BMC Oral Health

RESEARCH

Open Access

Evaluation of ChatGPT-4's performance on pediatric dentistry questions: accuracy and completeness analysis

Berkant Sezer[1*] and Alev Eda Okutan[2]

Abstract

Background This study aimed to evaluate the accuracy and completeness of Chat Generative Pre-trained Transformer-4 (ChatGPT-4) responses to frequently asked questions (FAQs) posed by patients and parents, as well as curricular questions related to pediatric dentistry. Additionally, it sought to determine whether the ChatGPT-4's performance varied across different question topics.

Methods Responses from ChatGPT-4 to 30 FAQs by patients and parents and 30 curricular questions covering six pediatric dentistry topics (fissure sealants, fluoride, early childhood caries, oral hygiene practices, development of dentition and occlusion, and pulpal therapy) were evaluated by 30 pediatric dentists. Accuracy was rated using a five-point Likert scale, while completeness was assessed via a three-point scale, capturing distinct aspects of response quality. Statistical analyses included Fisher's Exact test, Mann–Whitney U test, Kruskal–Wallis test, and Bonferroni-adjusted post hoc comparisons.

Results ChatGPT-4's responses demonstrated high overall accuracy across all question types. Mean accuracy scores were 4.21 ± 0.55 for FAQs and 4.16 ± 0.70 for curricular questions, indicating that responses were generally rated as "good" to "excellent" by pediatric dentists, with no statistically significant difference between the two groups ($p = 0.942$). Completeness scores were moderate overall, with means of 2.51 ± 0.40 (median: 3) and 2.61 ± 1.53 (median: 3) for FAQs and curricular questions, respectively ($p = 0.563$), reflecting a generally acceptable response coverage. Accuracy scores for curricular questions varied significantly by topic ($p = 0.007$), with the highest score for fissure sealants (4.45 ± 0.62; median: 5) and the lowest for pulpal therapy (3.93 ± 0.93; median: 4).

Conclusion From a clinical perspective, ChatGPT-4 demonstrates promising accuracy and acceptable completeness in pediatric dental communication. However, its performance in certain curricular areas—particularly fluoride and pulpal therapy—warrants cautious interpretation and requires professional oversight.

Keywords Artificial intelligence, Generative artificial intelligence, Large language models, Chatbot, Fissure sealant, Fluoride, Dental pulp, Dental caries, Oral hygiene, Dental occlusion

*Correspondence:
Berkant Sezer
dt.berkantsezer@gmail.com; berkant.sezer@comu.edu.tr
[1]Department of Pediatric Dentistry, School of Dentistry, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye
[2]Private Practice in Pediatric Dentistry, Istanbul, Türkiye

Background

Artificial intelligence-based Large Language Models (LLMs) are advanced systems that utilize deep learning algorithms trained on billions of data points. They learn language structure from vast corpora of text—such as websites, books, and articles—and generate human-like responses [1]. In recent years, LLMs have rapidly gained importance due to their enhanced language understanding and generation capabilities, enabling novel applications in healthcare, education, and scientific communication, largely driven by major advances in Natural Language Processing (NLP) [1]. The first Generative Pre-trained Transformer (GPT) model, based on LLM technology and released by OpenAI in 2018, marked a major milestone in artificial intelligence. It was soon followed by more advanced versions that achieved widespread adoption due to their improved language capabilities [2, 3]. Today, LLMs are employed across diverse fields such as medical consulting, law, academic writing, and customer service, and have even become integrated into daily life as personal assistants and chatbots [4].

The development of LLMs began prior to the COVID-19 pandemic; however, the restrictions imposed during the pandemic significantly accelerated the adoption of artificial intelligence technologies in healthcare [5]. This rapid shift towards digital tools highlighted the growing demand for accessible and responsive communication methods among patients, parents, and healthcare providers. The advancement of GPT models has been driven by the availability of larger datasets and more powerful computing resources. Additionally, widespread usage has rapidly enhanced these models' ability to understand and generate human-like language [6, 7].

The use of LLMs in healthcare is rapidly expanding, particularly in applications such as clinical decision support, patient education, medical documentation, and accelerating research processes. LLMs possess the capability to process vast datasets, comprehend patient information, and provide personalized treatment recommendations [8]. When integrated with comprehensive tools, GPT models can assist healthcare professionals interpreting medical images, case-related information, and symptoms, thereby increasing diagnostic accuracy and reducing interpretation time [9].

Dentistry is an important sub-discipline in which NLP and LLMs can be effectively applied. These models are increasingly used in ths field to improve both administrative efficiency and patient communication, as well as to support clinical decision-making and professional education [10, 11]. Moreover, LLMs are preferred tools for efficiently accessing accurate technical information to aid dentists' continuing education, and for academic purposes such as course material development and examination preparation [12]. The growing integration of LLMs into healthcare communication has sparked rising interest in their role within dentistry. Recent studies have evaluated their performance in responding to clinically relevant questions across various dental specialties, including prosthodontics, endodontics, orthodontics, and evidence-based dentistry [13–17]. While these tools offer promise for streamlining patient communication and supporting education, multiple evaluations have highlighted that their accuracy and consistency remain suboptimal. These findings emphasize the need for cautious implementation and further validation before LLMs can be routinely integrated into dental practice [15].

In pediatric dentistry, recent studies have explored the applicability of LLMs in various contexts. These investigations have assessed the use of LLMs for addressing frequently asked questions (FAQs) from parents and patients [18, 19], supporting evidence-based decision-making and the dissemination of technical knowledge in clinical settings [20, 21], and providing structured responses for behavior guidance and trauma management scenarios [22, 23]. More recent research has evaluated the utility of LLMs in addressing parental concerns related to pediatric dental trauma [24], as well as their effectiveness as patient information tools in orthodontic contexts [25, 26]. This expanding body of literature highlights the need for comprehensive and structured evaluations that examine both the capabilities and limitations of LLMs across the diverse content areas within pediatric dentistry.

The aim of this study was to comprehensively evaluate the performance of ChatGPT-4 in pediatric dentistry by assessing the accuracy and completeness of its responses to both FAQs posed by patients and parents, as well as curricular clinical questions formulated by experts. Unlike prior studies that addressed patient education or academic content in isolation, our research uniquely combines both dimensions, spanning six core domains of pediatric dentistry. This dual-scope design, paired with structured expert scoring, fills a critical gap in the literature by reflecting real-world informational needs from both lay and professional perspectives. By offering a topic-specific, context-aware evaluation of ChatGPT-4, this study provides novel insights into its potential and limitations, thereby contributing meaningfully to the growing body of research on artificial intelligence in pediatric dentistry.

Methods

This study employed a cross-sectional descriptive design to evaluate the accuracy and completeness of ChatGPT-4's responses to selected pediatric dentistry questions. Although no specific reporting guidelines currently exist for chatbot performance in dentistry, the methodology was informed by the STROBE guidelines and general

principles from the EQUATOR Network. Ethical review and approval were waived as the study did not involve human or animal subjects [25–29]. The objective was to assess ChatGPT-4's answers to FAQs posed by patients and parents, as well as curricular questions commonly used in academic pediatric dentistry. To obtain a balanced set of questions from both clinical and educational domains, two seperate groups of licensed pediatric dentists participated. One group, consisting of clinical pediatric dentists, submitted patient-/parent-facing FAQs, while the other group—comprising university-based pediatric dentistry academicians—provided curricular questions frequently used in educational settings.

Questions development and content

Sourcing of frequently asked questions (FAQs)

Firstly, the study's lead researchers (B.S. and A.E.O.) independently searched Google, Microsoft Bing, Yahoo, and Yandex between April 12, 2024, and April 15, 2024. They used using English-language search phrases such as "frequently asked questions," "pediatric dentistry," and "pedodontics." For each search engine, they examined websites on the first ten pages in detail and compiled a list of FAQs related to pediatric dentistry found on these pages. This preliminary step aimed to create a foundational pool of real-world FAQs content that reflects the types of questions parents and patients commonly encounter online. It served as a reference to validate and guide the collection of clinician-submitted questions.

The search included a mix of academic and non-academic sources. Frequently accessed websites such as the American Academy of Pediatric Dentistry (AAPD), European Academy of Paediatric Dentistry (EAPD), and International Association of Paediatric Dentistry (IAPD), as well as public-facing clinic websites, were reviewed. A total of 86 unique FAQs were compiled through this process. Only patient- and parent-facing questions were collected in this stage; no curricular questions were included in the web-derived list.

The researchers then jointly reviewed the lists and created a single consensus list of questions. All web searches, communications, and survey materials—including question prompts, instructions, and evaluation forms—were prepared and administered in English. In Türkiye, English proficiency is a formal requirement for dental specialists during their specialty or doctoral training. This ensured that all participating pediatric could understand and respond to the English content used throughout the study.

Subsequently, ten pediatric dentists were randomly selected from the official member registry of the Turkish Society of Pediatric Dentistry. To enhance geographical diversity, each participant was from a different city in Türkiye. These clinical pediatric dentists were contacted via email and informed about the study. Each of the ten participants was asked to submit 30 FAQs commonly received from patients and parents during routine clinical encounters. The Collection of these questions took place between April 20 and April 30, 2024, with all responses obtained electronically. This step aimed to capture real-world patient and parent concerns across pediatric dental settings.

Collection of curricular questions

Ten pediatric dentistry academicians were randomly selected from different universities across Türkiye to contribute to the development of curricular questions. This selection ensured inclusion of widely accepted educational and examination content used in both undergraduate and postgraduate programs. Each academician was affiliated with a dental school and registered in the Council of Higher Education (YÖK) database as a faculty member in a pediatric dentistry department. To achieve regional diversity, efforts were made to include academicians from different cities. These individuals were contacted via email, informed about the study, and asked to submit 30 curricular questions they most frequently use when preparing educational content or examinations for undergraduate and postgraduate pediatric dentistry courses. Responses were collected electronically between May 1 and May 11, 2024.

Question categorization and selection criteria

The responses from both groups of pediatric dentists, along with the outputs obtained from the web search, were thoroughly evaluated by the two lead authors of this study (B.S. and A.E.O.), both pediatric dentists. The final FAQs list used in the study did not directly include web-derived questions. Instead, the web search served as a supplementary reference to validate and cross-check the relevance of questions submitted by pediatric dentists based on their clinical experience with patients or parents.

Through inductive content analysis, six thematic categories were identified based on recurring topics: fissure sealants, fluoride, early childhood caries, oral hygiene practices, development of dentition and occlusion, and pulpal therapy. The questions were grouped accordingly, maintaining a distinction between FAQs by patients and parents and curricular questions.

Within each topic, questions were ranked by frequency of submission. The five most frequently submitted questions in each group (FAQs vs. curricular) were selected to ensure topical balance and representativeness. This approach guaranteed that the final question set reflected those most commonly encountered or prioritized by clinicians and educators.

In this study, "FAQs by patients and parents" refers to questions commonly encountered during routine pediatric dental appointments or listed on public-facing websites targeting non-professional audiences. Although these questions were not collected directly from parents or patients, they were submitted by experienced pediatric dentists based on their daily interactions and the FAQs commonly asked by parents during routine visits. This indirect method was adopted to realistically capture parental concerns across diverse clinical settings. In contrast, "curricular questions" denote items typically used by academic pediatric dentists when preparing educational content or assessments. The categorized questions used in the study are presented in Table 1.

Answers collection

All responses were obtained by submitting the selected questions to ChatGPT-4 Omni (ChatGPT-4o; OpenAI, San Francisco, CA, USA), which was released on May 13, 2024, and was the most recent publicly available version at the time of data collection (June 2024). All prompts were submitted through the same registered ChatGPT Plus account, using a consistent browser and device. To ensure a clean and unbiased starting point, all search history and cookies were cleared once at the beginning of the session. However, cookies and history were not cleared between prompts to preserve session consistency and enhance reproducibility. While this approach aimed to minimize variability, the precise influence of cookies on ChatGPT response variability remains unclear, and definitive scientific evidence on this matter is currently lacking.

To reduce potential bias introduced by search engine memory and personalization algorithms, a new Google account was created to access ChatGPT-4 via the web interface. The "Continue with Google" option was used to log in, replicating a typical user experience. Before entering the questions, all search history and cookies were cleared using the "Clear Browsing Data" option found under the "Privacy and Security" settings of the browser. Since data collection was performed via the web-based interface of ChatGPT Plus, no user-accessible controls for system-level parameters such as temperature, system prompts, or API settings were available. Therefore, default platform settings were used throughout the study. Each question was manually entered on June 8, 2024, in Istanbul, Türkiye. For consistency, a new chat window was opened for each question to ensure that previous responses did not influence subsequent answers. All responses were provided in English, recorded in real time, and saved along with their corresponding questions and timestamps. A sample screenshot of a ChatGPT-4 response is included as Fig. 1, illustrating the response format.

Answers evaluation

A four-section online questionnaire was created using Google Forms on June 1, 2024. Detailed information about the study was provided in the first section of the form. The participants were licensed pediatric dentists randomly selected from the official member registry of the Turkish Society of Pediatric Dentistry. To avoid evaluator bias, the 20 pediatric dentists who participated in the question development phase were excluded from the sampling pool. Although the evaluators were not the same individuals who developed the curricular questions, many had university affiliations or teaching responsibilities, ensuring familiarity with academic content without compromising scoring objectivity.

Subsequently, 60 individuals were selected from the remaining eligible members using simple random sampling and contacted via email. The invitation included detailed information about the study. The first section of the form collected demographic information such as gender, age, years of clinical experience, and prior use of artificial intelligence for patient information or technical purposes. These variables provided descriptive insight into the evaluator sample but were excluded from primary statistical analyses, as they were not relevant to evaluating model performance independent of evaluator characteristics.

The second section requested the evaluation of ChatGPT-4's responses to 30 FAQs by patients and parents, while the third section requested evaluation of answers to 30 curricular questions. The final section contained two questions: "After these answers, would you recommend the use of artificial intelligence in pediatric dentistry to your patients?" and "After these answers, would you recommend/use artificial intelligence in the field of pediatric dentistry education?"

Of the 60 invited pediatric dentists, 30 agreed to participate and were included. This evaluator group consisted of both clinically practicing and university-based pediatric dentists. To enhance scoring consistency, each evaluator was sent the same set of questions and ChatGPT-4 responses in randomized order two weeks after the initial assessment. If discrepancies appeared between the first and second scoring, those responses were sent again for a third and final evaluation. The third-round score was recorded as final. This three-step procedure was implemented to improve intra-rater consistency across the evaluators. The entire scoring process occurred between July 1 and September 30, 2024, concurrently with the primary data collection.

In the second and third sections of the questionnaire, participants evaluated the ChatGPT-4's responses using two criteria: (i) a five-point Likert scale for assessing "accuracy," and (ii) a three-point Likert scale for assessing "completeness." Definitions for each point on the Likert

Table 1 Topics and questions used to evaluate the accuracy and completeness of ChatGPT-4's responses in pediatric dentistry

Frequently Asked Questions by Patients and Parents	Curricular Questions
Topic 1. Fissure sealants	
Q1. What are fissure sealants?	Q1. What are the types of fissure sealants?
Q2. How does the dentist apply fissure sealant to the teeth?	Q2. What is the content of resin fissure sealants?
Q3. Are there any alternatives to using a fissure sealant in teeth?	Q3. What are the indications of glass ionomer fissure sealants?
Q4. At what ages should fissure sealants be applied?	Q4. What are the reasons for failure in fissure sealants?
Q5. Are there any harmful ingredients in fissure sealants?	Q5. To what extent do fissure sealants reduce the risk of dental caries formation?
Topic 2. Fluoride	
Q1. What is fluoride?	Q1. What should be the optimal amount of fluoride in drinking water?
Q2. Should fluoride toothpaste be used?	Q2. What are the treatment options for dental fluorosis?
Q3. Does fluoride inhibit growth and development?	Q3. How to apply fluoride gel?
Q4. What amount of fluoride should be in toothpaste for which age group?	Q4. What is acute fluoride toxicity?
Q5. What is dental fluorosis?	Q5. What is chronic fluoride toxicity?
Topic 3. Early childhood caries	
Q1. What is early childhood caries?	Q1. What are the microorganisms responsible for early childhood caries?
Q2. What should nutrition be like to prevent early childhood caries?	Q2. What is the window of infectivity associated with early childhood caries?
Q3. How should oral care be done to prevent early childhood caries?	Q3. What are the treatment options for early childhood caries?
Q4. Does bottle feeding pose a risk for early childhood caries?	Q4. Is it possible for microorganisms that cause dental caries to be transmitted from mother to child?
Q5. Is there a genetic predisposition to early childhood caries?	Q5. What are demineralization and remineralization?
Topic 4. Oral hygiene practices	
Q1. What age should children start brushing their teeth?	Q1. Which is the most effective tooth brushing technique for children?
Q2. What type of toothpaste should children use?	Q2. What are the characteristics of healthy gums in the primary dentition?
Q3. Which type of toothbrush should children use?	Q3. What are the causes of gingival bleeding in adolescents?
Q4. At what age can a child be left to brush their own teeth without supervision?	Q4. What are the beneficial foods for maintaining oral health?
Q5. In what cases is bleeding gums observed in children?	Q5. What are Nasymth membrane stains?
Topic 5. Development of dentition and occlusion	
Q1. Is it normal to have a gap between primary teeth?	Q1. What is the ugly-duckling period?
Q2. What are the causes of teeth grinding in children?	Q2. What are the treatment methods for long-term thumb sucking?
Q3. What do space maintainers do?	Q3. What is Leeway space?
Q4. Until what age should thumb sucking be stopped in children?	Q4. What are the types of primary dentition occlusal relationships?
Q5. Is it a problem for the permanent tooth to erupt without exfoliating the related primary tooth?	Q5. What factors are involved in the etiology of bruxism in children?
Topic 6. Pulpal therapy	
Q1. Is root canal treatment performed on primary teeth?	Q1. What are the obturation materials and techniques for primary tooth root canal treatment?
Q2. In what cases is root canal treatment required for primary teeth?	Q2. What are the materials used in primary tooth pulpotomy?
Q3. What materials are used in primary tooth root canal treatment?	Q3. What are the indications for regenerative endodontic treatment?
Q4. Does root canal treatment of a primary tooth damage the permanent tooth that will replace it?	Q4. What is lesion sterilization tissue repair?
Q5. Is primary tooth root canal treatment costly?	Q5. In what cases is Cvek pulpotomy applied?

scales are provided in Table 2. These scales were adapted from the Global Quality Scale and prior studies [18, 28, 30, 31], with minor modifications to suit the pediatric dentistry context and the specific nature of the questions evaluated. An overview of the entire methodological workflow—including question development, response generation, expert evaluation, and analysis—is illustrated in Fig. 2.

Statistical analyses

Data were initially recorded in Microsoft Excel (version 16; Microsoft Corp., WA, USA) and analyzed using

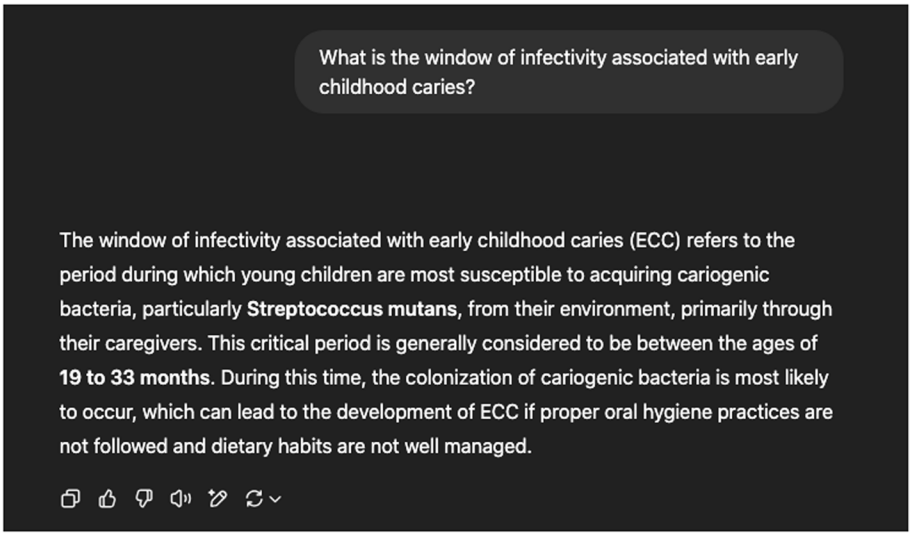

Fig. 1 Sample response generated by ChatGPT-4 to one of the evaluated questions included in the study

Table 2 Definitions of the likert scales used for accuracy and completeness evaluation

Score	Accuracy	Completeness
1	Poor accuracy, poor flow of information, with most information missing; not useful for patient information or curricular purposes.	Incomplete response addressing only some aspects of the question; important parts missing.
2	Generally poor accuracy; some information is presented, but many important issues are missing; very limited utility for patient or curricular purposes.	Adequate response that covers all aspects of the question, but only provides the minimum required information.
3	Moderate accuracy with inadequate flow; some important information is adequate, but other parts are insufficient; somewhat useful for patient or curricular.	Comprehensive answer that addresses all aspects of the question and offers additional information or context beyond expectations.
4	Good accuracy and flow in general; most relevant information is sufficient and useful for patient or curricular applications.	–
5	Excellent accuracy and flow; very useful for both patient information and curricular purposes.	–

IBM SPSS Statistics, version 26 (IBM Corp., Chicago, IL, USA). Continuous variables (participants' age and years of clinical experience) were presented as means and standard deviations, as well as medians, minimums, and maximums; categorical variables were summarized as counts and percentages. Among these, age and years of professional experience were found to follow a normal distribution. For these variables, Levene's test was used to assess homogeneity of variances, and descriptive statistics were applied accordingly. For the main outcome variables—accuracy (five-point Likert scale) and completeness (three-point Likert scale)—the data did not follow a normal distribution; therefore, non-parametric tests were employed. The Mann-Whitney U test was used to compare evaluation scores between FAQs by patients and parents and curricular questions. The Kruskal-Wallis test was used to compare accuracy and completeness scores across the six pediatric dentistry topics. When statistically significant differences were found, Bonferroni-adjusted post hoc comparisons were conducted. Associations between categorical variables were analyzed using Fisher's Exact test. A two-tailed p-value < 0.05 was considered statistically significant for all analyses. Effect sizes (r and η^2) were calculated for the main statistical analyses to provide a more comprehensive understanding of the magnitude and practical relevance of observed differences.

Results

A total of 60 questions were included in the study, comprising five FAQs posed by patients and parents, as well as five curricular questions across six pediatric dentistry topics. All questions and ChatGPT-4's answers to these questions are presented in the Supplementary Material.

A total of 30 pediatric dentists participated in the study, consisting of 19 females and 11 males. The distribution of participants by gender, age, years of professional experience, and experience with artificial intelligence for patient information or curricular questions are shown in Table 3.

Figure 3 illustrates the distribution of scores assigned by pediatric dentists to ChatGPT-4's responses to FAQs by patients and parents, while Fig. 4 presents the distribution of scores given to ChatGPT-4's curricular questions

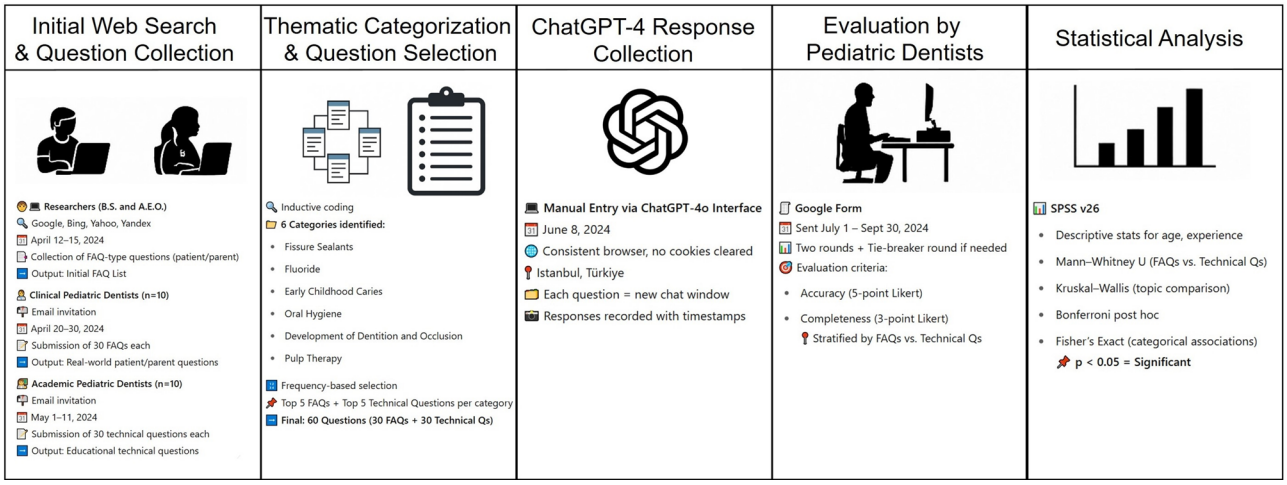

Fig. 2 Flowchart illustrating the complete study design, including the processes of question development (from both patients/parents FAQs and curricular questions), response generation by ChatGPT-4, and expert evaluation methodology used for accuracy and completeness assessment

Table 3 Demographic characteristics of respondents pediatric dentists

N (%)		Age		Professional experience year		Artificial intelligence experience for patient information			Artificial intelligence experience for curricular questions		
		Mean ± SD	*p*-value*	Mean ± SD	*p*-value*	Yes *N* (%)	No *N* (%)	*p*-value†	Yes *N* (%)	No *N* (%)	*p*-value†
Female	19 (63.3)	34.58 ± 6.34	0.612	8.11 ± 6.46	0.840	3 (15.8)	16 (84.2)	1.000	4 (21.1)	15 (78.9)	0.108
Male	11 (36.7)	33.45 ± 4.66		8.55 ± 4.03		1 (9.1)	10 (90.9)		6 (54.5)	5 (45.5)	

N Number, *SD* Standard Deviation

*Independent samples t-test

†Fisher exact test

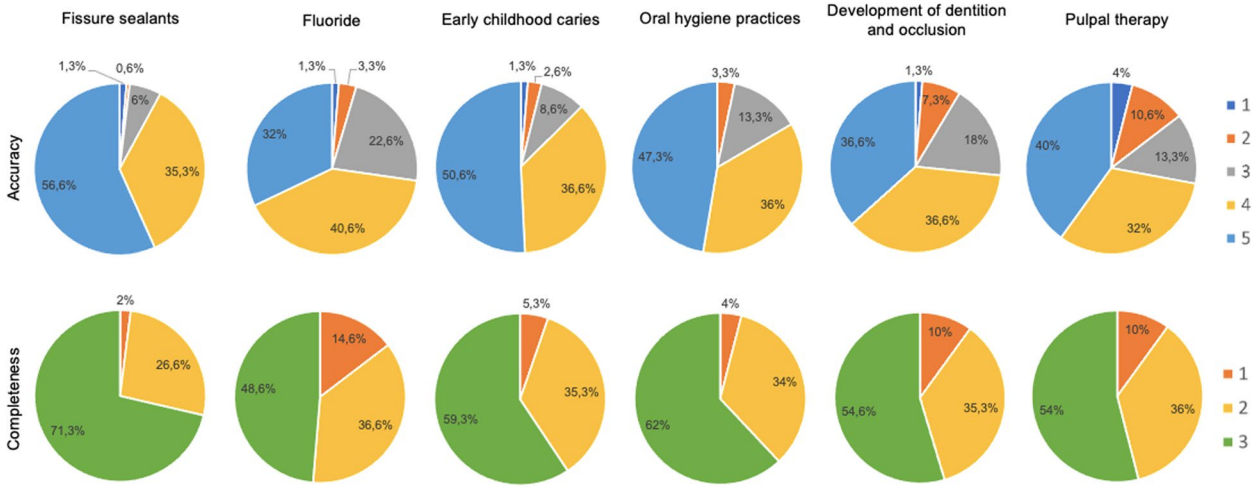

Fig. 3 Distribution of accuracy and completeness scores assigned by pediatric dentists to ChatGPT-4's responses to frequently asked questions posed by patients and parents. Upper row charts show accuracy score distribution on a five-point Likert scale; lower row charts show completeness score distribution on a three-point Likert scale

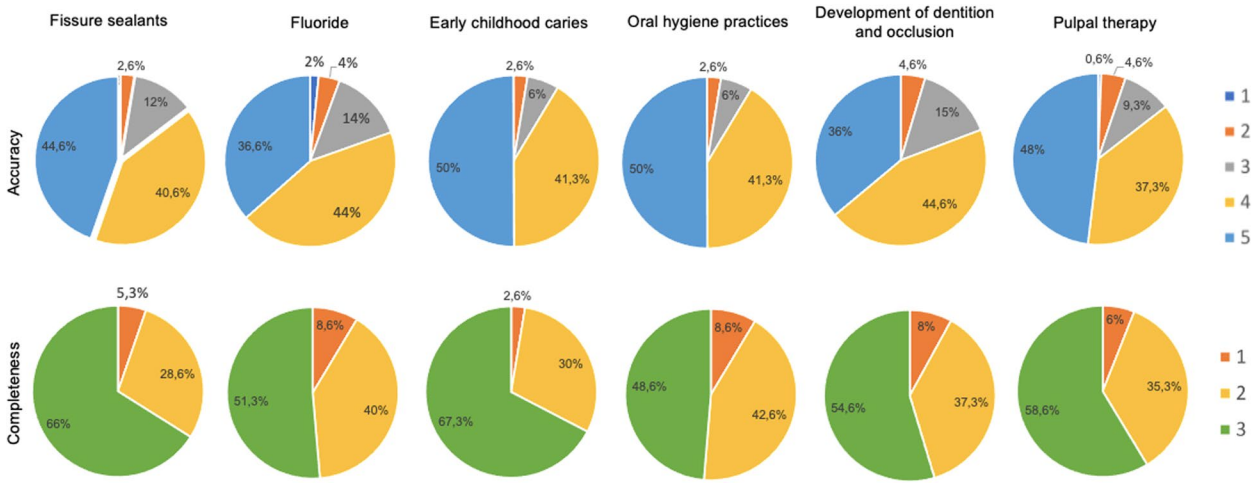

Fig. 4 Distribution of accuracy and completeness scores assigned by pediatric dentists to ChatGPT-4's responses to expert-generated curricular questions across six pediatric dentistry topics. Upper row charts show accuracy score distribution on a five-point Likert scale; lower row charts show completeness score distribution on a three-point Likert scale

Table 4 Analysis of accuracy scores of frequently asked questions by patients and parents and curricular questions on different topics

Topics	Question Types	Number of responses	Score Mean ± SD	Median (min.-max.)	*p*- value*	*r* (effect-size)
Fissure sealants	FAQs by patients and parents	150	4.27 ± 0.51	4 (2–5)	0.081	0.23
	Curricular questions	150	4.45 ± 0.62	5 (1–5)		
Fluoride	FAQs by patients and parents	150	4.11 ± 0.51	4 (1–5)	0.356	0.12
	Curricular questions	150	3.99 ± 0.56	4 (1–5)		
Early childhood caries	FAQs by patients and parents	150	4.39 ± 0.55	4.5 (2–5)	0.692	0.05
	Curricular questions	150	4.33 ± 0.62	5 (1–5)		
Oral hygiene practices	FAQs by patients and parents	150	4.11 ± 0.53	4 (1–5)	0.171	0.18
	Curricular questions	150	4.27 ± 0.63	4 (2–5)		
Development of dentition and occlusion	FAQs by patients and parents	150	4.12 ± 0.55	4 (2–5)	0.503	0.09
	Curricular questions	150	4 ± 0.68	4 (1–5)		
Pulpal Therapy	FAQs by patients and parents	150	4.27 ± 0.63	4 (1–5)	0.242	0.15
	Curricular questions	150	3.93 ± 0.93	4 (1–5)		

SD Standard Deviation, *FAQs* Frequently Asked Questions, *min* minimum, *max* maximum

*Mann Whitney U test

in the form of a pie charts. The graphs in the upper rows depict the distribution of accuracy scores, whereas the graphs in the lower rows display the distribution of completeness scores.

There were no statistically significant differences between the mean accuracy scores assigned by participants for ChatGPT-4's responses to FAQs by patients and parents and to curricular questions across different topics ($p > 0.05$) (Table 4). Except for the accuracy scores given to curricular questions on fluoride and pulpal therapy, the mean accuracy score of all other responses was above 4. In addition to the reported *p*-values, effect sizes (*r*) were calculated for all Mann-Whitney U tests assessing accuracy, ranging from 0.05 to 0.23. These values indicate small to small-to-moderate effects, with the highest effect size observed for fissure sealants. This additional analysis provides a more nuanced understanding of the comparative performance between FAQs and curricular questions.

Similarly, no statistically significant differences were found between the mean completeness scores assigned by participants for the responses provided to FAQs by patients and parents and curricular questions across various topics ($p > 0.05$) (Table 5). The mean completeness score of all responses in all topics exceeded 2. Effect sizes (*r*) for the Mann-Whitney U tests related to completeness scores ranged from 0.02 to 0.10, indicating small effects across all pediatric dentistry topics. This additional analysis provides a more nuanced understanding of the lack of significant differences in completeness between FAQs and curricular questions.

When all questions were evaluated, irrespective of the topic within the scope of FAQs by patients and parents and curricular questions, there were no statistically

Table 5 Analysis of completeness scores of frequently asked questions by patients and parents and curricular questions on different topics

Topics	Question Types	Number of responses	Score Mean ± SD	Median (min.-max.)	*p*-value*	*r* (effect size)
Fissure sealants	FAQs by patients and parents	150	2.61 ± 0.32	3 (1–3)	0.109	0.09
	Curricular questions	150	2.69 ± 0.42	3 (1–3)		
Fluoride	FAQs by patients and parents	150	2.43 ± 0.41	3 (1–3)	0.381	0.05
	Curricular questions	150	2.34 ± 0.52	2 (1–3)		
Early childhood caries	FAQs by patients and parents	150	2.65 ± 0.37	3 (1–3)	0.270	0.06
	Curricular questions	150	2.54 ± 0.39	3 (1–3)		
Oral hygiene practices	FAQs by patients and parents	150	2.4 ± 0.42	2 (1–3)	0.084	0.10
	Curricular questions	150	2.58 ± 0.45	3 (1–3)		
Development of dentition and occlusion	FAQs by patients and parents	150	2.47 ± 0.41	3 (1–3)	0.782	0.02
	Curricular questions	150	2.45 ± 0.47	3 (1–3)		
Pulpal Therapy	FAQs by patients and parents	150	2.53 ± 0.42	3 (1–3)	0.686	0.02
	Curricular questions	150	2.44 ± 0.52	3 (1–3)		

SD Standard Deviation, *FAQs* Frequently Asked Questions, *min* minimum, *max* maximum

*Mann Whitney U test

Table 6 Analysis of accuracy and completeness scores for total frequently asked questions by patients and parents and curricular questions

Evaluation type	Question type	Number of responses	Score Mean±SD	Median (min.-max.)	*p*-value*	*r* (effect size)
Accuracy	FAQs by patients and parents	900	4.21±0.55	4 (1-5)	0.942	0.01
	Curricular questions	900	4.16±0.70	4 (1-5)		
Completeness	FAQs by patients and parents	900	2.51±0.40	3 (1-3)	0.563	0.21
	Curricular questions	900	2.61±1.53	3 (1-3)		

SD Standard Deviation, *FAQs* Frequently Asked Questions

*Mann Whitney U test

significant differences between the groups regarding accuracy and completeness ($p > 0.05$) (Table 6). Effect sizes (r) were also calculated for these comparisons, yielding values of approximately 0.01 for accuracy and 0.21 for completeness, indicating negligible to small-to-moderate effects consistent with the nonsignificant p-values.

Table 7 presents comparisons of the mean scores of ChatGPT-4 responses evaluated by pediatric dentists for both accuracy and completeness across different topics. A statistically significant difference was identified between the scores of ChatGPT-4 responses to curricular questions ($p = 0.007$). The highest score in curricular topics was for fissure sealant (4.45 ± 0.62), while the lowest score was for pulpal therapy (3.93 ± 0.93). According to post hoc Bonferroni correction, the mean accuracy score of responses to curricular questions on fissure sealants was statistically significantly higher than those for fluoride ($p = 0.001$), development of dentition and occlusion ($p = 0.004$), and pulpal therapy ($p = 0.013$). Additionally, accuracy scores for early childhood caries were significantly higher than those for fluoride ($p = 0.018$). There were no statistically significant differences among the other groups ($p > 0.05$). Effect sizes were calculated using eta squared (η^2) to better quantify the magnitude of differences observed. For curricular questions, the significant differences in accuracy scores corresponded to a moderate effect size ($\eta^2 \approx 0.25$), indicating a meaningful distinction across topics. In contrast, FAQs accuracy differences showed a negligible effect size ($\eta^2 \approx 0.09$), consistent with the lack of statistical significance. Completeness scores for both groups exhibited very small effect sizes ($\eta^2 \approx 0.03$ and 0.04), underscoring minimal practical differences. These findings highlight that while some accuracy differences in curricular topics are of moderate practical relevance, other comparisons reflect limited impact.

Regarding the final section of the questionnaire, 26 participants (86.7%) answered "yes," 2 (6.7%) "undecided," and 2 (6.7%) "no" to the question, "After these answers, would you recommend the use of artificial intelligence in pediatric dentistry to your patients?" For the question, "After these answers, would you recommend/use artificial intelligence in pediatric dentistry education?" 21 participants (70%) answered "yes," 9 (30%) were "undecided," and none answered "no."

Table 7 Analysis of the difference between the pediatric dentistry related topics in accuracy and completeness scores in frequently asked questions by patients and parents and curricular questions

	Accuracy						Completeness					
	FAQs by patiens and parents			Curricular questions			FAQs by patiens and parents			Curricular questions		
Topics	Score Mean ± SD	*p*-value*	η^2	Score Mean ± SD	*p*-value†	η^2	Score Mean ± SD	*p*-value*	η^2	Score Mean ± SD	*p*-value*	η^2
Fissure sealants	4.27 ± 0.51	0.154	0.09	4.45 ± 0.62^{A}	**0.007**	0.25	2.61 ± 0.32	0.124	0.03	2.69 ± 0.42	0.164	0.04
Fluoride	4.11 ± 0.51			3.99 ± 0.56^{C}			2.43 ± 0.41			2.34 ± 0.52		
Early childhood caries	4.39 ± 0.55			4.33 ± 0.62^{AB}			2.65 ± 0.37			2.54 ± 0.39		
Oral hygiene practices	4.11 ± 0.53			4.27 ± 0.63^{ABC}			2.4 ± 0.42			2.58 ± 0.45		
Development of dentition and occlusion	4.12 ± 0.55			4 ± 0.68^{BC}			2.47 ± 0.41			2.45 ± 0.47		
Pulpal therapy	4.27 ± 0.63			3.93 ± 0.93^{BC}			2.53 ± 0.42			2.44 ± 0.52		

SD Standard Deviation, *FAQs* Frequently Asked Questions. *Kruskal Wallis test. †Kruskal Wallis test with post hoc adjusted Bonferroni test. η^2 means effect size for Kruskal Wallis test. Bold *p*-value means statistically significant. Different superscript capital letters indicate significant difference within relevant column ($p<0.05$)

Discussion

The results of this study showed that ChatGPT-4's responses received an average accuracy score of around 4, reflecting good accuracy and logical coherence, with most answers providing sufficient and relevant information for both patient-facing FAQs and curricular questions. The mean completeness score, exceeding 2, further indicates that the majority of responses adequately covered the essentials aspects of the questions. Although no statistically significant differences were found between FAQs and curricular questions in overall scores, a trend of higher performance on FAQs was observed. Moreover, significant variation emerged across pediatric dentistry topics, with notably lower accuracy scores in fluoride and pulpal therapy questions compared to fissure sealants. These findings underscore the influence of content type and topic complexity on the performance of LLMs, highlighting the need for further investigation into their limitations in addressing nuanced clinical scenarios. These results were supported by effect size analyses, which provided additional insight into the practical significance of observed differences, strengthening the reliability of these conclusions.

Several studies have evaluated ChatGPT's ability to answer patient-centered questions, demonstrating its capacity to provide information within the field of dentistry. For example, a study assessing patient-centered questions related to interceptive orthodontics reported high levels of accuracy and completeness in the artificial intelligence's responses [32]. Both ChatGPT-3.5 and ChatGPT-4 were evaluated for answering FAQs about fixed orthodontic treatment, with both models showing good quality but low to moderate reliability and challenging readability [25]. More recently, pediatric dentistry-specific studies have emerged. Gökcek Taraç and Nale [24] assessed the accuracy and relevance of chatbot responses to parental questions following pediatric dental trauma, concluding that while some limitations existed, chatbots could provide helpful guidance. Boyan et al. [33] evaluated LLMs' responses to preventive pediatric dentistry questions and found them generally informative, despite occasional gaps. Guven et al. [34] examined chatbot responses to traumatic dental injury-related questions, noting that ChatGPT-4 and Google's Gemini produced more accurate and higher-quality answers than ChatGPT-3.5, although readability remained difficult. Similarly, Elkarmi et al. [35] evaluated ChatGPT's responses to FAQs on early childhood caries; while most answers were rated as useful or very useful, readability and actionability were low. Gugnani et al. [19] explored ChatGPT's usefulness from a maternal perspective, concluding that the chatbot generally provided clear and logical information, although some expert supplementation was still necessary. The findings of the present

study align with these prior works, indicating that ChatGPT-4's answers to FAQs posed by patients and parents regarding pediatric dentistry are highly accurate.

Studies evaluating ChatGPT's performance in answering expert-prepared questions provide further insight into its capacity to deliver dental information. Molena et al. [36] conducted a pilot study assessing ChatGPT's responses to general dentistry questions prepared by experts and found that while the artificial intelligence generally provided accurate answers, some responses were incomplete or contained inaccuracies. This highlights the potential of artificial intelligence models as supportive clinical decision tools when used under expert supervision, rather than as standalone decision-makers [36]. Another study evaluated the quality of ChatGPT's information on oral surgery, preventive dentistry, and oral cancer, revealing coherent delivery of complex information in preventive dentistry but lower accuracy in other domains [37]. In the present study, responses to curricular questions scored lower in accuracy and completeness compared to FAQs posed by patients and parents. Tokgöz Kaplan and Cankar [38] compared ChatGPT and Google's Gemini in answering clinical and curricular questions about dental avulsion. Four pediatric dentists evaluated accuracy and completeness, finding that ChatGPT provided more detailed responses, albeit with some outdated information, whereas Gemini offered shorter, more protocol-aligned answers. Both models showed limitations in technical depth, underscoring the importance of expert oversight. Similarly, Kusaka et al. [21] found that chatbot answers to pediatric clinical questions scored higher on systemic diseases or parental consultations topics but lower on dental anomalies, reflecting weaker technical performance. Rokhshad et al. [20] compared multiple chatbots against pediatric dentists using true/false questions and reported that although chatbots had acceptable internal consistency, their overall accuracy was significantly inferior to human experts. Collectively, these findings align with current literature and suggest that ChatGPT-4 may perform less effectively on highly technical content. While differences in scores were not statistically significant, the trend indicates that ChatGPT-4 offers acceptable but inconsistent performance for curricular pediatric dentistry questions. As in prior studies, expert supervision remains critical for technically complex topics.

Recent findings in pediatric dental literature further support our topic-specific performance observations. Kusaka et al. [21] reported that chatbot responses received the highest ratings for general topics such as "consultations from guardians" and "systemic diseases," while questions related to "dental abnormalities" received the lowest scores, indicating weaker performance in more technical areas. Similarly, Bayraktar and Nahir [18] found that among the academic questions evaluated, ChatGPT's poorest performance was in dental trauma, a topic requiring complex, case-specific pediatric expertise. These studies suggest that artificial intelligence systems generally struggle with content demanding clinical depth and individualized reasoning. This helps explain the lower performance we observed for nuanced topics such as fluoride and pulpal therapy. While fluoride remains a globally debated issue with varying perspectives, pulpal therapy involves multifactorial clinical decision-making [39–41], which poses challenges for artificial intelligence-generated responses. Conversely, ChatGPT-4 performed well in the domain of fissure sealants, likely due to the standardized nature of this preventive treatment and its widespread use in pediatric dentistry [42]. This suggests that artificial intelligence-based chatbots may be more effective in addressing topics with well-established guidelines and consistent clinical practices.

One of the key strengths of this study lies in its systematic, topic-based evaluation of ChatGPT-4's responses within pediatric dentistry—a clinically relevant and frequently consulted field by both patients/parents and dental professionals. As general-purpose generative artificial intelligence platforms become increasingly integrated into everyday life, particularly in their basic, publicly accessible forms, assessing their accuracy across various pediatric dentistry topics is crucial for guiding future model improvements and ensuring safe clinical integration. This study fills an important gap in the literature, especially given the limited number of comprehensive evaluations currently available. Moreover, the study design incorporated a balanced mix of FAQs posed by patients and parents alongside curricular questions developed by experts, allowing for a nuanced comparison of ChatGPT-4's strengths and limitations in addressing diverse levels of content complexity.

However, several limitations should be acknowledged. First, the evaluators in this study were not formally calibrated or trained prior to assessing the responses. Although Likert scale descriptors guided scoring and efforts were made to ensure consistency, the absence of formal calibration may have introduced subjective variation and potential bias. Furthermore, while the use of a single registered Plus account minimized session-based variability, it does not eliminate broader reproducibility concerns inherent to generative artificial intelligence systems, which can vary based on user type, location, and session history. This limitation affects the generalizability of our findings across different usage contexts. Another limitation relates to the inherent subjectivity of interpreting Likert scale-based scores. Although intra-rater consistency was improved through repeated scoring, future studies could benefit from employing multidimensional assessment tools that separately evaluate accuracy,

clarity, and clinical applicability. The use of a three-point Likert scale for completeness—although supported by prior literature—may have constrained the resolution in necessary to distinguish nuanced differences in response quality. Additionally, inter-rater reliability was not formally calculated. Furthermore, the sample size of pediatric dentists participating in the evaluation was modest, which may limit the statistical power and generalizability of the findings. Future studies with larger and more diverse evaluator populations are warranted to validate and extend these results. Moreover, the FAQs were collected indirectly via pediatric dentists rather than directly from parents or patients. While this method ensured clinical relevance and geographic representation, it may have introduced subjectivity based on clinicians' interpretations of parental or patient-related concerns. In addition, since the FAQ items were based on pediatric dentists' personal experiences, availability bias may have occurred, potentially leading to the overrepresantation of frequently encountered or relatively simple questions. This could have inadvertently inflated performance scores by favoring topics that are more familiar or less complex. In addition, this study did not include benchmarking against domain-specific LLMs or human-generated reference answers, which may have limited the scope of comparative evaluation. Finally, the performance of ChatGPT-4 in responding to case-specific or patient-tailored clinical scenarios was not assessed. Therefore, the model's ability to handle personalized, context-dependent inquiries remains untested. Future research should explore the responsiveness of artificial intelligence to individualized, real-world pediatric dental cases to better understand its utility in clinical decision-making.

This study highlights both the potential and limitations of general-purpose generative artificial intelligence platforms, such as ChatGPT-4, in supporting patient education and clinical communication in pediatric dentistry. ChatGPT-4 performs relatively well when answering FAQs posed by patients and parents, as well as expert-level curricular questions, especially in well-established topics like fissure sealants. However, it shows notable limitations when addressing more complex or context-sensitive topics, such as pulpal therapy and fluoride-related guidance. Although high average Likert scores suggest favorable perceptions of accuracy and completeness, these should not be taken as indicators of clinical safety or readiness for autonomous use. It is essential to recognize that artificial intelligence-generated content lacks clinical judgment and individual patient context, and therefore cannot replace the expertise of a trained dental professional.

Despite these constraints, general-purpose LLMs like ChatGPT-4 can still meaningfully contribute to patient engagement by providing accessible baseline information, alleviating pre-appointment anxiety, and enhancing understanding. When accompanied by appropriate oversight and clear disclaimers about their non-clinical nature, such tools can serve as valuable adjuncts to conventional care, particularly in the areas of communication and health literacy. However, recent literature highlights the necessity for rigorous ethical safeguards when integrating artificial intelligence systems into healthcare—especially in pediatric settings. Risk-averse deployment frameworks and provable explainability protocols have been proposed to improve transparency, prevent harm, and ensure alignment with core human values prior to clinical application [43, 44]. These frameworks promote embedding cautionary mechanisms and implementing structured human oversight to foster accountability and trust in high-stakes environments. Additionally, including two perception-based questionnaire items—evaluating participants' willingness to recommend or use artificial intelligence in clinical and educational contexts—provided further insight into the perceived utility of ChatGPT-4. While these subjective perspectives help contextualize the objective scores, they should be interpreted cautiously and not as endorsements of clinical validity.

Conclusion

This study presents a comprehensive, topic-specific evaluation of ChatGPT-4's performance in pediatric dentistry, highlighting both its potential and limitations. By assessing responses to both FAQs posed by patients and parents alongside expert-generated curricular questions across six key subdomains, the study offers a unique dual-perspective insight into the model's capabilities. While ChatGPT-4 generally produced coherent and informative answers—especially for well-established and standardized topics like fissure sealants—its reliability diminished in technically complex or context-sensitive areas, such as fluoride guidance and pulpal therapy. These findings emphasize that despite favorable average scores, with mean accuracy nearing 4 out of 5 and completeness exceeding 2 out of 3, ChatGPT-4's outputs should not be regarded as clinically validated or suitable for autonomous application without expert oversight. As generative artificial intelligence tools continue to develop and integrate into healthcare, further research is vital to enhance their contextual reasoning abilities and establish clear parameters for their safe, ethical, and expert-guided use in pediatric dental care.

While the observed effect sizes were generally small, suggesting modest practical differences between FAQs and curricular responses, these findings emphasize that statistical significance alone does not guarantee clinical relevance. Therefore, careful expert oversight remains essential when considering the integration of ChatGPT-4

into pediatric dental practice to ensure patient safety and optimal care outcomes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12903-025-06791-9.

Supplementary Material 1.

Acknowledgements
The figures in the article were created with the assistance of the ChatGPT-4 Omni (OpenAI, San Francisco, USA) large language model-based artificial intelligence software.

Clinical trial number
Not applicable.

Authors' contributions
B.S. and A.E.O. conceived the ideas; B.S. and A.E.O. collected the data; B.S. analysed the data; B.S. and A.E.O. led figure/table development; B.S. and A.E.O. led the writing. All author reviewed the final submission.

Funding
No funding was obtained for this study.

Data availability
The datasets used and/or analysed during the current study are available from the corresponding author (B.S.) on reasonable request.

Declarations

Ethics approval and consent to participate
No ethical approval and consent to participate was needed because this is not a human study, but only online information was used.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 9 January 2025 / Accepted: 13 August 2025
Published online: 24 September 2025

References

1. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large Language models release for medical applications: 1-year timeline and perspectives. J Med Syst. 2024;48:22. https://doi.org/10.1007/s10916-024-02045-3.
2. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med. 2024;99:22–7. https://doi.org/10.1097/ACM.0000000000005439.
3. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthc (Basel). 2023;11:887. https://doi.org/10.3390/healthcare11060887.
4. Zhang P, Kamel Boulos MN. Generative AI in medicine and healthcare: promises, opportunities and challenges. Future Internet. 2023;15:286. https://doi.org/10.3390/fi15090286.
5. Zendaoui N, Bouchemal N, Benabdelhafid M. AI-LMS: AI-based long-term monitoring system for patients in pandemics: COVID-19 case study. In: Mosbah M, Kechadi T, Bellatreche L, Gargouri F, editors. Model and data engineering. Cham: Springer; 2023. p. 14396. https://doi.org/10.1007/978-3-031-49333-1_20.
6. Keesara S, Jonas A, Schulman K. Covid-19 and health care's digital revolution. N Engl J Med. 2020;382:e82. https://doi.org/10.1056/NEJMp2005835.
7. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21:1–67.
8. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inf Assoc. 2023;30:1237–45. https://doi.org/10.1093/jamia/ocad072.
9. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT. Diagn Interv Imaging. 2023;104:269–74. https://doi.org/10.1016/j.diii.2023.02.003.
10. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. J Dent Res. 2020;99:769–74. https://doi.org/10.1177/0022034520915714.
11. Huang H, Zheng O, Wang D, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large Language model. Int J Oral Sci. 2023;15:29. https://doi.org/10.1038/s41368-023-00239-y.
12. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large Language models such as ChatGPT for dental medicine. J Esthet Restor Dent. 2023;35:1098–102. https://doi.org/10.1111/jerd.13046.
13. Freire Y, Santamaría Laorden A, Orejas Pérez J, Gómez Sánchez M, Díaz-Flores García V, Suárez A. ChatGPT performance in prosthodontics: Assessment of accuracy and repeatability in answer generation. J Prosthet Dent. 2024;131:659.e1-659.e6. https://doi.org/10.1016/j.prosdent.2024.01.018.
14. Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: evaluating the consistency and accuracy of endodontic question answers. Int Endod J. 2024;57:108–13. https://doi.org/10.1111/iej.13985.
15. Giannakopoulos K, Kavadella A, Aaqel Salim A, 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. J Med Internet Res. 2023;25:e51580. https://doi.org/10.2196/51580.
16. Abu Arqub S, Al-Moghrabi D, Allareddy V, Upadhyay M, Vaid N, Yadav S. Content analysis of AI-generated (ChatGPT) responses concerning orthodontic clear aligners. Angle Orthod. 2024;94:263–72. https://doi.org/10.2319/071123-484.1.
17. Gheisarifar M, Shembesh M, Koseoglu M, Fang Q, Afshari FS, Yuan JC, Sukotjo C. Evaluating the validity and consistency of artificial intelligence chatbots in responding to patients' frequently asked questions in prosthodontics. J Prosthet Dent. 2025. https://doi.org/10.1016/j.prosdent.2025.03.009. S0022-3913(25)00243-4.
18. Bayraktar Nahir C. Can ChatGPT be guide in pediatric dentistry? BMC Oral Health. 2025;25:9. https://doi.org/10.1186/s12903-024-05393-1.
19. Gugnani N, Pandit IK, Gupta M, Gugnani S, Kathuria S. Parental concerns about oral health of children: is ChatGPT helpful in finding appropriate answers? J Indian Soc Pedod Prev Dent. 2024;42:104–11. https://doi.org/10.4103/jisppd.jisppd_110_24.
20. 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. J Dent. 2024;144:104938. https://doi.org/10.1016/j.jdent.2024.104938.
21. 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 (Basel). 2024;14:2818. https://doi.org/10.3390/diagnostics14242818.
22. 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 Pediatr Parent. 2025;8:e62738. https://doi.org/10.2196/62738.
23. Johnson AJ, Singh TK, Gupta A, Sankar H, Gill I, Shalini M, et al. Evaluation of validity and reliability of AI chatbots as public sources of information on dental trauma. Dent Traumatol. 2025;41:187–93. https://doi.org/10.1111/edt.13000.
24. Gökcek Taraç M, Nale T. Artificial intelligence in pediatric dental trauma: do artificial intelligence chatbots address parental concerns effectively? BMC Oral Health. 2025;25:736. https://doi.org/10.1186/s12903-025-06105-z.
25. Perez-Pino A, Yadav S, Upadhyay M, Cardarelli L, Tadinada A. The accuracy of artificial intelligence-based virtual assistants in responding to routinely asked questions about orthodontics. Angle Orthod. 2023;93:427–32. https://doi.org/10.2319/100922-691.1.

26. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? BMC Med Inf Decis Mak. 2024;24:211. https://doi.org/10.1186/s12911-024-02619-8.
27. Acar AH. Can natural Language processing serve as a consultant in oral surgery? J Stomatol Oral Maxillofac Surg. 2024;125:101724. https://doi.org/10.1016/j.jormas.2023.101724.
28. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatol Oral Maxillofac Surg. 2023;124:101471. https://doi.org/10.1016/j.jormas.2023.101471.
29. Şişman AÇ, Acar AH. Artificial intelligence-based chatbot assistance in clinical decision-making for medically complex patients in oral surgery: a comparative study. BMC Oral Health. 2025;25:351. https://doi.org/10.1186/s12903-025-05732-w.
30. De Vito A, Colpani A, Moi G, Babudieri S, Calcagno A, Calvino V, et al. Assessing chatgpt's potential in HIV prevention communication: A comprehensive evaluation of accuracy, completeness, and inclusivity. AIDS Behav. 2024;28:2746–54. https://doi.org/10.1007/s10461-024-04391-2.
31. Coskun B, Yagiz BN, Ocakoglu B, Dalkilic G, Pehlivan E. Assessing the accuracy and completeness of artificial intelligence Language models in providing information on methotrexate use. Rheumatol Int. 2024;44:509–15. https://doi.org/10.1007/s00296-023-05473-5.
32. Hatia A, Doldo T, Parrini S, Chisci E, Cipriani L, Montagna L, et al. Accuracy and completeness of ChatGPT-Generated information on interceptive orthodontics: A multicenter collaborative study. J Clin Med. 2024;13:735. https://doi.org/10.3390/jcm13030735.
33. Guan B, Xu M, Zhang H, Ma S, Zhang S. Accuracy of large Language models for answering pediatric preventive dentistry questions. J Prev Treat Stomatol Dis. 2025;33:313–9. https://doi.org/10.12016/j.issn.2096-1456.202440370.
34. Guven Y, Ozdemir OT, Kavan MY. Performance of artificial intelligence chatbots in responding to patient queries related to traumatic dental injuries: A comparative study. Dent Traumatol. 2025;41:338–47. https://doi.org/10.1111/edt.13020.
35. Elkarmi R, Abu-Ghazaleh S, Sonbol H, Haha O, Al-Haddad A, Hassona Y. ChatGPT for parents' education about early childhood caries: A friend or foe? Int J Paediatr Dent. 2025;35:717–24. https://doi.org/10.1111/ipd.13283.
36. 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. https://doi.org/10.7759/cureus.65658.
37. Alsayed AA, Aldajani MB, Aljohani MH, Alamri H, Alwadi MA, Alshammari BZ, et al. Assessing the quality of AI information from ChatGPT regarding oral surgery, preventive dentistry, and oral cancer: an exploration study. Saudi Dent J. 2024;36:1483–9. https://doi.org/10.1016/j.sdentj.2024.09.009.
38. Tokgöz Kaplan T, Cankar M. Evidence-Based potential of generative artificial intelligence large Language models on dental avulsion: ChatGPT versus gemini. Dent Traumatol. 2025;41(2):178–86. https://doi.org/10.1111/edt.12999.
39. Veneri F, Vinceti M, Generali L, Giannone ME, Mazzoleni E, Birnbaum LS, et al. Fluoride exposure and cognitive neurodevelopment: systematic review and dose-response meta-analysis. Environ Res. 2023;221:115239. https://doi.org/10.1016/j.envres.2023.115239.
40. Coll JA, Seale NS, Vargas K, Marghalani AA, Al Shamali S, Graham L. Primary tooth vital pulp therapy: A systematic review and meta-analysis. Pediatr Dent. 2017;39:16–123.
41. Saxena N, Hugar SM, Soneta SP, Joshi RS, Dialani PK, Gokhale N. Evaluation of the treatment protocols in the management of pulpally involved young permanent teeth in children: A systematic review and meta-analysis. Int J Clin Pediatr Dent. 2022;15:S103–13. https://doi.org/10.5005/jp-journals-10005-2218.
42. Ahovuo-Saloranta A, Forss H, Walsh T, Nordblad A, Mäkelä M, Worthington HV. Pit and fissure sealants for preventing dental decay in permanent teeth. Cochrane Database Syst Rev. 2017;7:CD001830. https://doi.org/10.1002/14651858.CD001830.pub5.
43. Thurzo A, Thurzo V. Embedding fear in medical AI: A Risk-Averse framework for safety and ethics. AI. 2025;6(5):101. https://doi.org/10.3390/ai6050101.
44. Thurzo A, Provable AI. Ethics and explainability in medical and educational AI agents: trustworthy ethical firewall. Electronics. 2025;14(7):1294. https://doi.org/10.3390/electronics14071294.

Publisher's Note