Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluating the Usefulness, Quality, Reliability, and Readability of Large Language Model Responses About Pediatric Laser Dentistry: A Parent-Oriented Study

Berkehan Aykanat

berkehanaykanat1994@gmail.com

Burdur Mehmet Akif Ersoy University

Emine Şuranur Ayaz

Recep Tayyip Erdogan University

Article

Keywords: Artificial intelligence, Large language model, Chatbot, Pediatric dentistry, Dental laser, Patient education

Posted Date: June 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9865941/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Parents increasingly use large language model (LLM)-based chatbots to seek health information, yet their reliability in pediatric laser dentistry remains unclear. This study evaluated the usefulness, quality, reliability, readability, and length of responses generated by four paid versions of LLM-based chatbots to parent-oriented questions on pediatric laser dentistry. Twenty content-validated open-ended questions (S-CVI/Ave, 0.86) were submitted once daily for seven days to ChatGPT-5.5 Thinking, Google Gemini 3.1 Pro, Claude Opus 4.7, and DeepSeek-V4, yielding 560 responses. Two blinded pediatric dentists assessed usefulness, Global Quality Scale (GQS), and CLEAR scores; readability and length were also measured. Models were compared using Friedman tests with Bonferroni-corrected Wilcoxon post hoc tests, and Kendall's W was calculated as the effect size. Inter-rater reliability was excellent. The models differed significantly across all outcomes. ChatGPT achieved the most favorable usefulness, GQS, and CLEAR scores, whereas DeepSeek produced the longest and easiest-to-read responses but the lowest expert-rated content quality and reliability. Claude and Gemini showed intermediate performance. Day-to-day differences in expert-rated outcomes were statistically significant but practically negligible. All models required reading levels above those recommended for patient education. Current LLMs may support parental education but should not replace professional pediatric dental consultation.

1. Introduction

The rapid development of artificial intelligence-based large language models (LLMs) has changed the way health-related information is accessed, interpreted, and communicated. These systems can generate human-like responses to user questions and are increasingly used by patients, caregivers, students, and healthcare professionals as easily accessible sources of information [1, 2]. In healthcare-related contexts, however, the usefulness of LLM-generated content depends not only on fluency but also on factual accuracy, reliability, clinical appropriateness, and readability [2–4]. Recent studies in dentistry have shown that although LLMs have promising potential, their responses may vary according to the model, topic, prompt structure, and evaluation method [3, 4].

This issue is particularly relevant in pediatric dentistry, where parents and caregivers often seek information before making decisions about their children's dental treatment. Previous studies evaluating LLM-generated responses to parent- or patient-oriented pediatric dental questions have reported that these systems may support caregiver education, but they may also produce incomplete, overly complex, or potentially misleading information [5–8]. Therefore, topic-specific evaluation is necessary before LLMs can be considered reliable adjuncts for parental education in pediatric dental care [6–8].

In pediatric dentistry, reliable parent-oriented information is particularly important for technology-assisted and minimally invasive approaches such as dental lasers, which are increasingly presented to parents as modern alternatives to conventional procedures. The American Academy of Pediatric Dentistry supports the judicious use of lasers in selected restorative and soft tissue procedures for pediatric patients [9]. However, the clinical value of laser use depends on appropriate case selection,

wavelength, target tissue, operating parameters, and clinician training [9, 10]. Therefore, parent-oriented information about dental lasers should clearly explain both the potential benefits and the limitations of this technology, rather than presenting lasers as painless, universally superior, or suitable for every pediatric dental procedure [9].

Despite the growing number of studies evaluating LLMs in dental education, clinical decision support, and patient communication, evidence regarding chatbot-generated information on pediatric laser dentistry remains limited [6, 8, 11, 12]. This topic is clinically relevant because parents may consult AI chatbots to understand whether laser-assisted treatment is safe, comfortable, appropriate for primary teeth, suitable for anxious or uncooperative children, or comparable with conventional dental procedures [5, 9].

Therefore, this study aimed to evaluate and compare the responses generated by four LLM-based chatbot platforms—ChatGPT-5.5 Thinking, Google Gemini 3.1 Pro, Claude Opus 4.7, and DeepSeek-V4—to parent-oriented questions regarding the use of dental lasers in pediatric dentistry. The responses were assessed in terms of usefulness, overall quality, content quality and informational reliability, readability, and response length. Moreover, because LLM outputs may vary over time, responses were collected across seven consecutive days to provide additional information on short-term temporal variability. The null hypothesis was that no significant differences would be observed among the LLMs in the assessed outcomes.

2. Materials and Methods

2.1 Study design

This study was designed to evaluate the responses generated by four large language model-based chatbots to parent-oriented questions regarding the use of dental lasers in pediatric dentistry. The responses were assessed in terms of usefulness, overall quality, content quality and informational reliability, readability, and response length.

This study did not involve patients, caregivers, patient records, biological materials, clinical images, or identifiable personal or health-related data. Therefore, formal ethics committee approval was not obtained. Human expert involvement was limited to the content validation of the question set. All content-validation experts were informed about the purpose and procedures of the content validation process, and informed consent was obtained from them prior to their participation.

2.2 Development and content validation of parent-oriented questions

An initial pool of 25 open-ended parent-oriented questions was developed to reflect common concerns regarding the use of dental lasers in pediatric dentistry. The initial questions were generated by reviewing frequently asked online parent/patient questions as well as relevant pediatric dentistry and dental laser

sources. The question pool covered general information about dental lasers, indications and limitations of laser-assisted treatment, comparison with conventional dental treatment methods, caries removal and restorative procedures, pain and local anesthesia expectations, dental anxiety and comfort, treatment of primary teeth, management of deep caries, gingival problems, tongue-tie and lip-tie procedures, root canal treatment, laser safety, postoperative care, and common misconceptions about laser-assisted dental procedures.

The content validity of the initial question pool was evaluated by a panel of five experts in pediatric dentistry. Each expert independently assessed each question using a 4-point relevance scale. For each item, the item-level content validity index (I-CVI) was calculated as the proportion of experts rating the item as relevant. Ratings of 3 or 4 on the 4-point relevance scale were considered content-valid, whereas ratings of 1 or 2 were considered not content-valid.

Questions with I-CVI values ≥ 0.80 were directly retained in the final question set. Questions with I-CVI values between 0.60 and 0.79 were revised according to expert feedback, whereas questions with I-CVI values below 0.60 were excluded from the question pool. After expert evaluation and revision, the final question set consisted of 20 open-ended parent-oriented questions. The scale-level content validity index using the average method (S-CVI/Ave) was 0.86, indicating acceptable content validity of the final question set.

All final questions were formulated in simple parent-oriented language to ensure understandability for individuals without professional dental knowledge. For each parent-oriented question, a predefined gold-standard answer framework was developed before response evaluation. Each framework included essential content elements expected in an appropriate answer and examples of critical errors or misleading claims. The gold-standard frameworks were primarily based on the American Academy of Pediatric Dentistry policy on the use of lasers for pediatric dental patients and were supplemented by previous literature and reference textbooks on dental laser applications [9, 13–15]. These frameworks were used to guide expert evaluation, particularly for the CLEAR domains of completeness, lack of false information, and evidence-supported content. The complete list of parent-oriented questions, gold-standard key points, and critical error examples is provided in Supplementary Table S1.

2.3 Large language models evaluated

Four large language model-based chatbot platforms were evaluated in this study: ChatGPT-5.5 Thinking, Google Gemini 3.1 Pro, Claude Opus 4.7, and DeepSeek-V4. All models were accessed through their official web interfaces using paid subscription-based accounts between April 30 and May 6, 2026. Paid subscription-based versions were selected to evaluate the performance of the most advanced user-facing configurations available for each platform during the study period. The model names displayed in the interfaces during the data collection period were recorded to improve methodological transparency and reproducibility.

No external plug-ins, file uploads, browsing extensions, application programming interfaces, or additional prompt-engineering tools were used. All chatbots were used with their default settings, and the same standardized parent-oriented questions, written in English, were submitted to each model without modification.

2.4 Prompting procedure and response collection

Data collection was conducted over seven consecutive days, from April 30 to May 6, 2026. Each of the 20 standardized parent-oriented questions was submitted once per day to each of the four LLMs. Accordingly, the total number of responses was calculated as follows:
20 questions × 4 LLMs × 7 days = 560 responses.

Each question was entered into a new and independent chat session in a private/incognito browser window to minimize the potential influence of previous prompts, conversation history, personalization, or contextual carryover. Where available, memory or personalization features were disabled, and no previous conversation history was used. No follow-up questions were asked, and no response regeneration was requested. Only the first complete response generated by each LLM was recorded. The wording of the questions was kept identical across all LLMs and all data collection days.

All responses were copied verbatim and recorded in a spreadsheet. For each response, the model name, question number, date of data collection, and full response text were documented. Before expert evaluation, all model identifiers were anonymized and coded as Model 1, Model 2, Model 3, and Model 4. The order of responses was randomized to ensure blinded assessment by the evaluators.

2.5 Assessment of LLM responses

The responses were independently assessed by two pediatric dentistry specialists who were blinded to the identity of the LLMs. Before the main evaluation, both evaluators completed a calibration session using an independent pilot set of 80 LLM-generated responses obtained before the main data collection. These pilot responses were generated using the same parent-oriented question format and scoring framework but were not included in the final dataset, inter-rater reliability analysis, or statistical comparisons. The calibration session was used only to standardize the interpretation and application of the assessment tools. Any discrepancies identified during calibration were discussed until consensus was reached. After calibration, the 560 study responses were anonymized, randomized, and independently evaluated. During the main evaluation, discrepancies between the two evaluators were resolved through discussion and consensus.

The assessment addressed three domains: usefulness; overall quality (GQS); and content quality and informational reliability (CLEAR).

Usefulness was assessed using a modified 4-point scale. Responses were scored as follows: 1 = very useful, indicating that the response provided comprehensive and correct information similar to what a specialist would provide; 2 = useful, indicating that the response provided accurate information but

missed some important details; 3 = partially useful, indicating that the response contained some correct information but also included incomplete, unclear, or potentially misleading information; and 4 = not useful, indicating that the response was incorrect, misleading, or unsuitable for parental information [16].

The overall quality of the responses was evaluated using the Global Quality Scale (GQS). The GQS is a 5-point scale, where 1 indicates poor quality, poor flow, missing information, and limited usefulness, while 5 indicates excellent quality, good flow, comprehensive information, and high usefulness for parental education [17].

Content quality and informational reliability of the responses were evaluated using the CLEAR tool. CLEAR assesses five domains: Completeness, Lack of false information, Evidence-supported, Appropriateness, and Relevance. Each domain was scored from 1 to 5, yielding a total CLEAR score ranging from 5 to 25. Higher scores indicated better content quality and informational reliability. Total CLEAR scores were interpreted as poor content for scores of 5–11, average content for scores of 12–18, and very good content for scores of 19–25. The “Lack of false information” and “Evidence-supported” domains were scored with reference to the predefined reference answers based on current pediatric dentistry and laser dentistry knowledge [16, 18].

The readability of the chatbot-generated responses was evaluated to determine whether the information was suitable for a parental audience. Since the questions and responses were generated in English, readability was assessed using Readable software (Readable, https://readable.com), an online readability assessment tool. Each response was analyzed separately using the following readability indices: Flesch Reading Ease (FRE), Flesch–Kincaid Grade Level (FKGL), and SMOG Index. FRE scores range from 0 to 100, with higher scores indicating easier readability. FKGL and SMOG Index estimate the school grade level required to understand the text. In addition, response length was assessed by calculating the word count and character count of each response. These parameters were used to compare the length of the LLM-generated answers [5].

Inter-rater reliability between the two evaluators was assessed using the intraclass correlation coefficient for usefulness, GQS, and CLEAR total scores. In cases of disagreement between the two evaluators, the final score was determined by consensus.

2.6 Statistical analysis

All statistical analyses were performed using SPSS software. Descriptive statistics were calculated for all variables. Ordinal and continuous variables were presented as median and interquartile range (Q1–Q3), and categorical variables as frequencies and percentages.

Because the same questions were submitted to each LLM across seven consecutive days, the matched structure of the data was considered. Comparisons among the four LLMs were performed using Friedman tests, with each day–question combination treated as a matched block. Pairwise comparisons were conducted using Wilcoxon signed-rank tests with Bonferroni correction, and Kendall’s W was calculated as the effect size.

Temporal variability was assessed using Friedman tests across the seven data collection days, with each LLM–question combination treated as a matched block. Inter-rater reliability was evaluated using single-measure intraclass correlation coefficients based on a two-way random-effects model with absolute agreement. A p-value of < 0.05 was considered statistically significant. For post hoc analyses, the Bonferroni-adjusted significance threshold was set at $p < 0.0083$.

3. Results

A total of 560 LLM-generated responses were included in the analysis, corresponding to 20 parent-oriented questions submitted to four LLMs over seven consecutive days.

Inter-rater reliability was calculated using the independent pre-consensus scores of the two evaluators. Agreement was excellent for all expert-based scoring outcomes across 560 responses. The single-measure intraclass correlation coefficients were 0.976 for usefulness (95% CI, 0.972–0.980; $p < 0.001$), 0.979 for GQS (95% CI, 0.975–0.982; $p < 0.001$), and 0.995 for CLEAR total score (95% CI, 0.995–0.996; $p < 0.001$).

The expert-based evaluation scores differed significantly among the four LLMs (Table 1). For usefulness, Friedman analysis showed a significant difference among the models ($\chi^2 = 139.380$, $p < 0.001$, Kendall's $W = 0.332$). Since lower usefulness scores indicate better performance, ChatGPT showed the most favorable usefulness score.

Table 1
Expert-based evaluation scores of the four LLMs

Outcome	ChatGPT	Claude	DeepSeek	Gemini	χ^2	p value	Kendall's W
Usefulness	1.0 (1.0–2.0)	2.0 (2.0–2.0)	2.0 (2.0–3.0)	2.0 (2.0–2.0)	139.380	< 0.001	0.332
Global Quality Scale	5.0 (4.0–5.0)	4.0 (3.0–4.0)	3.0 (2.25–4.0)	4.0 (3.0–4.0)	128.457	< 0.001	0.306
CLEAR total score	22.0 (19.0–23.0)	17.0 (14.0–19.0)	15.0 (11.0–18.0)	16.0 (13.25–18.0)	176.577	< 0.001	0.420

Data are presented as median (Q1–Q3). Friedman tests were used to compare the four LLMs. Kendall's W was reported as the effect size. For usefulness, lower scores indicate better performance. For Global Quality Scale and CLEAR total scores, higher scores indicate better performance. CLEAR, Completeness, Lack of false information, Evidence-supported, Appropriateness, and Relevance; LLM, large language model; Q1, first quartile; Q3, third quartile.

GQS scores also differed significantly among the LLMs ($\chi^2 = 128.457$, $p < 0.001$, Kendall's $W = 0.306$). ChatGPT had the highest GQS score, followed by Claude and Gemini, whereas DeepSeek had the lowest

median score. Similarly, CLEAR total scores differed significantly among the models ($\chi^2 = 176.577$, $p < 0.001$, Kendall's W = 0.420), with ChatGPT achieving the highest median CLEAR total score.

Post hoc Wilcoxon signed-rank tests with Bonferroni correction showed that ChatGPT differed significantly from all other models in usefulness, GQS, and CLEAR total scores. For usefulness and GQS, the difference between Claude and Gemini was not statistically significant after Bonferroni correction. For CLEAR total score, all pairwise comparisons were statistically significant.

[Table 1 **near here]**

Readability indices and response length parameters differed significantly among the four LLMs (Table 2). FRE scores showed a significant difference among the models ($\chi^2 = 60.673$, $p < 0.001$, Kendall's W = 0.144). DeepSeek produced the highest FRE score, indicating easier readability.

Table 2
Readability and response length of the four LLMs

Outcome	ChatGPT	Claude	DeepSeek	Gemini	χ^2	p value	Kendall's W
Flesch Reading Ease	32.35 (21.73–40.90)	38.30 (22.28–47.05)	39.70 (26.68–50.20)	32.65 (20.03–42.95)	60.673	< 0.001	0.144
Flesch–Kincaid Grade Level	13.20 (11.35–14.78)	12.45 (10.60–14.28)	11.10 (9.53–12.80)	12.80 (11.30–14.78)	76.484	< 0.001	0.182
SMOG Index	15.20 (13.73–16.40)	14.20 (12.70–15.80)	13.20 (12.30–14.68)	14.50 (13.40–15.50)	76.276	< 0.001	0.182
Word count	283.50 (211.25–342.25)	328.50 (264.25–388.50)	608.50 (489.25–730.50)	422.00 (359.00–488.50)	281.779	< 0.001	0.671
Character count	1525.00 (1170.75–1948.25)	1695.00 (1365.50–2084.00)	3061.00 (2596.00–3802.50)	2244.00 (1920.25–2740.50)	284.751	< 0.001	0.678

Data are presented as median (Q1–Q3). Friedman tests were used to compare the four LLMs. Kendall's W was reported as the effect size. For Flesch Reading Ease, higher scores indicate easier readability. For Flesch–Kincaid Grade Level and SMOG Index, lower scores indicate easier readability. Word count and character count were used to assess response length. LLM, large language model; Q1, first quartile; Q3, third quartile; SMOG, Simple Measure of Gobbledygook.

FKGL scores also differed significantly among the LLMs ($\chi^2 = 76.484$, $p < 0.001$, Kendall's W = 0.182). DeepSeek produced the lowest grade-level score, indicating that its responses required a lower reading level. Similarly, SMOG Index scores differed significantly among the models ($\chi^2 = 76.276$, $p < 0.001$, Kendall's W = 0.182), with DeepSeek showing the lowest median SMOG Index.

Response length also differed significantly among the LLMs. Word count ($\chi^2 = 281.779$, $p < 0.001$, Kendall's W = 0.671) and character count ($\chi^2 = 284.751$, $p < 0.001$, Kendall's W = 0.678) showed large effect sizes. DeepSeek generated the longest responses, whereas ChatGPT generated the shortest responses. Pairwise comparisons showed significant differences between all model pairs for both word count and character count.

[Table 2 **near here]**

Temporal variability across the seven consecutive data collection days is presented in Table 3. Significant day-based differences were observed for usefulness, GQS, and CLEAR total score. However, the corresponding Kendall's W values were very small, indicating limited temporal variability. No significant day-based differences were observed for word count, character count, FRE, FKGL, or SMOG Index.

Table 3
Temporal variability of LLM-generated responses across seven consecutive data collection days

Outcome	χ^2	p value	Kendall's W
Usefulness	28.741	< 0.001	0.060
Global Quality Scale	30.826	< 0.001	0.064
CLEAR total score	30.092	< 0.001	0.063
Word count	3.611	0.729	0.008
Character count	4.103	0.663	0.009
Flesch Reading Ease	7.262	0.297	0.015
Flesch–Kincaid Grade Level	10.600	0.102	0.022
SMOG Index	7.236	0.300	0.015

Temporal variability was assessed using Friedman tests across seven consecutive data collection days, with each LLM–question combination treated as a matched block. Kendall's W was reported as the effect size. Although usefulness, Global Quality Scale, and CLEAR total scores showed statistically significant day-based differences, Kendall's W values indicated very small effect sizes. CLEAR, Completeness, Lack of false information, Evidence-supported, Appropriateness, and Relevance; LLM, large language model; SMOG, Simple Measure of Gobbledygook.

[Table 3 **near here]**

Pairwise post hoc comparisons among the four models are presented in Table 4. ChatGPT differed significantly from Claude, DeepSeek, and Gemini across all expert-based outcomes. Among the remaining models, all expert-based comparisons involving DeepSeek were significant, whereas Claude and Gemini did not differ significantly in usefulness ($p = 0.148$) or GQS ($p = 0.028$); the difference between these two models remained significant after adjustment only for the CLEAR total score ($p = 0.008$). For readability, DeepSeek showed the most distinct profile, differing significantly from the other

models across FRE, FKGL, and SMOG Index. In contrast, ChatGPT and Gemini were comparable across all readability indices, and Claude and Gemini did not differ significantly in FKGL or SMOG Index. For response length, all six pairwise comparisons were significant for both word count and character count.

Table 4
Pairwise post hoc comparisons among the four LLMs

Outcome	ChatGPT vs Claude	ChatGPT vs DeepSeek	ChatGPT vs Gemini	Claude vs DeepSeek	Claude vs Gemini	DeepSeek vs Gemini
Usefulness	< 0.001	< 0.001	< 0.001	< 0.001	0.148	0.005
Global Quality Scale	< 0.001	< 0.001	< 0.001	< 0.001	0.028	0.007
CLEAR total score	< 0.001	< 0.001	< 0.001	< 0.001	0.008	0.005
Flesch Reading Ease	< 0.001	< 0.001	0.859	0.008	< 0.001	< 0.001
Flesch–Kincaid Grade Level	0.017	< 0.001	0.270	< 0.001	0.206	< 0.001
SMOG Index	0.001	< 0.001	0.016	< 0.001	0.555	< 0.001
Word count	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Character count	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

Pairwise comparisons were performed using Wilcoxon signed-rank tests with Bonferroni correction. Since six pairwise comparisons were performed for each outcome, the Bonferroni-adjusted significance threshold was set at $p < 0.0083$. p values below this threshold were considered statistically significant. SMOG, Simple Measure of Gobbledygook.

[Table 4 **near here]**

4. Discussion

To the best of our knowledge, this is the first study to systematically evaluate the usefulness, overall quality, content quality and informational reliability, readability, and response length of four advanced, paid LLM chatbots—ChatGPT-5.5 Thinking, Google Gemini 3.1 Pro, Claude Opus 4.7, and DeepSeek-V4—on parent-oriented questions regarding pediatric laser dentistry, using a matched-block design across seven consecutive days. The null hypothesis was rejected: global model comparisons revealed statistically significant differences among the four LLMs for every expert-based outcome and for every readability and response-length parameter assessed. The pattern of results was consistent and interpretable. ChatGPT outperformed all other models in usefulness, GQS, and CLEAR total score; DeepSeek showed the most accessible readability profile and generated the longest responses but also

exhibited the lowest expert-rated overall quality and content reliability; and Claude and Gemini occupied an intermediate position, with no significant difference between them in usefulness or GQS after Bonferroni correction. Day-to-day variability across the seven evaluation days was statistically significant for expert-based outcomes only, but with very small effect sizes (Kendall's $W \approx 0.06$), indicating practically stable short-term performance.

ChatGPT differed significantly from all other models across the expert-rated outcomes. Its median usefulness score of 1.0 corresponded to the "very useful" category, and its median GQS of 5.0 and CLEAR total score of 22.0 placed its responses in the highest quality and "very good content" ranges. In contrast, the median CLEAR total scores of Claude (17.0), Gemini (16.0), and DeepSeek (15.0) all fell within the "average content" range, although these three models still differed significantly from one another. This distinction is clinically relevant because parent-oriented information on pediatric laser dentistry should provide a balanced explanation of indications, expected benefits, limitations, safety considerations, the possible need for local anesthesia, and the importance of clinician training. Responses that remain only within the average content range may omit or underemphasize these issues, potentially limiting their value for parental education.

The pattern observed for ChatGPT is consistent with several recent comparative studies that have reported the superior performance of newer-generation OpenAI models in dental contexts. Dermata et al. found that ChatGPT-4 produced the highest expert-rated answers among six LLMs evaluated in pediatric dentistry [19]. Sezer and Okutan reported that ChatGPT-4o performed well on frequently asked questions posed by parents and on curricular pediatric dentistry items [12]. Kocaoğlu et al. similarly identified ChatGPT-4o and Gemini as the top performers in pediatric dental sedation, with Claude and Copilot performing less consistently [6]. However, the literature also contains studies in which DeepSeek- or Gemini-based models outperformed ChatGPT for parent- or patient-oriented dental queries. For frequently asked endodontic questions, DeepSeek-V3.1 achieved significantly higher accuracy than both ChatGPT-5 and Gemini 2.5 Flash [20]; for molar–incisor hypomineralization, DeepSeek R1 received higher expert-rated quality than ChatGPT-4o [21]; and for primary tooth pulpotomy, Gemini Pro produced more reliable information (DISCERN) than ChatGPT-4 [22]. Such heterogeneity across topics and model versions—occurring even with current-generation models—illustrates that LLM performance is highly task-dependent, and that platform recommendations cannot be generalized across pediatric dental subdomains. The advantage of ChatGPT in the present study should therefore be interpreted as a topic-specific finding for parent-oriented questions on pediatric laser dentistry, evaluated with the specific tools used here.

The relatively lower performance of DeepSeek on the expert-rated metrics warrants closer attention. Although DeepSeek generated the longest responses and achieved the most favorable readability scores, expert evaluation showed that this linguistic accessibility and verbosity were not matched by higher content quality, balance, or informational reliability. This dissociation—where the model producing the longest and most readable responses received the lowest expert-rated scores—indicates that response length and readability should not be interpreted as direct indicators of content quality. This

interpretation is consistent with previous pediatric dental chatbot research suggesting that linguistic complexity and information quality represent distinct dimensions of AI-generated patient education content [6, 22]. Importantly, the median usefulness score of 2.0, with an upper quartile reaching 3.0, indicates that a non-trivial proportion of DeepSeek responses contained partially correct but also potentially misleading or incomplete elements—an outcome of clinical relevance when the audience consists of parents seeking guidance for a child's treatment.

Claude and Gemini occupied an intermediate and practically similar profile, with no statistically significant difference between them in usefulness or GQS after Bonferroni correction, and only a small but statistically significant difference in CLEAR total scores. From a caregiver-education perspective, this convergence suggests that, in their current versions, these two models perform comparably as supplementary information sources, although neither reaches the level of detail and balance offered by ChatGPT [23, 24].

Comparing responses produced over seven consecutive days revealed statistically significant day-to-day differences for usefulness, GQS, and CLEAR total scores. However, the corresponding Kendall's W values were very small (all ≈ 0.06), and no statistically significant temporal variability was detected for word count, character count, FRE, FKGL, or SMOG. Taken together, these findings indicate that, while statistical testing in a matched repeated-measures dataset may detect small fluctuations, the very low Kendall's W values indicate that the practical magnitude of temporal change was negligible. From a caregiver perspective, this means that the information a parent would receive on a given day is unlikely to differ meaningfully from what they would receive on another day within the same week, at least for the four model versions and study period evaluated here. The result is consistent with prior longitudinal benchmarks in dental traumatology [25, 26], endodontics [27], and molar–incisor hypomineralization [21], where short-term variability was statistically detectable but practically limited. It is important, nevertheless, to acknowledge that LLMs are subject to silent updates, decoding stochasticity, and version drift, all of which can change response characteristics over longer time horizons. A seven-day window captures short-term stochastic variability but cannot characterize medium- or long-term drift after major model updates.

The present findings are broadly consistent with the wider dental LLM literature on three points: (i) model performance is strongly version- and task-dependent, with newer-generation models generally—but not uniformly—showing stronger expert-rated performance, and the leading model varying by dental topic and evaluation method; (ii) the readability of LLM-generated dental content consistently exceeds recommended patient-education levels [28], limiting its direct suitability for lay audiences; and (iii) although reliability and content-quality scores are improving, current LLM outputs still require expert oversight and, in many cases, professional curation before being used as patient-facing educational material. At the same time, the present study extends the literature in two important ways. First, to our knowledge, no previous study has specifically evaluated LLM performance on parent-oriented questions concerning pediatric laser dentistry, a clinically relevant and underexplored domain. Second, the use of a seven-day matched-block design with rigorous Bonferroni-corrected post hoc comparisons and Kendall's

W effect sizes provides a more robust evaluation of both inter-model differences and short-term temporal stability than the cross-sectional designs that dominate the field.

Several clinical and public-health implications emerge from these findings. Although LLM-based chatbots may provide readily available preliminary information about pediatric laser dentistry, none of the four evaluated models should currently be considered a substitute for assessment by a qualified pediatric dentist. Even ChatGPT, the best-performing model in the present analysis, did not produce content that was uniformly classified as "very useful" across all questions, and its readability profile was the most demanding among the four chatbots tested. Clinicians should therefore anticipate that parents may arrive at consultations with information shaped by AI-generated content. Asking parents whether they have used AI tools and clarifying common misconceptions—particularly the perception of laser dentistry as universally painless, anesthesia-free, or appropriate for every procedure—may become increasingly important during informed-consent discussions [9]. The persistent gap between current AI-generated content and recommended readability levels for patient education materials also underscores the need for parent-targeted prompting strategies, clinician- or institution-led simplification of AI-generated information, and the development of domain-tuned models supported by guideline-aware retrieval from authoritative pediatric dental sources. In parallel, professional societies and pediatric dental departments may consider issuing curated, evidence-based FAQs on laser dentistry written at appropriate readability levels, which could serve both as direct caregiver resources and as reference content for future retrieval-augmented AI systems.

This study has several limitations that should be acknowledged. First, only English-language queries and only paid versions of four LLMs were evaluated; performance may differ in other languages, in free-tier versions, or in regionally adapted variants. Second, each question was submitted once per day in a freshly initiated session without follow-up prompting; performance under prompt-engineering or conversational scenarios that mirror real parent-chatbot interactions may differ. Third, the seven-day evaluation window characterizes short-term stability but cannot detect drift following major model updates, which may occur outside the data-collection period. Fourth, although the question pool was developed through a structured process and validated by five pediatric dentistry experts (S-CVI/Ave = 0.86), it inevitably reflects clinician-perceived rather than directly observed parent inquiry behavior. Fifth, our evaluators were specialists in pediatric dentistry rather than parents themselves; parent-reported perceived usefulness and readability may diverge from expert-rated metrics. Sixth, the readability formulas used assess surface linguistic complexity but do not capture visual hierarchy, examples, or cultural relevance, all of which influence real-world comprehension. Finally, the CLEAR and modified usefulness tools, although increasingly used in the dental LLM literature, have not been formally validated for AI-generated content in pediatric laser dentistry.

5. Conclusion

For parent-oriented questions on pediatric laser dentistry in English, ChatGPT-5.5 Thinking provided the most useful, highest-quality, and most reliable content among the four LLMs evaluated, though with the

most demanding readability; DeepSeek-V4 produced the most readable and longest responses but the lowest expert-rated quality, while Claude Opus 4.7 and Google Gemini 3.1 Pro performed comparably in between. Across all four models, content remained harder to read than recommended for patient-education materials, and short-term performance was practically stable over seven days. Current LLMs are therefore plausible adjuncts to, but not substitutes for, professional pediatric dental consultation regarding laser-assisted treatment in children.

Declarations

Ethics approval and consent to participate

This study did not involve patients, caregivers, patient records, biological materials, clinical images, or identifiable personal or health-related data. Therefore, formal ethics committee approval was not obtained. Human expert involvement was limited to the content validation of the question set. All content-validation experts were informed about the purpose and procedures of the content validation process, and informed consent was obtained from them prior to their participation.

Competing interests

The authors declare no competing interests.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

B.A.: Conceptualization, methodology, investigation, data curation, formal analysis, validation, writing – original draft, writing – review and editing, project administration, and supervision. E.S.A.: Investigation, data curation, writing – original draft, and writing – review and editing. Both authors contributed to the interpretation of the findings, critically reviewed the manuscript, and approved the final version.

Data Availability

The dataset supporting the findings of this study, including anonymized model-coded responses, expert consensus scores, readability metrics, and response-length variables, is provided as Supplementary Dataset S2. The parent-oriented questions, gold-standard key points, and examples of critical errors used for response evaluation are provided in Supplementary Table S1.

References

1. Aydin, S., Karabacak, M., Vlachos, V. & Margetis, K. Large language models in patient education: a scoping review of applications in medicine. *Front. Med. (Lausanne)*. **11**, 1477898. 10.3389/fmed.2024.1477898 (2024).
2. Busch, F. et al. Current applications and challenges in large language models for patient care: a systematic review. *Commun. Med. (Lond)*. **5**, 26. 10.1038/s43856-024-00717-2 (2025).
3. Sivaramakrishnan, G. et al. Assessing the power of AI: a comparative evaluation of large language models in generating patient education materials in dentistry. *BDJ Open.* **11**, 59. 10.1038/s41405-025-00349-1 (2025).
4. Lv, X. et al. Leveraging Large Language Models for Improved Patient Access and Self-Management: Assessor-Blinded Comparison Between Expert- and AI-Generated Content. *J. Med. Internet Res.* **26**, e55847. 10.2196/55847 (2024).
5. Sağlam, C. et al. Evaluating the efficacy of large language models in providing information for parental inquiries regarding primary care of pediatric oral and dental health. *J. Clin. Pediatr. Dentistry*. **50**, 132–141. 10.22514/jocpd.2026.042 (2026).
6. Kocaoğlu, M. H., Demirel, A. & Kaya, İ. Accuracy, quality, and readability analyses of responses from large language models to questions on pediatric dental sedation. *BMC Oral Health*. **26** 10.1186/s12903-026-08026-x (2026).
7. Karamüftüoğlu, N., Varol, E. A. & Bal, C. Exploring artificial intelligence chatbots in pediatric fluoride education: a cross-sectional study. *Sci. Rep.* **16**, 182. 10.1038/s41598-025-28857-y (2025).
8. Akitomo, T., Hamada, M., Hamaguchi, S., Kusaka, S. & Nomura, R. Generative Artificial Intelligence and Large Language Models in Paediatric Dentistry: A Scoping Review. *Int. Dent. J.* **76**, 109401. 10.1016/j.identj.2025.109401 (2026).
9. American Academy of Pediatric Dentistry. in. *The Reference Manual of Pediatric Dentistry* 141–144 (American Academy of Pediatric Dentistry, 2025).
10. FDI World Dental Federation. *Lasers in dentistry*, < (2024). https://www.fdiworlddental.org/sites/default/files/2024-11/DPS4-Lasers-in-Dentistry.pdf
11. Bayraktar Nahir, C. Can ChatGPT be guide in pediatric dentistry? *BMC Oral Health*. **25** 10.1186/s12903-024-05393-1 (2025).
12. Sezer, B. & Okutan, A. E. Evaluation of ChatGPT-4's performance on pediatric dentistry questions: accuracy and completeness analysis. *BMC Oral Health*. **25**, 1427. 10.1186/s12903-025-06791-9 (2025).
13. Grzech-Leśniak, K. Lasers in Dentistry–Overview Based on FDI Policy Statement. *Int. Dent. J.* **76**, 109369 (2026).
14. Zhang, O. L. et al. Advanced Lasers and Their Applications in Dentistry. *Dent. J. (Basel)*. **13** 10.3390/dj13010037 (2025).
15. Coluzzi, D. J. & Parker, S. P. A. *Lasers in Dentistry: Current Concepts* 2nd edn (Springer, 2024).

16. Al-Haddad, A., Alrabadi, M., Saadeh, O., Alrabadi, G. & Hassona, Y. The evaluation of tooth whitening from a perspective of artificial intelligence: a comparative analytical study. *Front. Digit. Health*. **7**, 1710159 (2025).
17. Ozdemir, Z. M. & Yapici, E. Evaluating the accuracy, reliability, consistency, and readability of different large language models in restorative dentistry. *J. Esthetic Restor. Dentistry*. **37**, 1740–1752 (2025).
18. Şen, S. C., Atagün, Ö. S., Ustaoğlu, G., Yıldız, Z. H. & Kayacı, R. N. Do AI Chatbots Tell the Truth About Dentin Hypersensitivity? A Comparative Evaluation of Quality, Accuracy, and Readability. *Cumhuriyet Dent. J.* **29**, 138–147 (2026).
19. Dermata, A., Arhakis, A., Makrygiannakis, M. A., Giannakopoulos, K. & Kaklamanos, E. G. Evaluating the evidence-based potential of six large language models in paediatric dentistry: a comparative study on generative artificial intelligence. *Eur. Arch. Paediatr. Dent.* **26**, 527–535. 10.1007/s40368-025-01012-x (2025).
20. Taşyürek, M., Adıgüzel, Ö., Gündoğar, M., Goncharuk-Khomyn, M. & Ortaç, H. Comparative evaluation of the responses from ChatGPT-5, gemini 2.5 Flash, and DeepSeek-V3. 1 chatbots to patient inquiries about endodontic treatment in terms of accuracy, understandability, and readability. *Int. Dent. Res.* **15**, 123–135 (2025).
21. Uçar Gündoğar, Z. & Sarıoğlu, D. Comparative assessment of quality, consistency, and reference accuracy of MIH-related clinical information generated by ChatGPT-4o and DeepSeek R1. *BMC Oral Health*. **26** 10.1186/s12903-026-07911-9 (2026).
22. Aydin Varol, E., Ozturk, Z., Bal, C. & Karamuftuoglu, N. Comparison of two different artificial intelligence chatbots that provide information to patients and parents about primary tooth pulpotomy treatments. *BMC Oral Health*. **26** 10.1186/s12903-025-07449-2 (2025).
23. Gökcek Taraç, M. & Nale, T. Artificial intelligence in pediatric dental trauma: do artificial intelligence chatbots address parental concerns effectively? *BMC Oral Health*. **25**, 736. 10.1186/s12903-025-06105-z (2025).
24. Sözen, E., Akpınar, H. & Yaman, S. Comparative performance of large language models for patient-oriented support in dental trauma emergencies. *BMC Oral Health*. **26** 10.1186/s12903-026-08037-8 (2026).
25. Lisboa, R. M. et al. Performance of five free large language models in dental trauma: a 30-day longitudinal benchmark study. *Front. Oral Health*. **6**, 1737114. 10.3389/froh.2025.1737114 (2025).
26. Mustuloğlu, Ş. & Deniz, B. P. Evaluation of Chatbots in the Emergency Management of Avulsion Injuries. *Dent. Traumatol.* **41**, 437–444. 10.1111/edt.13041 (2025).
27. Şimşek, E. & Büker, M. Effect of language difference and time on the accuracy of artificial intelligence chatbots responses to questions about vital pulp therapy. *BMC Oral Health*. **26** 10.1186/s12903-026-08060-9 (2026).
28. Eltorai, A. E., Yang, N. P. T., Daniels, H., Born, C. T. & A. H. & Readability of Trauma-Related Patient Education Materials From the American Academy of Orthopaedic Surgeons. *Trauma. Mon*. **21**, e20141. 10.5812/traumamon.20141 (2016).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS1.pdf
- SupplementaryDatasetS2.xlsx