Submitted: 08 April, 2025 Accepted: 12 May, 2025 Published: 03 January, 2026 DOI:10.22514/jocpd.2026.014

ORIGINAL RESEARCH

Journal of Clinical Pediatric Dentistry

Bridging the information gap in pediatric dentistry: a comparison of ChatGPT-4o, Google Gemini Advanced, and expert responses based on evaluations by parents and pediatric dentists

İsmail Haktan Çelik[1,*], Hasan Camcı[2], Farhad Salmanpour[2]

[1]Department of Pediatric Dentistry, Afyonkarahisar Health Sciences University, 03030 Afyonkarahisar, Turkey
[2]Department of Orthodontics, Afyonkarahisar Health Sciences University, 03030 Afyonkarahisar, Turkey

***Correspondence**
haktan.celik@afsu.edu.tr
(İsmail Haktan Çelik)

Abstract

Background: This study aimed to evaluate the accuracy and adequacy of responses provided by ChatGPT-4o and Google Gemini Advanced to common pediatric dentistry questions posed by parents, and to compare these responses with those given by pediatric dentistry experts. **Methods**: Fifty-seven questions were extracted from the "Frequently Asked Questions by Parents" section of the International Association of Paediatric Dentistry (IAPD) website. Based on a preliminary survey of 20 pediatric dentists, the 15 most frequently asked questions were selected. For each question, three responses (from experts, ChatGPT-4o and Google Gemini Advanced) were collected and assessed for readability using the Flesch-Kincaid test. The responses were then randomized and included in a survey completed by 47 pediatric dentists and 101 parents, who rated the adequacy of each answer on a scale from 1 (insufficient) to 10 (very sufficient). **Results**: Pediatric dentists consistently rated expert answers higher than Artificial Intelligence (AI)-generated responses, with significant differences observed for 13 out of 15 questions ($p < 0.05$). In contrast, parents showed varying levels of satisfaction, with no significant differences found in their ratings for eight of the questions. In some instances, AI-generated answers were rated comparably or even higher than expert responses by parents. **Conclusions**: The alignment between expert opinions and AI-generated responses remained inconsistent. While pediatric dentists generally found expert answers more satisfactory, parents occasionally preferred chatbot-generated answers depending on the question. These findings suggested that AI-powered chatbots hold promise for the future of patient education in pediatric dentistry, though expert oversight remains essential.

Keywords

Pediatric dentistry; Orthodontics; Artifical intelligence; ChatGPT-4o; Google Gemini Advanced

1. Introduction

Artificial Intelligence (AI) refers to computer systems capable of performing tasks that normally require human intelligence. Studies in this field aim to develop intelligent machines that can simulate human-like thinking and decision-making processes, combining disciplines such as computer science, mathematics and psychology. AI systems can analyze large data sets, recognize patterns, make predictions, and improve what they have learned over time. Additionally, through technologies such as natural language processing, speech recognition and computer vision, they also have the ability to interact with humans [1].

A subfield of AI, Large Language Models (LLMs), contributes to the provision of more efficient and accessible healthcare services in fields such as medicine and dentistry [2]. LLMs are neural network-based systems trained on large text datasets such as Wikipedia, digital books, scientific articles and web pages [3]. Using deep learning algorithms, they aim to generate meaningful, coherent, and human-like responses based on a given text [4]. Unlike traditional search engines, they save time by presenting information directly in text form, sparing users from browsing through various sources, and offering a user-friendly experience [5].

One of the most well-known models in this field, ChatGPT, was first publicly released by OpenAI at the end of 2022 (version 3.5, San Francisco, CA, USA) and quickly reached millions of users. Subsequently, ChatGPT-4 was introduced in 2023, followed by the GPT-4o model, which can also process visual inputs, in 2024 [6]. ChatGPT stands out

for its easy accessibility, user-friendly interface and creative, fluent responses [7]. Similarly, Google introduced Bard based on Language Model for Dialogue Applications (LaMDA), followed by the Gemini model based on Pathways Language Model (PaLM) 2 in February 2024. Gemini is notable for its multimodal structure that can analyze text, audio and visual content simultaneously, and it has reached a broad user base due to its feature of providing free and unlimited interaction [8]. While Gemini stands out with its information density and comprehensive responses, ChatGPT models distinguish themselves with fluency and creative content generation [9]. Both models are developed through interactive learning methods based on user feedback [10].

The implementation of these large language models (LLMs), particularly in the form of chatbots, has also introduced various potential applications in specialized fields such as pediatric dentistry. In recent years, AI-powered chatbots have begun to be used in areas such as oral health education and counseling for children. For example, a chatbot named COSC (Chatbot for Oral Self-Care) developed in Taiwan aimed to improve the tooth brushing habits of children aged 6–12. In a pilot study, the usability score of this chatbot was reported as 79.91 out of 100, and user satisfaction was found to be quite high [11]. In another study conducted in the United States, the responses of ChatGPT and similar chatbots to 30 true/false questions related to pediatric dentistry were compared with those of expert dentists and students. In this study, pediatric dentists achieved an accuracy rate of 96.7%, while ChatGPT performed the best among the chatbots with a 78% accuracy rate. However, it was emphasized that chatbots cannot yet replace human clinicians in clinical decision-making processes [12]. In another study conducted in South Korea, the responses of ChatGPT-3.5 and Gemini chatbots to national exam questions related to pediatric dentistry were evaluated. Although both models had similar accuracy rates, they failed to reach the minimum score required to pass the exam. These results indicate that while chatbots may be useful in educational and counseling contexts, they still have significant limitations in clinical decision-making processes [13].

In order to better understand the potential contributions and current limitations of AI-supported chatbots in pediatric dentistry, this study aims to compare the information provided by two different AI-supported chatbots—ChatGPT-4o and Google Gemini Advanced—in response to frequently asked questions by parents in the field of pediatric dentistry, with the responses given by pediatric dentistry specialists. The main objective of the study is to evaluate the extent to which the information provided by these AI systems aligns with expert opinions in terms of accuracy and clinical adequacy, and to analyze the parents' perception of the adequacy of these responses.

2. Material and methods

Before initiating the study, approval was obtained from the Afyonkarahisar Health Sciences University Scientific Research Ethics Committee (ID: 2025/1). The study was conducted in accordance with the principles outlined in the Declaration of Helsinki [14]. Written and verbal informed consent was obtained from the parents and pediatric dentists. Inclusion criteria required that parents had at least a high school education. Exclusion criteria included parents who were primary or secondary school graduates or illiterate, those whose children had not received treatment in our pediatric dentistry clinic, and those who did not volunteer to participate in the study. In the first stage, on 20 January 2025, a total of 57 questions were selected from the "Questions Frequently Asked by Parents" section, categorized by age groups, on the website of the International Association of Paediatric Dentistry (IAPD) (https://iapdworld.org/parents/). Twenty pediatric dentists working in our clinic independently reviewed these questions and were asked to identify the ones most frequently asked by parents based on their clinical experience. Based on the frequency of selections made by the experts, a list of the 32 most frequently chosen questions was created. To ensure both representativeness and impartiality, these 32 questions were sequentially numbered and randomly ordered using an online tool (https://www.random.org/lists/). The first 15 questions resulting from this process were selected for evaluation in the study (**Supplementary material**).

Three different response options were prepared for each selected question. For the first response option, the original answers available on the IAPD website were evaluated on 27 January 2025, by two pediatric dentists and one orthodontist, each with at least five years of clinical experience. The experts were asked to assess the accuracy and adequacy of these responses and to make revisions where necessary. The answers finalized through this evaluation process were accepted as expert responses.

For the second and third response options, the same questions were submitted without any modifications on 28 January 2025, to two different AI-based chatbot systems: ChatGPT-4o (https://chat.openai.com/) and Google Gemini Advanced (https://gemini.google.com/app?hl=tr). The three responses obtained for each question (provided by the experts, ChatGPT-4o and Google Gemini Advanced) were randomly assigned to the questions using the online randomization service offered by Random.org (**Supplementary material**).

Before the questions were included in the survey, the Flesch-Kincaid reading ease test was administered to assess the readability of all responses obtained from the three different sources and to investigate the impact of readability on the participants' evaluations. This test was used to determine the understandability of the responses and how their complexity level affected the evaluation process. After the necessary adjustments were made, the survey was created on the Google Forms platform. The survey consisted of 45 scoring questions (15 questions, each with three different responses) and four demographic questions. Participants were asked to rate the adequacy of the three different responses provided for each question on a scale from 1 (less adequate) to 10 (very adequate). The survey and informed consent form were sent to participants via email and WhatsApp Messenger (v2.24.2.10, WhatsApp Inc, Menlo Park, CA, USA). The sample size was calculated using G*Power software (version 3.0.10, Franz Faul, Christian-Albrechts-Universität, Kiel, SH, Germany), which indicated that a minimum of 111 participants

was required (effect size = 0.3; significance level = 0.05; power = 0.90).

All statistical procedures were performed using IBM SPSS software (version 27.0.1.0, Armonk, NY, USA). In the first stage, Cronbach's alpha coefficient was calculated to assess the internal consistency of the survey, and reliability was tested. Next, the Jarque-Bera test was applied to determine whether the data followed a normal distribution. Descriptive statistics, including mean and standard deviation, were calculated for each variable. For normally distributed data, one-way analysis of variance (ANOVA) was used to assess differences between groups, followed by pairwise comparisons using the Bonferroni *post-hoc* test; corrected *p*-values were reported in this analysis. For non-normally distributed data, group comparisons were made using the Kruskal-Wallis test, with pairwise comparisons examined using the Tamhane *post-hoc* test. In pairwise comparisons, the Independent Samples *T*-Test was used for normally distributed data, while the Mann-Whitney U test was used for non-normally distributed data. A statistical significance level of $p < 0.05$ was set. Additionally, the Flesch-Kincaid readability test was applied to evaluate the readability level of the texts.

3. Results

In this study, the findings obtained from the survey, in which patient guardians and pediatric dentists evaluated the responses provided by the specialist, ChatGPT-4o and Google Gemini Advanced, are presented below. A total of 148 individuals participated in the study, consisting of 47 pediatric dentists and 101 parents. The mean age of the pediatric dentists was 30.14 ± 3.65 years, while the mean age of the parents was 26.88 ± 6.49 years. The Cronbach's alpha reliability test yielded a value of 0.957, indicating excellent internal consistency. The results of the Flesch-Kincaid readability test were 57.11 ± 15.42 for expert responses, 53.26 ± 17.25 for ChatGPT-4o responses and 53.59 ± 18.81 for Google Gemini Advanced responses.

The results of the statistical analysis regarding the evaluation of pediatric dentists' responses by specialists, ChatGPT-4o and Google Gemini Advanced are presented in Table 1. To provide a visual overview of the adequacy scores given by pediatric dentists, a bar chart comparing the mean ratings for expert, ChatGPT-4o and Google Gemini Advanced responses across all 15 questions is presented in Fig. 1. It was observed that pediatric dentists predominantly rated the specialist responses higher compared to the responses from ChatGPT-4o and Google Gemini Advanced. Specifically, for questions such as Q1 (7.64 ± 2.09, $p < 0.001$) and Q5 (8.81 ± 1.17, $p = 0.001$), specialist responses received significantly higher scores, while ChatGPT-4o responses (Q1: 5.72 ± 2.36, Q5: 7.39 ± 2.02) received the lowest ratings. Additionally, no significant differences were observed between the groups for Q3 and Q10.

The evaluations of patient guardians regarding the responses from the specialist, ChatGPT-4o and Google Gemini Advanced are presented in Table 2. To further illustrate how patient guardians evaluated the responses from each source, Fig. 2 presents a comparative bar chart showing the mean adequacy scores for the expert, ChatGPT-4o and Google Gemini Advanced responses across all 15 questions. According to the data, it was found that patient guardians generally preferred the expert responses. However, in some instances, particularly for certain questions, the answers from ChatGPT-4o and Google Gemini Advanced received scores that were either close to or even higher than those of the expert responses. For question Q2, the expert answers (6.06 ± 3.01) received the lowest score, while ChatGPT-4o (7.09 ± 2.89) and Google Gemini Advanced (6.91 ± 2.39) were rated higher ($p = 0.021$). Conversely, for questions such as Q6 ($p = 0.008$) and Q13 ($p = 0.003$), expert answers were rated higher than those from the artificial intelligence models.

When comparing the scores given by pediatric dentists and patient guardians, significant differences were observed for some questions. As shown in Table 3, pediatric dentists rated the expert answers significantly higher than patient guardians, particularly for Q2 (7.50 ± 2.31 *vs.* 6.06 ± 3.01) and Q12 (9.03 ± 1.03 *vs.* 8.38 ± 1.90). On the other hand, in some cases, patient guardians provided more favorable evaluations than pediatric dentists, such as for the ChatGPT-4o answers to Q1 (7.07 ± 2.47) ($p = 0.005$). This finding suggests that while expert answers were generally rated highly by patient guardians, artificial intelligence responses provided acceptable information in certain instances and were rated more positively than expected by patient guardians. Tables 4 and 5 present the results for the evaluation of ChatGPT-4o and Google Gemini Advanced responses by pediatricians and patient guardians. Pediatric dentists rated ChatGPT-4o responses lower for some questions (Q1, Q9, Q10), whereas patient guardians gave these responses higher scores. Similarly, while pediatric dentists Google Gemini Advanced responses are more favorably rated for certain questions, patient guardians are rated Gemini Advanced responses higher for others (Q2, Q10). In conclusion, the study indicates that while pediatricians generally consider expert responses to be more competent, AI-based systems' responses were also positively received by patient guardians for specific questions, such as Q2.

4. Discussion

In the digitalized world, people are increasingly trying to obtain information about health issues from the internet through written or video content [15]. Although this may seem appealing to patients, many researchers continue to investigate how much both textual and video content overlap with expert opinions [16, 17]. With the recent spread of artificial intelligence-supported chatbots, patients have begun to turn to these chatbots for questions about their health issues [11, 18]. For this reason, researchers from many branches and dentistry have begun to conduct studies measuring the adequacy levels of chatbots in health-related questions [19, 20]. However, to our knowledge, there is a limited number of studies investigating the adequacy of current chatbots in pediatric dentistry [12, 21]. Therefore, the aim of this study was to compare the answers given by two current chatbots, ChatGPT-4o and Google Gemini Advanced, to questions frequently asked by parents in pediatric dentistry with the answers given by specialist doctors to the same questions.

TABLE 1. Evaluation results by pediatric dentists of responses provided by experts, ChatGPT-4o, and Google Gemini Advanced to questions asked.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
Expert's responses	7.64 ± 2.09^{A}	7.50 ± 2.31^{A}	8.08 ± 1.59^{A}	8.94 ± 1.17^{A}	8.81 ± 1.17^{A}	8.00 ± 1.88^{AB}	8.83 ± 1.40^{A}	8.39 ± 1.48^{A}	8.42 ± 1.61^{A}	7.81 ± 1.53^{A}	8.03 ± 1.42^{A}	9.03 ± 1.03^{A}	7.78 ± 1.38^{A}	8.58 ± 1.02^{A}	8.31 ± 1.04^{A}
ChatGPT-4o responses	5.72 ± 2.36^{B}	5.44 ± 2.96^{B}	7.94 ± 1.37^{A}	8.31 ± 1.33^{AB}	7.39 ± 2.02^{B}	7.75 ± 1.34^{B}	8.28 ± 1.41^{AB}	7.83 ± 1.68^{AB}	6.36 ± 2.02^{B}	7.22 ± 1.68^{A}	7.19 ± 1.35^{B}	7.53 ± 1.11^{B}	5.22 ± 2.11^{B}	7.08 ± 1.54^{B}	7.19 ± 1.86^{B}
Gemini Advance responses	7.58 ± 2.10^{A}	5.22 ± 2.58^{B}	8.39 ± 1.38^{A}	7.69 ± 1.31^{B}	7.94 ± 1.47^{B}	8.89 ± 1.09^{A}	7.53 ± 1.59^{B}	7.44 ± 1.56^{B}	7.94 ± 1.45^{A}	7.69 ± 1.41^{A}	7.83 ± 1.40^{AB}	7.28 ± 1.26^{B}	7.47 ± 1.23^{A}	7.56 ± 1.80^{B}	8.94 ± 1.07^{A}
p	<0.001ϰ	<0.001Δ	0.327ϰ	<0.001Δ	0.001ϰ	0.001ϰ	<0.001ϰ	0.017ϰ	<0.001ϰ	0.238Δ	0.033Δ	<0.001Δ	<0.001Δ	<0.001Δ	<0.001Δ

Note. Values are presented as mean ± standard deviation.
$p < 0.05$.
Q: Question; Δ: p values for one way ANOVA test; ϰ: p values for Kruskal Wallis test. ^{A,B}: Different capital letters (A and B) represent statistically significant differences between groups within the same columns.

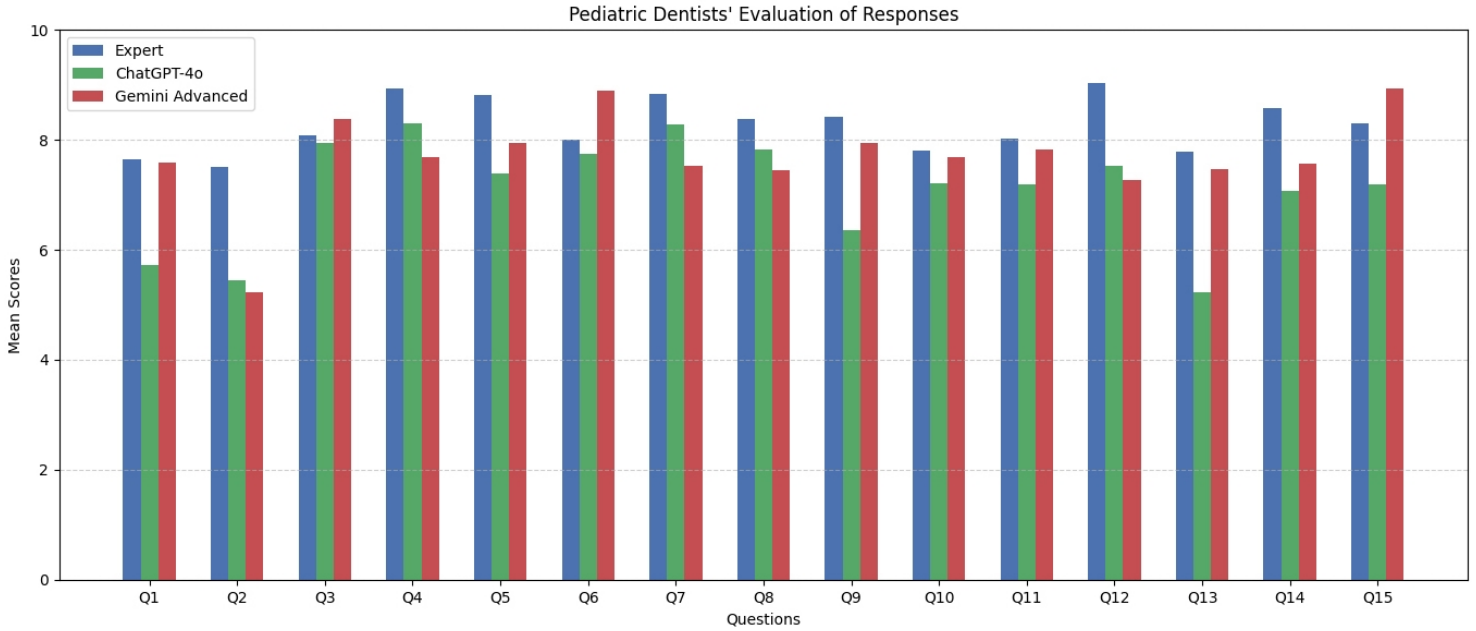

FIGURE 1. Pediatric dentists' evaluation of responses.

TABLE 2. Evaluation results by pediatric patients' parents of responses provided by experts, ChatGPT-4o and Google Gemini advanced to questions asked.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
Expert's responses	7.69 ± 2.32^{A}	6.06 ± 3.01^{A}	7.78 ± 2.40^{A}	8.49 ± 1.74^{A}	8.66 ± 1.69^{A}	8.62 ± 1.55^{A}	8.61 ± 1.48^{A}	8.23 ± 1.80^{A}	8.37 ± 1.58^{A}	8.19 ± 1.82^{A}	8.04 ± 1.81^{A}	8.38 ± 1.90^{A}	8.08 ± 1.84^{A}	8.48 ± 1.67^{A}	8.39 ± 1.57^{AB}
ChatGPT-4o responses	7.07 ± 2.47^{AB}	7.09 ± 2.89^{B}	8.03 ± 2.08^{A}	8.24 ± 2.02^{A}	8.30 ± 1.79^{A}	7.97 ± 2.01^{B}	8.14 ± 1.73^{AB}	7.87 ± 1.94^{A}	7.54 ± 2.09^{B}	8.19 ± 1.93^{A}	7.52 ± 2.09^{A}	7.80 ± 1.89^{A}	7.10 ± 2.33^{B}	7.74 ± 1.96^{B}	7.89 ± 1.98^{B}
Gemini Advanced responses	6.43 ± 3.23^{B}	6.91 ± 2.39^{B}	8.27 ± 2.05^{A}	7.70 ± 2.15^{A}	7.89 ± 2.08^{A}	7.87 ± 1.89^{B}	7.92 ± 1.84^{B}	8.03 ± 1.74^{A}	8.46 ± 1.62^{A}	8.36 ± 1.78^{A}	7.94 ± 2.03^{A}	7.85 ± 1.92^{A}	7.81 ± 1.95^{A}	7.65 ± 1.10^{B}	8.79 ± 1.55^{A}
p	0.005Δ	0.021Δ	0.706ϰ	0.522ϰ	0.126ϰ	0.008Δ	0.014Δ	0.379Δ	<0.001Δ	0.889ϰ	0.145Δ	0.060Δ	0.003Δ	0.005ϰ	0.001Δ

Note. Values are presented as mean ± standard deviation.
$p < 0.05$.
Q: Question; Δ: p values for one way ANOVA test; ϰ: p values for Kruskal Wallis test; ^{A,B}: Different capital letters (A and B) represent statistically significant differences between groups within the same columns.

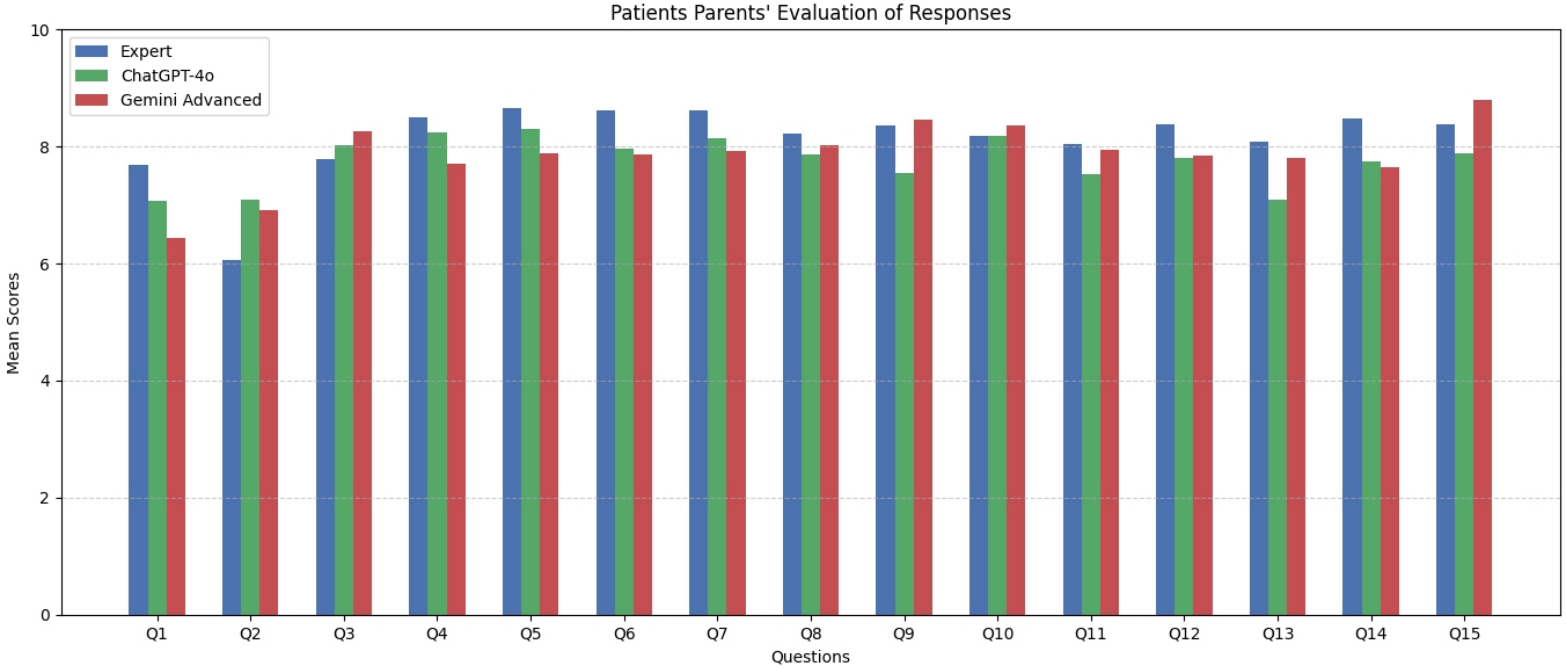

FIGURE 2. Patients parents' evaluation of responses.

TABLE 3. Statistical results on the evaluation of expert responses by pediatric dentists and parents.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
Pediatric dentist	7.64 ± 2.09	7.50 ± 2.31	8.08 ± 1.59	8.94 ± 1.17	8.81 ± 1.17	8.00 ± 1.88	8.83 ± 1.40	8.39 ± 1.48	8.42 ± 1.61	7.81 ± 1.53	8.03 ± 1.42	9.03 ± 1.03	7.78 ± 1.38	8.58 ± 1.02	8.31 ± 1.04
Parents	7.69 ± 2.32	6.06 ± 3.01	7.78 ± 2.40	8.49 ± 1.74	8.66 ± 1.69	8.62 ± 1.55	8.61 ± 1.48	8.23 ± 1.80	8.37 ± 1.58	8.19 ± 1.82	8.04 ± 1.81	8.38 ± 1.90	8.08 ± 1.84	8.48 ± 1.67	8.39 ± 1.57
p	0.656‡	0.004§	0.303§	0.305‡	0.678‡	0.049‡	0.445‡	0.976‡	0.875‡	0.223§	0.971§	0.013§	0.307§	0.644‡	0.718§

Note. Values are presented as mean ± standard deviation.
$p < 0.05$.
Q: Question; §: p values for independent sample student t test; ‡: p values for Mann Whitney U test.

TABLE 4. Statistical results on the evaluation of ChatGPT-4o responses by pediatric dentists and parents.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
Pediatric dentists	5.72 ± 2.36	5.44 ± 2.96	7.94 ± 1.37	8.31 ± 1.33	7.39 ± 2.02	7.75 ± 1.34	8.28 ± 1.41	7.83 ± 1.68	6.36 ± 2.02	7.22 ± 1.68	7.19 ± 1.35	7.53 ± 1.11	5.22 ± 2.11	7.08 ± 1.54	7.19 ± 1.86
Parents	7.07 ± 2.47	7.09 ± 2.89	8.03 ± 2.08	8.24 ± 2.02	8.30 ± 1.79	7.97 ± 2.01	8.14 ± 1.73	7.87 ± 1.94	7.54 ± 2.09	8.19 ± 1.93	7.52 ± 2.09	7.80 ± 1.89	7.10 ± 2.33	7.74 ± 1.96	7.89 ± 1.98
p	0.005‡	0.004‡	0.783‡	0.537§	0.013‡	0.466‡	0.638‡	0.706§	0.004‡	0.001§	0.292‡	0.305‡	<0.001‡	0.072‡	0.069‡

Note. Values are presented as mean ± standard deviation.
$p < 0.05$.
Q: Question; ‡: p values for independent sample student t test; §: p values for Mann Whitney U test.

TABLE 5. Statistical results on the evaluation of Google Gemini advanced responses by pediatric dentists and parents.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
Pediatric dentists	7.58 ± 2.10	5.22 ± 2.58	8.39 ± 1.38	7.69 ± 1.31	7.94 ± 1.47	8.89 ± 1.09	7.53 ± 1.59	7.44 ± 1.56	7.94 ± 1.45	7.69 ± 1.41	7.83 ± 1.40	7.28 ± 1.26	7.47 ± 1.23	7.56 ± 1.80	8.94 ± 1.07
Parents	6.43 ± 3.23	6.91 ± 2.39	8.27 ± 2.05	7.70 ± 2.15	7.89 ± 2.08	7.87 ± 1.89	7.92 ± 1.84	8.03 ± 1.74	8.46 ± 1.62	8.36 ± 1.78	7.94 ± 2.03	7.85 ± 1.92	7.81 ± 1.95	7.65 ± 1.10	8.79 ± 1.55
p	0.017‡	<0.001‡	0.580§	0.986‡	0.866‡	<0.001‡	0.260‡	0.077‡	0.039§	0.045‡	0.731‡	0.047‡	0.236‡	0.794‡	0.515‡

Note. Values are presented as mean ± standard deviation.
$p < 0.05$.
Q: Question; ‡: p values for independent sample student t test; §: p values for Mann Whitney U test.

The aim of this study is to fill a gap in the pediatric dentistry literature by investigating the adequacy and accuracy of artificial intelligence-supported chatbots in the field of pediatric dentistry.

The current study has highlighted both the potential and limitations of rapidly evolving AI-supported chatbot technology in specialized fields such as pediatric dentistry. In some cases, no significant difference was found between expert opinions and chatbot responses, indicating that chatbots have the potential to provide answers similar to those of experts for certain questions. However, this competence was not consistent across all questions. This finding underscores the limitation that chatbot responses may not always offer satisfactory answers for every query. We suggest that to improve the accuracy of these AI-based applications, it is essential for them to gain more clinical experience from seasoned physicians and to receive feedback from patients regarding the disease and treatments being addressed.

The findings of our study revealed that pediatric dentists consistently rated expert responses higher, whereas parents tended to evaluate the responses provided by AI-supported chatbots more positively. A possible explanation for this difference is that large language models such as ChatGPT-4o and Google Gemini Advanced are designed to present information in a simplified, clear, and accessible manner in order to optimize user experience [22]. The linguistically simplified and easily understandable nature of AI-generated responses may have led parents—who lack specialized knowledge—to assign higher scores. On the other hand, the lower ratings given by pediatric dentists to the responses provided by ChatGPT-4o and Google Gemini Advanced, compared to expert answers, may stem from perceived deficiencies in content accuracy. In a study conducted in Türkiye by Şişmanoğlu and colleagues, the answers given by ChatGPT-4.0 and Gemini Advanced to questions from the Dentistry Specialty Exam (DUS) were analyzed. Although both chatbots achieved scores sufficient to pass the exam, they were found to fall short, particularly in areas requiring clinical knowledge [23]. This suggests that participants with expert knowledge are more likely to identify information gaps in AI-generated responses and, therefore, adopt a more critical perspective in their evaluations.

Another notable finding is that the adequacy scores of the responses varied depending on the type of chatbot. For instance, ChatGPT-4o received higher scores for some questions, while Google Gemini Advanced scored higher for others. This highlights the unpredictability of which chatbot will provide a more adequate response to a given question. The variation in adequacy scores between the two chatbots may stem from differences in their algorithms for data processing and text generation. A study by Qi *et al.* [3] comparing ChatGPT-4V and Google Gemini, identified these differences and concluded that each model has its own unique strengths and specialized niches. In a similar study, Omar and colleagues reported comparable results [24]. On the other hand, Koç and Tiryaki found that the performances of Google Gemini Advanced and ChatGPT-4 were quite similar in their study. However, the researchers compared the two models specifically in the context of medical imaging. The authors attributed this similarity to the fact that ChatGPT is often more understandable and user-friendly in certain scenarios, such as data presentation, while Google Gemini is better equipped to analyze larger data sets and provide more detailed information [25].

The literature presents varied findings regarding the performance of AI-supported chatbots in healthcare applications. For instance, Schmidt *et al*. [26] reported that ChatGPT-4 can serve as an effective support tool for preoperative patient education, providing comprehensive and readable information to patients considering penile prosthesis implantation. Similarly, in the study conducted by Reyhan *et al.* [27], which evaluated the reliability of answers provided by five different chatbots to questions about keratoconus, the researchers found that while all models produced satisfactory results, Google Gemini and Microsoft Copilot offered more reliable and higher-quality answers.

Huo *et al.* [20] evaluated the performance of four different chatbots (ChatGPT-4, Copilot, Google Bard and Perplexity) in the context of surgical management for gastroesophageal reflux disease. Their findings indicated that Google Bard and ChatGPT-4 achieved high accuracy rates, while Microsoft Copilot and Perplexity showed low accuracy rates [20]. Similarly, Lim *et al.* [19] assessed the accuracy of ChatGPT-3.5, Claude, Gemini, and CoPilot in providing preoperative recommendations for abdominoplasty. They found that Claude and ChatGPT-3.5 demonstrated high reliability, while Microsoft Copilot and Google Gemini exhibited lower reliability [19].

In the field of dentistry, similar inconsistencies in research results are observed. For example, Avşar and Ertan reported that both ChatGPT-3.5 and Google Gemini had limited knowledge regarding prosthetic dental treatments, and both applications yielded similar results [28]. In a study by Guven *et al.* [21] which compared the competencies of three different chatbots (ChatGPT-3.5, ChatGPT-4.0 and Google Gemini) regarding traumatic dental injuries, it was found that ChatGPT-3.5 provided misleading and inaccurate answers, whereas ChatGPT-4.0 and Google Gemini offered more accurate and comprehensive responses. The authors attributed these differences to architectural variations in the chatbots' algorithms and the scope and size of their training data. In contrast, a pilot study by Rokhshad *et al.* [12] compared the responses of nine different chatbots (Google Bard, ChatGPT-4, ChatGPT-3.5, Llama, Sage, Claude 2 100k, Claude-instant, Claude-instant-100k and Google Palm) in the context of pediatric dentistry, concluding that the use of chatbots in clinical pediatric dentistry may not be recommended. Previous studies focused mainly on expert evaluations, but the satisfaction of patients regarding the responses was not explored. This study, however, sought to provide a comprehensive analysis by including both expert evaluations and patient opinions.

When examining studies conducted in the fields of medicine and dentistry, it becomes evident that there is no clear consensus on the reliability of chatbots in healthcare. Different chatbots have demonstrated varying levels of success across different specialties. For instance, ChatGPT provided sufficiently reliable answers in the context of abdominoplasty but was found to be less competent when it came to the topic of dental prostheses [19, 28].

In the long term, chatbots have the potential to reach higher

levels of accuracy as they continue to train and update themselves. This highlights their future capacity to provide more reliable answers in the healthcare field. Another potential solution for improving chatbot reliability and satisfaction could be the development of specialized chatbots designed for specific tasks, similar to how doctors specialize in particular fields of medicine. For example, Pupong and colleagues developed a chatbot named FunDee, which provides oral health care services. They demonstrated that this chatbot achieved high user satisfaction in oral health education [29].

Another important factor in studies evaluating chatbot performance is readability and understandability. The Flesch test was used to determine reading ease, while the Flesch-Kincaid test assessed reading level. The Flesch score ranges from 0 to 100, with higher scores indicating easier readability [30]. Nahir claimed that the readability of GPT responses in pediatric dentistry was low [31]. However, according to the current study, the readability levels of responses provided by chatbots were similar to those of expert opinions. This similarity suggests that the participants' ability to understand the responses was comparable across the different sources. As a result, the participants seemed to be more influenced by the quality, accuracy, and persuasiveness of the content when evaluating the responses. On the other hand, Nahir's study was based on ChatGPT-3.5. We believe that the discrepancy between the results can be attributed to the improved readability and understandability of responses in ChatGPT-4o compared to GPT-3.5. Additionally, AI-supported chatbots have the potential to enhance the clarity of their responses over time as they continue to evolve and improve.

In general, the results indicate that improving the performance and accuracy of chatbots requires optimizing the questions through prompt engineering methods. However, in real-world conditions, users typically ask their questions without any optimization. To better reflect real-world scenarios, no prompt optimization was applied in this study.

This study has several limitations. First, the evaluation was restricted to 15 questions frequently asked by parents, as listed on the IAPD website. While these questions reflect common parental concerns addressed to pediatric dentists, this narrow scope may limit the applicability of the findings to broader or more diverse dental care settings. Additionally, only two chatbot models—ChatGPT-4o and Google Gemini Advanced—were assessed, excluding other current or emerging AI-based chatbot systems that might perform differently under similar conditions. Another limitation lies in the absence of prompt optimization. Although this approach was deliberately chosen to replicate real-life usage, it may have hindered the chatbots from demonstrating their full potential. Future studies should aim to cover a wider range of dental topics, include larger and more diverse participant samples, evaluate various chatbot platforms, and adopt a mixed-methods approach combining both quantitative and qualitative data for a more comprehensive analysis.

5. Conclusions

The concordance between expert opinions and artificial intelligence (AI)-generated responses to questions related to pediatric dentistry and orthodontics remains unpredictable. While pediatric dentists generally find expert opinions to be more satisfactory, the level of satisfaction with AI-generated responses among parents varies depending on the specific question. The fact that the scores for all three response sources are relatively close suggests that AI has promising potential to provide satisfactory answers in the long term. There is a need for more advanced studies that include a larger number of participants, address a broader range of topics, and compare the outputs of multiple AI-based chatbots using a wider set of questions.

AVAILABILITY OF DATA AND MATERIALS

Data and materials are available at the Pediatric Department in the Faculty of Dentistry, Afyonkarahisar Health Sciences University.

AUTHOR CONTRIBUTIONS

İHÇ and FS—conceived the study. İHÇ—designed the research process, collected and organized the data. FS—performed the statistical analysis and interpreted the results. HC—conducted the literature review and contributed to the manuscript editing. All three authors collaboratively wrote the manuscript, contributed to editorial revisions, and approved the final version of the manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The research protocol was reviewed and approved by the Afyonkarahisar Health Sciences University Scientific Research Ethics Committee prior to the initiation of the study (Approval ID: 2025/1). Written and verbal informed consent was obtained from the parents and pediatric dentists.

ACKNOWLEDGMENT

The authors would like to thank the Pediatric Dentistry Department of the Faculty of Dentistry, Afyonkarahisar Health Sciences University, for their support in facilitating data access and clinical coordination throughout the study.

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/2006258298729185280/attachment/Supplementary%20material.docx.

REFERENCES

[1] Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, *et al*. Artificial intelligence for mental health and mental illnesses: an overview. Current Psychiatry Reports. 2019; 21: 1–18.

[2] Kılınç DD, Mansız D. Examination of the reliability and readability of Chatbot Generative Pretrained Transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. American Journal of Orthodontics and Dentofacial Orthopedics. 2024; 165: 546–555.

[3] Wang Q, Erqsous M, Barner KE, Mauriello ML. LATA: A pilot study on LLM-assisted thematic analysis of online social network data generation experiences. Proceedings of the ACM on Human-Computer Interaction. 2025; 9: 1–28.

[4] Fergus S, Botha M, Ostovar M. Evaluating academic answers generated using ChatGPT. Journal of Chemical Education. 2023; 100: 1672–1675.

[5] Giannakopoulos K, Kavadella A, Salim AA, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: comparative mixed methods study. Journal of Medical Internet Research. 2023; 25: e51580.

[6] Daraqel B, Wafaie K, Mohammed H, Cao L, Mheissen S, Liu Y, *et al.* The performance of artificial intelligence models in generating responses to general orthodontic questions: ChatGPT *vs.* Google Bard. American Journal of Orthodontics and Dentofacial Orthopedics. 2024; 165: 652–662.

[7] 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. Applied Sciences. 2025; 15: 893.

[8] Masalkhi M, Ong J, Waisberg E, Lee AG. Google DeepMind's Gemini AI versus ChatGPT: a comparative analysis in ophthalmology. Eye. 2024; 38: 1412–1417.

[9] Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, *et al*. A survey on evaluation of large language models. ACM Trans Intell Syst Technol. 2024; 15: 1–45.

[10] Shukla M, Goyal I, Gupta B, Sharma J. A comparative study of ChatGPT, Gemini, and Perplexity. International Journal of Innovative Research in Computer Science & Technology. 2024; 12: 10–15.

[11] Chang WJ, Chang PC, Chang YH. The gamification and development of a chatbot to promote oral self-care by adopting behavior change wheel for Taiwanese children. Digital Health. 2024; 10: 20552076241256750.

[12] 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.

[13] Jung YS, Chae YK, Kim MS, Lee HS, Choi SC, Nam OH. Evaluating the accuracy of artificial intelligence-based chatbots on pediatric dentistry questions in the Korean National Dental Board Exam. Journal of the Korean Academy of Pediatric Dentistry. 2024; 51: 299–309

[14] Association WM. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA. 2013; 310: 2191–2194.

[15] Madathil KC, Rivera-Rodriguez AJ, Greenstein JS, Gramopadhye AK. Healthcare information on YouTube: a systematic review. Health Informatics Journal. 2015; 21: 173–194.

[16] Oey CG, Livas C. The informative value and design of orthodontic practice websites in The Netherlands. Progress in Orthodontics. 2020; 21: 1–7.

[17] Kılınç DD, Sayar G. Assessment of reliability of YouTube videos on orthodontics. Turkish Journal of Orthodontics. 2019; 32: 145.

[18] Goodman RS, Patrinely JR, Stone CA Jr, Zimmerman E, Donald RR, Chang SS, *et al.* Accuracy and reliability of chatbot responses to physician questions. JAMA Network. 2023; 6: e2336483.

[19] Lim B, Seth I, Cuomo R, Kenney PS, Ross RJ, Sofiadellis F, *et al*. Can AI answer my questions? Utilizing artificial intelligence in the perioperative assessment for abdominoplasty patients. Aesthetic Plastic Surgery. 2024; 48: 4712–4724.

[20] Huo B, Calabrese E, Sylla P, Kumar S, Ignacio RC, Oviedo R, *et al*. The performance of artificial intelligence large language model-linked chatbots in surgical decision-making for gastroesophageal reflux disease. Surgical Endoscopy. 2024; 38: 2320–2330.

[21] 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. 2024; 41: 338–347.

[22] Slavych BK, Atcherson SR, Zraick R. Using ChatGPT to improve health communication and plain language writing for students in communication sciences and disorders. Perspectives of the ASHA Special Interest Groups. 2024; 9: 599–612.

[23] Sismanoglu S, Capan BS. Performance of artificial intelligence on Turkish dental specialization exam: can ChatGPT-4.0 and Gemini advanced achieve comparable results to humans? BMC Medical Education. 2025; 25: 1–10.

[24] Omar M, Nassar S, Hijazi K, Glicksberg BS, Nadkarni GN, Klang E. Generating credible referenced medical research: a comparative study of OpenAI's GPT-4 and Google's Gemini. Computers in Biology and Medicine. 2025; 185: 109545.

[25] Kalluri K, Kokala A. Performance benchmarking of generative AI models: Chatgpt-4 *vs.* Google Gemini AI. International Research Journal of Modernization in Engineering Technology and Science. 2024; 6: 4673–4677.

[26] Schmidt J, Lichy I, Kurz T, Peters R, Hofbauer S, Plage H, *et al*. ChatGPT as a support tool for informed consent and preoperative patient education prior to penile prosthesis implantation. Journal of Clinical Medicine. 2024; 13: 7482.

[27] Reyhan AH, Mutaf Ç, Uzun İ, Yüksekyayla F. A performance evaluation of large language models in keratoconus: a comparative study of ChatGPT-3.5, ChatGPT-4.0, Gemini, Copilot, Chatsonic, and Perplexity. Journal of Clinical Medicine. 2024; 13: 6512.

[28] Bilgin Avsar D, Ertan AA. Comparative evaluation of ChatGPT-3.5 and Gemini in answering prosthodontics questions from the dental specialty exam: a cross-sectional study. Turkiye Klinikleri Journal of Dental Sciences. 2024; 30:668–673.

[29] Pupong K, Hunsrisakhun J, Pithpornchaiyakul S, Naorungroj S. Development of chatbot-based oral health care for young children and evaluation of its effectiveness, usability, and acceptability: mixed methods study. JMIR Pediatrics and Parenting. 2025; 8: e62738.

[30] Flesch R. A new readability yardstick. Journal of Applied Psychology. 1948; 32: 221–233.

[31] Bayraktar Nahir C. Can ChatGPT be a guide in pediatric dentistry? BMC Oral Health. 2025; 25: 1–8.

How to cite this article: İsmail Haktan Çelik, Hasan Camcı, Farhad Salmanpour. Bridging the information gap in pediatric dentistry: a comparison of ChatGPT-4o, Google Gemini advanced, and expert responses based on evaluations by parents and pediatric dentists. Journal of Clinical Pediatric Dentistry. 2026; 50(1): 147-155. doi: 10.22514/jocpd.2026.014.