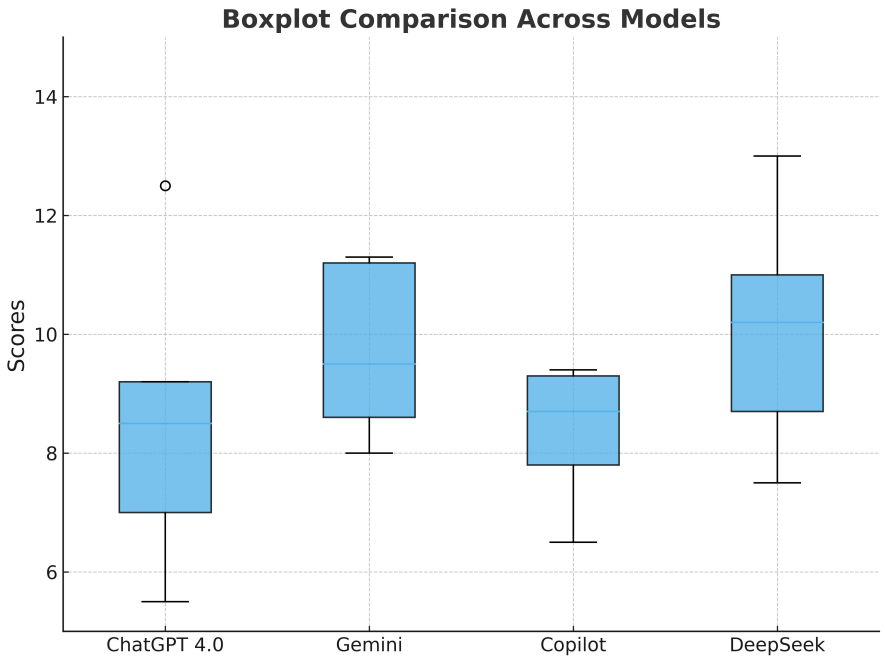

FIGURE 2. Comparison of the distribution of readability scores: Average Reading Level Consensus. Kruskal-Wallis test + Dunn's *post hoc* test (Bonferroni correction).

TABLE 4. Comparison of the readability scores of large language models.

	ChatGPT 4.0 Mean (SD)	Google Gemini Mean (SD)	Microsoft Copilot Mean (SD)	DeepSeek-R1 Mean (SD)	*p*
Flesch-Kincaid Grade Level	74.45 ± 10.58	64.75 ± 6.62	69.40 ± 6.49	63.95 ± 10.07	<0.001*
Gunning Fog Index	6.30 ± 1.98	7.94 ± 1.34	7.26 ± 1.37	9.07 ± 1.90	<0.001*
SMOG Index	7.11 ± 1.16	7.96 ± 0.79	6.97 ± 1.50	8.58 ± 1.99	0.004*

*Kruskal-Wallis test + Dunn's post hoc test (Bonferroni correction). *p < 0.05 considered statistically significant.*
Comparison of the readability scores of large language models across three representative indices: Flesch-Kincaid Grade Level, Gunning Fog Index, and SMOG Index. These indices were selected as the most widely used and methodologically representative measures of text readability in health communication research. Full results for all eight readability indices are provided in ***Supplementary Table 2****.*
SD: Standard Deviation; SMOG: Simple Measure of Gobbledygook.

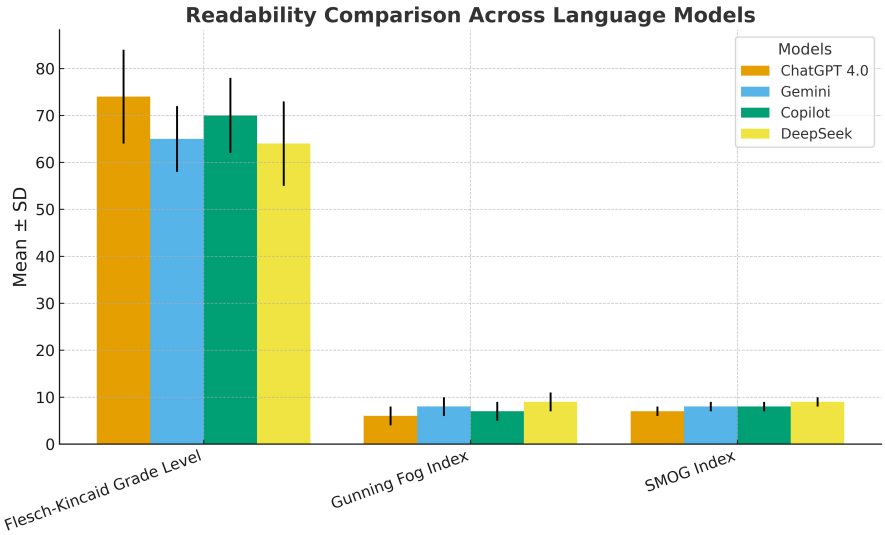

FIGURE 3. Comparison of the scores of readability scales. Kruskal-Wallis test + Dunn's *post hoc* test (Bonferroni correction). (R1: Flesch-Kincaid Grade Level, R2: Gunning Fog Index, R3: SMOG Index). SD: Standard Deviation; SMOG: Simple Measure of Gobbledygook.

TABLE 5. Overview of response length from LLMs to the questions.

LLM	Response length (words)			Response length (characters)		
	Mean (SD)	Minimum	Maximum	Mean (SD)	Minimum	Maximum
ChatGPT 4.0	163.65 ± 56.31	79	192	804.50 ± 279.35	597	1316
Google Gemini	367.90 ± 103.06	198	512	1897.85 ± 543.31	1625	2452
Microsoft Copilot	197.90 ± 89.65	83	241	969.55 ± 445.88	712	1249
DeepSeek-R1	111.60 ± 34.75	56	138	554.70 ± 173.98	391	763
p	<0.001*			<0.001*		

*One-way ANOVA (if normally distributed) or Kruskal-Wallis test (if non-normal). LLM: large language models; SD: standard deviation. *p < 0.05 considered statistically significant.*

3.6 Similarity analysis

To evaluate the originality of chatbot responses, the Similarity Index was assessed using Turnitin, a widely recognized and reliable plagiarism detection software. The results for each LLM are summarized in Table 6, with similarity percentages categorized into five ranges. ChatGPT 4.0 and Google Gemini had the lowest similarity rates, indicating a greater degree of originality in their responses. Statistical analysis using Fisher's Exact Test revealed significant differences in the similarity rates between the groups ($p = 0.02$).

4. Discussion

This study represents the first comprehensive evaluation of multiple LLMs within pediatric dentistry, providing an integrated analysis of accuracy, readability, and originality across ChatGPT 4.0, DeepSeek-R1, Google Gemini, and Microsoft Copilot. By using standardized parental inquiries, the present work not only quantifies the current performance of these platforms but also situates the findings within the broader landscape of AI adoption in healthcare communication, extending recent work on the clinical utility of generative AI [6–9].

A key finding of this study was the superior accuracy of ChatGPT 4.0 and DeepSeek-R1 compared with Google Gemini and Microsoft Copilot, particularly in oral health and dietary domains. While initial scoring employed a five-point Likert scale to capture nuanced differences in response quality, pilot analyses revealed a highly polarized distribution, with most ratings clustering at the extremes. To ensure methodological rigor and preserve statistical power, the scores were, therefore, collapsed into a binary classification ("Appropriate" *vs.* "Inappropriate"), following best practices in diagnostic accuracy research when intermediate categories lack adequate representation [13–15]. This analytic refinement does not diminish interpretative granularity, as full Likert data remain available for transparency and potential meta-analytic purposes. Importantly, the observed accuracy advantage aligns with previous studies demonstrating that recent LLMs benefit from expanded training datasets, advanced reinforcement learning, and larger parameter counts, yielding more precise and contextually appropriate outputs [23, 24]. The inclusion of DeepSeek-R1, a novel platform, further illustrates that emerging systems can achieve competitive accuracy, underscoring both the rapid technological evolution of generative AI and its potential relevance for pediatric dentistry applications [24, 25].

Nonetheless, the inconsistent citation practices observed across all evaluated LLMs remain a significant concern. Although AI-generated responses often contained accurate clinical content, the lack of verifiable, properly formatted references undermines their scientific reliability. Previous investigations have similarly emphasized that, without integration into validated medical databases, LLM outputs risk propagating information that is clinically reasonable but unsupported by primary evidence [23, 25]. Addressing this limitation will require not only technical advancements in real-time citation generation, but also the establishment of professional oversight frameworks to ensure compliance with evidence-based guidelines before AI systems can be responsibly deployed in patient education or clinical decision-making.

The readability analysis revealed notable differences in linguistic accessibility across platforms. ChatGPT 4.0 generated the most parent-friendly content, whereas DeepSeek-R1 produced responses with greater textual complexity. While advanced linguistic sophistication may benefit professional audiences, it could impede comprehension among caregivers with limited health literacy, consistent with prior studies advocating for tailored health communication strategies [25, 26]. To ensure methodological rigor without overwhelming the reader, eight established readability indices were calculated, but only the three most widely cited metrics; the Flesch-Kincaid Grade Level, the Gunning Fog Index, and the SMOG Index are presented in the main text for clarity, with the full set reported in **Supplementary Table 2** for transparency and potential meta-analytic use. Given the pivotal role of health literacy in shaping parental decision-making, optimizing AI-generated content for diverse educational backgrounds remains critical for maximizing clinical applicability.

Similarity analysis revealed that elevated similarity scores were often attributable to the use of standardized clinical terminology rather than genuine textual duplication, consistent with previous research on AI text generation [16]. However, occasional redundancy and unreferenced factual claims highlight the necessity for AI-specific originality metrics capable of distinguishing between clinically appropriate standardization and unnecessary repetition. Until such tools are developed, similarity indices derived from plagiarism detection systems originally designed for human-authored texts should be interpreted cautiously, as they may overestimate duplication or overlook nuanced forms of redundancy [16]. It is important to

TABLE 6. The similarity rates and percentages of different large language models.

LLM	0% Similarity	1–24% Similarity	25–49% Similarity	50–74% Similarity	75–100% Similarity
ChatGPT 4.0	5 responses	12 responses	2 responses	1 response	0 response
Google Gemini	6 responses	10 responses	3 responses	1 response	0 response
Microsoft Copilot	7 responses	3 responses	9 responses	1 response	0 response
DeepSeek-R1	8 responses	8 responses	3 responses	1 response	0 response

Fisher's Exact test.

recognize that plagiarism detection tools such as Turnitin were designed for human-authored texts and have not been formally validated for AI-generated content. As a result, their similarity scores should be interpreted cautiously, since common clinical terminology or widely accepted patient education phrases may lead to false positives, while repetitive or semantically derivative AI text may escape detection. Consequently, these metrics should be regarded as approximate indicators rather than definitive measures of originality, particularly in the context of AI-generated outputs.

Another noteworthy observation was the variability in response length and structure across platforms, suggesting a lack of standardized output parameters. Previous work in clinical informatics has similarly reported that LLMs often produce content of inconsistent scope and depth, reflecting randomized elements inherent in generative architectures [23–25]. Our qualitative error analysis indicated that most inaccuracies represented omissions rather than explicit factual contradictions, reinforcing the notion that current LLMs, while capable of producing generally accurate content, lack systematic mechanisms to ensure completeness. Future research should explore structured prompt engineering, context-aware summarization algorithms, and human-in-the-loop validation to mitigate these limitations and enhance output reliability.

Methodologically, the present study employed a standardized set of expert-validated pediatric dentistry questions to facilitate reproducibility and cross-model comparisons. However, practical parental inquiries exhibit far greater linguistic, cultural, and contextual variability than controlled datasets can capture. Prior studies in medical AI evaluation have emphasized that model performance frequently declines when exposed to heterogeneous, patient-generated inputs, highlighting the importance of external validation under authentic usage conditions [16, 25, 26]. Future work should therefore incorporate larger, multilingual, and demographically diverse datasets, enabling the assessment of model robustness across variable literacy levels, cultural contexts, and clinical scenarios.

Finally, while this study did not evaluate response latency due to platform release variability, previous research has suggested that promptness, accuracy, and readability collectively determine the practical usability of AI systems in healthcare communication [16, 25, 26]. As generative models continue to evolve, benchmarking future systems on all three parameters will provide a more comprehensive picture of their clinical utility.

The limitations of this study encompass several critical considerations. Firstly, although the twenty questions utilized in this research reflect common parental concerns, they do not comprehensively cover the full spectrum of dental issues that children may encounter. Future investigations evaluating LLM responses across a broader range of paediatric dentistry topics would be beneficial. Secondly, the assessment process in this study relied on the subjective judgments of four paediatric dentists. While this approach ensured a thorough evaluation, it inherently carries the potential for bias in determining the appropriateness of responses. Future studies with larger and more heterogeneous datasets may benefit from retaining full Likert granularity to capture nuanced differences among LLM outputs; however, dichotomization in the present study was statistically justified to ensure robust reliability metrics and clear interpretability. Future studies involving a larger panel of pediatric dentistry specialists would improve the reliability and validity of these assessments. Finally, the questions posed to LLMs in this study were formulated by professionals, meaning that response accuracy largely depended on the clarity and precision of the queries. It is essential that these platforms demonstrate the ability to generate reliable responses to inquiries from non-experts, including patients and parents. Continued advancements are needed to address this critical issue. It is important to note that the standardized set of twenty questions used in this study, while expert-validated and aligned with pediatric dentistry guidelines, does not capture the full linguistic diversity or contextual variability of real-world parental inquiries. As such, the present findings should be interpreted as preliminary evidence obtained under controlled conditions rather than as confirmation of real-world reliability. Future research should incorporate larger, more heterogeneous question sets derived from actual parental inputs to better assess the robustness, adaptability, and clinical applicability of LLM-generated responses.

AI-based LLMs hold promise for enhancing access to paediatric dental health information; however, their reliability hinges on alignment with established clinical guidelines. Further research is necessary to enhance the accuracy, clarity, and originality of AI-generated responses while minimizing dependence on unverified sources. As AI technology advances, continuous refinement is crucial to improving the credibility and readability of information while addressing concerns related to plagiarism and the use of non-validated references. These findings contribute to the growing body of research on AI's role in healthcare communication and provide valuable insights for parents seeking trustworthy dental guidance for their children.

Taken together, these findings underscore the transformative potential of LLMs for pediatric oral health education while simultaneously revealing persistent challenges in reference reliability, linguistic accessibility, output consistency, and prac-

tical generalizability. Addressing these limitations will require technological refinement, integration with authoritative medical databases, professional oversight, and extensive external validation across broader populations and clinical contexts. As generative AI systems become increasingly sophisticated, aligning their outputs with evidence-based pediatric dentistry guidelines will be essential to ensure that they evolve from experimental tools into trustworthy, clinically actionable resources for parents, practitioners, and researchers alike [16, 25, 26].

5. Conclusions

Within the constraints of this study, ChatGPT 4.0 delivered the strongest overall performance, while DeepSeek-R1, a newly introduced platform matched its accuracy and originality in several domains. This novel inclusion highlights DeepSeek's emerging potential for pediatric dentistry. These results provide initial evidence on how LLMs can address parental inquiries about children's oral health, emphasizing the urgent need for validation with larger, authentic datasets to guide their responsible clinical integration.

AVAILABILITY OF DATA AND MATERIALS

The data that support the findings of this study are available from the corresponding author, Asst. Prof. Aslı Aşık, PhD, DDS, upon reasonable request.

AUTHOR CONTRIBUTIONS

CS, AA, EK, HÇ, NE, AAY and DÇ—conceptualization; writing-review and editing. CS, AA, EK, HÇ and DÇ—methodology; data collection. CS, AA and DÇ—writing-original draft preparation.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study was exempt from ethical review as it involved no human participants, identifiable health records, or personal data, and only analyzed AI-generated textual content based on standardized dental health inquiries. All data analyzed were derived from simulated interactions between pre-designed pediatric dental questions and publicly available AI-based chatbot platforms. Accordingly, in line with institutional policy and international ethical standards, the requirement for ethical approval and informed consent was deemed unnecessary.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Ayça Ölmez for her valuable assistance with the statistical analyses.

FUNDING

This research received no external funding.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY MATERIAL

Supplementary material associated with this article can be found, in the online version, at https://oss.jocpd.com/files/article/2028749196687753216/attachment/Supplementary%20material.docx.

REFERENCES

[1] Wang M, Zhang Y, Li X, Liu X. Understanding and reducing delayed dental care for early childhood caries: a structural equation model approach. BMC Public Health. 2025; 25: 523.

[2] Giles E, Gray-Burrows KA, Bhatti A, Rutter L, Purdy J, Zoltie T, *et al*. "Strong Teeth": an early-phase study to assess the feasibility of an oral health intervention delivered by dental teams to parents of young children. BMC Oral Health. 2021; 21: 267.

[3] Chauhan A, Staples A, Forshaw E, Zoltie T, Nasser R, Gray-Burrows KA, *et al*. Exploring and enhancing the accessibility of children's oral health resources (called HABIT) for high risk communities. Frontiers in Oral Health. 2024; 5: 1392388.

[4] Abdelrahman HH, Hamza M, Essam W, Adham M, AbdulKafi A, Baniode M. Electronic oral health surveillance system for Egyptian preschoolers using District Health Information System (DHIS2): design description and time motion study. BMC Oral Health. 2024; 24: 807.

[5] Cooper D, Kim J, Duderstadt K, Stewart R, Lin B, Alkon A. Interprofessional oral health education improves knowledge, confidence, and practice for pediatric healthcare providers. Frontiers in Public Health. 2017; 5: 209.

[6] Gugnani N, Pandit IK, Gupta M, Gugnani S, Kathuria S. Parental concerns about oral health of children: is ChatGPT helpful in finding appropriate answers? Journal of Indian Society of Pedodontics and Preventive Dentistry. 2024; 42: 104–111.

[7] Avanzo M, Stancanello J, Pirrone G, Drigo A, Retico A. The evolution of artificial intelligence in medical imaging: from computer science to machine and deep learning. Cancers. 2024; 16: 3702.

[8] Pezoulas VC, Zaridis DI, Mylona E, Androutsos C, Apostolidis K, Tachos NS, *et al*. Synthetic data generation methods in healthcare: a review on open-source tools and methods. Computational and Structural Biotechnology Journal. 2024; 23: 2892–2910.

[9] 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: 527–535.

[10] Zhu J, Chen Z, Zhao J, Yu Y, Li X, Shi K, *et al*. Artificial intelligence in the diagnosis of dental diseases on panoramic radiographs: a preliminary study. BMC Oral Health. 2023; 23: 358.

[11] Zhang B, Campbell J. Beyond passive learning: artificial intelligence–driven spaced repetition learning and digital tools in oral and maxillofacial surgery residency. Journal of Oral and Maxillofacial Surgery. 2025; 83: 1060–1064.

[12] Guo X, Shao Y. AI-driven dynamic orthodontic treatment management: personalized progress tracking and adjustments—a narrative review. Frontiers in Dental Medicine. 2025; 6: 1612441.

[13] Mustuloğlu Ş, Deniz BP. Evaluation of chatbots in the emergency management of avulsion injuries. Dental Traumatology. 2025; 41: 437–444.

[14] Guven Y, Ozdemir OT, Kavan MY. Performance of artificial intelligence chatbots in responding to patient queries related to traumatic dental injuries: a comparative study. Dental Traumatology. 2025; 41: 338–347.

[15] 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. Pediatric Dentistry. 2024; 46: 337–344.

[16] Öztürk Z, Bal C, Çelikkaya BN. Evaluation of information provided by

ChatGPT versions on traumatic dental injuries for dental students and professionals. Dental Traumatology. 2025; 41: 427–436.
[17] American Academy of Pediatric Dentistry. Policy on the role of dental prophylaxis in pediatric dentistry. The Reference Manual of Pediatric Dentistry (pp. 95–97). American Academy of Pediatric Dentistry: Chicago, Ill. 2024.
[18] American Academy of Pediatric Dentistry. Policy on use of fluoride. The Reference Manual of Pediatric Dentistry (pp. 101–103). American Academy of Pediatric Dentistry: Chicago, Ill. 2024.
[19] American Academy of Pediatric Dentistry. Policy on dietary recommendations for infants, children, and adolescents. The Reference Manual of Pediatric Dentistry (pp. 109–113). American Academy of Pediatric Dentistry: Chicago, Ill. 2024.
[20] American Academy of Pediatric Dentistry. Periodicity of examination, preventive dental services, anticipatory guidance/counseling, and oral treatment for infants, children, and adolescents. The Reference Manual of Pediatric Dentistry (pp. 293–305). American Academy of Pediatric Dentistry: Chicago, Ill. 2024.
[21] American Academy of Pediatric Dentistry. Fluoride therapy. The Reference Manual of Pediatric Dentistry (pp. 351–357). American Academy of Pediatric Dentistry: Chicago, Ill. 2024.
[22] Malak A, Şahin MF. How useful are current chatbots regarding urology patient information? Comparison of the ten most popular chatbots' responses about female urinary incontinence. Journal of Medical Systems. 2024; 48: 102.
[23] Molena KF, Macedo AP, Ijaz A, Carvalho FK, Gallo MJD, Wanderley Garcia de Paula e Silva F, *et al*. Assessing the accuracy, completeness, and reliability of artificial intelligence-generated responses in dentistry: a pilot study evaluating the ChatGPT model. Cureus. 2024; 16: e65658.
[24] 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. 2024; 14: 2818.
[25] Rokhshad R, Zhang P, Mohammad-Rahimi H, Pitchika V, Entezari N, Schwendicke F. Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: a pilot study. Journal of Dentistry. 2024; 144: 104938.
[26] Bhatara S, Goswami M, Saxena A, Pathak P, Tuli S, Saxena B. The evolving role of social media in paediatric dentistry: a narrative review. Global Pediatrics. 2024; 9: 100221.

How to cite this article: Ceren Sağlam, Aslı Aşık, Elif Kuru, Handan Çelik, Nazan Ersin, Arzu Aykut Yetkiner, Dilşah Çoğulu. Evaluating the efficacy of large language models in providing information for parental inquiries regarding primary care of pediatric oral and dental health. Journal of Clinical Pediatric Dentistry. 2026; 50(2): 132-141. doi: 10.22514/jocpd.2026.042.