Dinc *et al. BMC Oral Health* (2026) 26:1200
https://doi.org/10.1186/s12903-026-08502-4

BMC Oral Health

RESEARCH Open Access

Are chatbots reliable sources of information regarding fluoride in pediatric dentistry?

Dilara Dinc[1*], Simin Kocaaydin[2] and Sabiha Ceren Ilisulu[3]

Abstract

Aim To evaluate the accuracy and consistency of responses generated by artificial intelligence (AI) chatbots in pediatric dentistry, specifically concerning fluoride usage.

Study design Descriptive cross-sectional study.

Methods Four AI chatbots (ChatGPT, Gemini, Claude, Copilot) and four groups of dental professionals (pediatric dentists, general dentists, pediatric dentistry PhD students, and fifth-year dental students) answered 23 true–false questions based on IAPD, AAPD and EAPD guidelines. Each chatbot was tested 28 times per question in separate sessions. Accuracy was analyzed across four categories: Individual Topical Fluoride Applications, Professional Topical Fluoride Applications, Systemic Fluoride Applications, and Fluorosis. All groups were statistically compared with each other to evaluate differences in response accuracy across AI chatbots and human participant categories.

Results Significant differences were observed in the accuracy of chatbot responses across fluoride application categories ($p < 0.05$). Claude achieved perfect accuracy in Systemic Fluoride Applications (100%), while the other AI models performed lower—with ChatGPT scoring the lowest (94.3%)—and Gemini showed the highest accuracy in Fluorosis-related questions (76.8%). Among professionals, pediatric dentists (82.3%) consistently had the highest accuracy.

Statistics Chi-square and Fisher's Exact tests were used to assess differences in response accuracy between groups. A p-value < 0.05 was considered statistically significant.

Conclusions Claude and Gemini demonstrated greater reliability in fluoride-related questions than ChatGPT and Copilot. However, expert oversight remains crucial in pediatric dental care.

Keywords Accuracy, Artificial intelligence, Chatbots, Fluoride, Pediatric dentistry

*Correspondence:
Dilara Dinc
dr.dt.dilaradinc@gmail.com; ddinc@biruni.edu.tr
[1]Faculty of Dentistry, Department of Pediatric Dentistry, Biruni University, 10. Yıl Street, No: 45 Topkapı, Zeytinburnu, Istanbul 34010, Turkey
[2]Faculty of Dentistry, Department of Pediatric Dentistry, Kent University, Istanbul, Turkey
[3]Faculty of Dentistry, Department of Pediatric Dentistry, Istanbul University-Cerrahpaşa, Istanbul, Turkey

Introduction

Fluoride has been used in dentistry for over a century and is now considered the key factor behind the significant global reduction in dental caries [1]. It helps prevent dental caries in three key ways: -by reinforcing enamel through the formation of fluorapatite, -by aiding in enamel remineralization, and -by influencing bacterial metabolism to decrease acid production from cariogenic bacteria [2]. Systemic fluoride, professional topical fluoride applications, and home-use products are well-established components of community caries prevention programs [3]. While fluoride offers significant benefits, excessive exposure can harm both dental and overall health, leaving patients uncertain about safe usage levels. Concerns also arise from its inclusion in food, hygiene products, and water sources [4]. Misinformation about fluoride is prevalent on social media, often distorting its benefits. Likewise, the avoidance of fluoride in dental practices seems to be a consequence of misleading information spread online [5].

Artificial intelligence (AI) is based on computers carrying out human tasks and is actively used in many fields today, including healthcare services [6]. With advancements in AI, analysis of large amounts of data has become possible, which speeds up the decision-making process by providing accurate information [7]. A frequent use of AI is chatbots, which utilize text or speech to mimic human-like interactions [8]. AI in dentistry has the potential to be used to reduce administration, simplify diagnostics, combine various data types, and eventually lower healthcare costs [6]. In pediatric dentistry, AI can be used for detecting early childhood caries, analyzing radiographic images, identifying dental anomalies, detecting dental plaque, tracing cephalometric landmarks, and assessing growth patterns. These applications improve diagnostic efficiency while reducing errors and treatment time [7].

Companies such as OpenAI, Google, Anthropic, and Microsoft have developed advanced chatbots that can provide human-like responses. OpenAI's ChatGPT delivers accurate and pertinent responses using a large database [9, 10]. Google's Gemini can comprehend multimodal input and adapt to different kinds of data such as text, audio, and video [10]. Anthropic's Claude that generates responses that mimic those of a human and assists with problem-solving [8]. Microsoft's Copilot, originally known as Bing Chat, is an AI tool that combines language models with organizational tools systems to offer recommendations and guidance [9].

Recent studies have shown that Claude and Gemini consistently perform at a high level in clinical and educational assessments [11, 12]. In addition, evaluations based on national dental examinations demonstrated that Gemini Advanced, Claude, and Copilot achieved accuracy rates comparable to passing scores, reflecting their growing potential in dentistry-related applications [13, 14]. Furthermore, Gemini Advanced, Claude, and ChatGPT-4 Omni were reported to achieve high diagnostic accuracy in pediatric traumatic dental injuries, highlighting their emerging applicability in clinical pediatric oral health [15]. These findings underline the rapid improvement of chatbot capabilities and justify the growing interest in evaluating their reliability within pediatric dentistry.

Although parents increasingly rely on these chatbots for health information, the accuracy, consistency, and clinical reliability of AI-generated advice regarding fluoride use in children remain unknown. The rapid rise of AI-assisted information-seeking behavior has created a critical need to evaluate how accurately these systems address fluoride-related questions—particularly because fluoride is both widely used and widely misunderstood.

The reliability of chatbots that parents can use to obtain information about fluoride applications in children is unknown. There is no research evaluating the responses of different chatbots to questions about fluoride. The aim of the study is to evaluate the accuracy and consistency of responses from AI chatbots in answering questions related to fluoride usage in pediatric dentistry. In addition, AI chatbots were compared to pediatric dentists, general dentists, PhD tudents in the departments of pediatric dentistry, and dental students in answering questions related to fluoride usage in pediatric dentistry.

The null hypothesis of the study is that there is no statistically significant difference in the accuracy and consistency of fluoride-related responses provided by AI chatbots and human dental professionals.

Material and method

Study design and participants

The present descriptive cross-sectional study was conducted between December 2024 to February 2025. The study population was categorized into two main groups.

The first group comprised artificial intelligence–based conversational agents, including ChatGPT 4.0 Mini (OpenAI, USA), Gemini 1.5 (Google DeepMind, UK), Claude 3.7 Sonnet (Anthropic, USA), and Microsoft 365 Copilot (Microsoft, USA).

The second group consisted of human respondents, namely pediatric dentists, general dentists, pediatric dentistry PhD students, and fifth-year dental students. This study included pediatric dentists with at least two years of experience in their specialty, general dentists with at least two years of clinical experience, and third-year PhD students in pediatric dentistry, as well as fifth-year dental students.

The age range of participants was between 20 and 60 years. Only actively practicing dentists and currently enrolled dental students were included in the study.

Dentists and specialists who were not currently engaged in clinical practice or who had been away from the profession were excluded.

Participants were recruited through the mailing lists of the Turkish Society of Pediatric Dentistry, the Turkish Dental Association, and the Turkish Dental Association Student Branch. The ethics committee approval form and the informed consent form were attached to all emails, and participants took part in the study on a voluntary basis.

Sample size

The sample size was determined using the G*Power 3.1.9.2 software (Universitat Kiel, Germany). With a study power of 80% and an alpha error probability of 0.05 with an effect size of 0.25, the minimum sample size needed was 112. In the present study, 120 participants were included.

Questionnaire

The questionnaire was based on the guidelines of the International Association of Pediatric Dentistry (IAPD) [16], American Association of Pediatric Dentistry (AAPD) [2], and European Association of Pediatric Dentistry (EAPD) [17] which encompass the widely accepted approaches for fluoride. A total of 23 questions were prepared for the questionnaire and the questions were divided into four subheadings: -individual topical fluoride applications (n = 8), -professional topical fluoride applications (n = 8), -systemic fluoride applications (n = 5), -fluorosis (n = 3). The questionnaire was sent to participants as online Google Forms document, and answers were evaluated according to the guidelines as true or false.

All question statements were presented to each chatbot in an identical format during every repetition. Each query was initiated as a new conversation session, and the prompts were submitted at different time intervals using three independent user accounts to minimize sequential influence and inter-response dependency. All chatbot evaluations were performed using standard default settings, with browsing or web-search capabilities disabled. Since the items were formatted as true–false questions, responses allowed clear correctness categorization based on a predefined guideline-based answer key. If a chatbot failed to provide a definitive correct or incorrect statement, the answer was coded as incorrect. This standardized protocol was implemented to ensure reliability, minimize bias, and enhance the reproducibility of the study. To reduce dependency among repeated chatbot outputs, each question was asked in a new conversation session and submitted at different time intervals using three independent user accounts. This methodological approach was implemented to minimize sequential response patterns and reduce the risk of model-driven dependence among responses. Figures 1, 2, and 3 provide representative examples of the chatbot-generated responses included in the analysis.

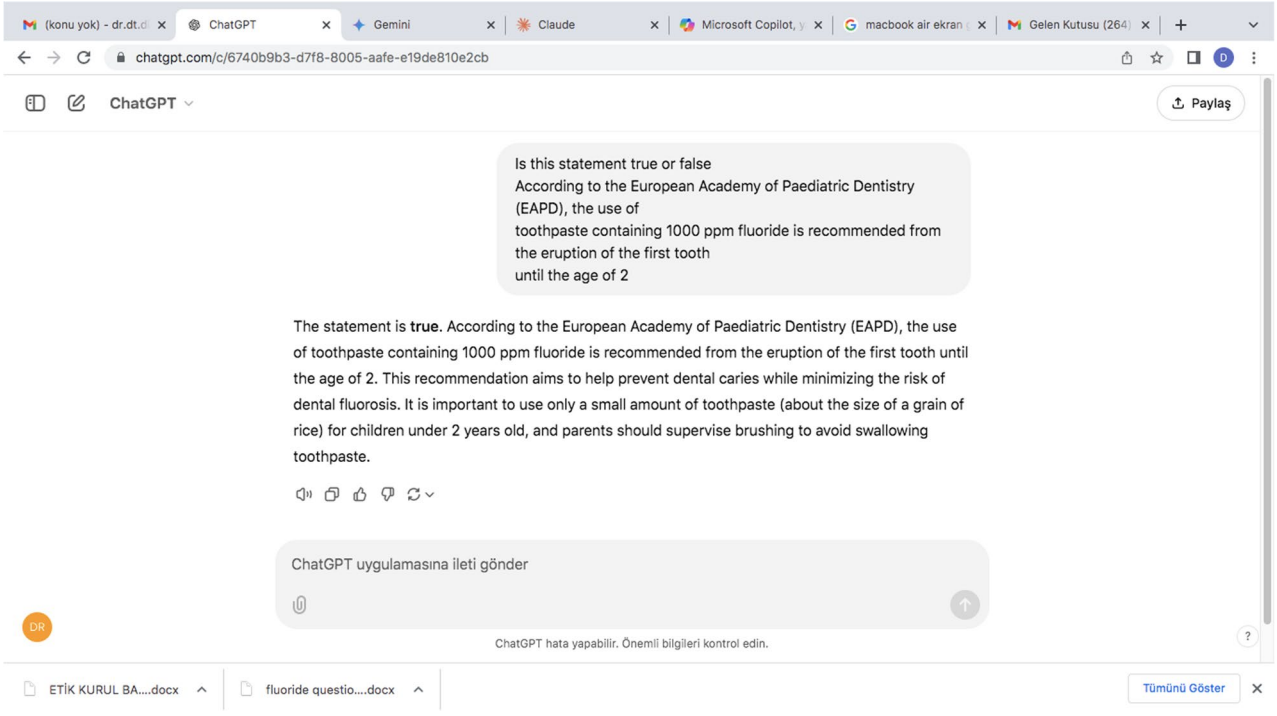

Fig. 1 ChatGPT answer to Question 2

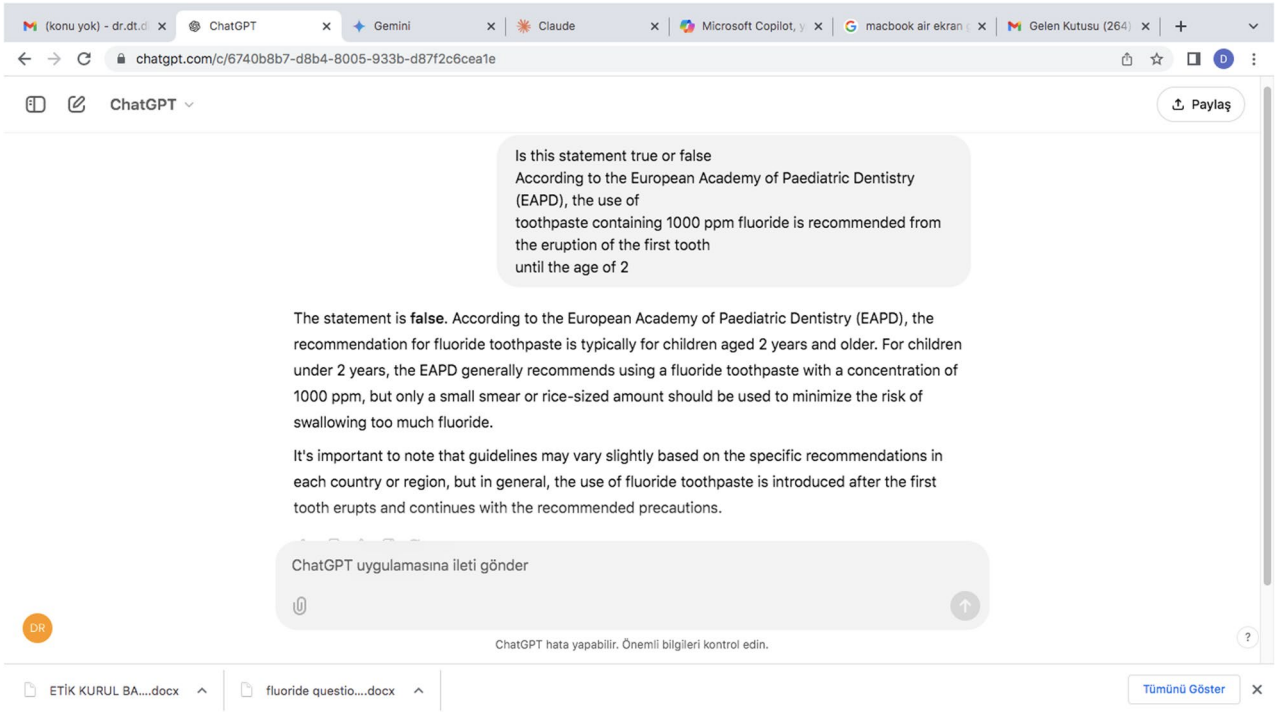

Fig. 2 ChatGPT another answer to Question 2

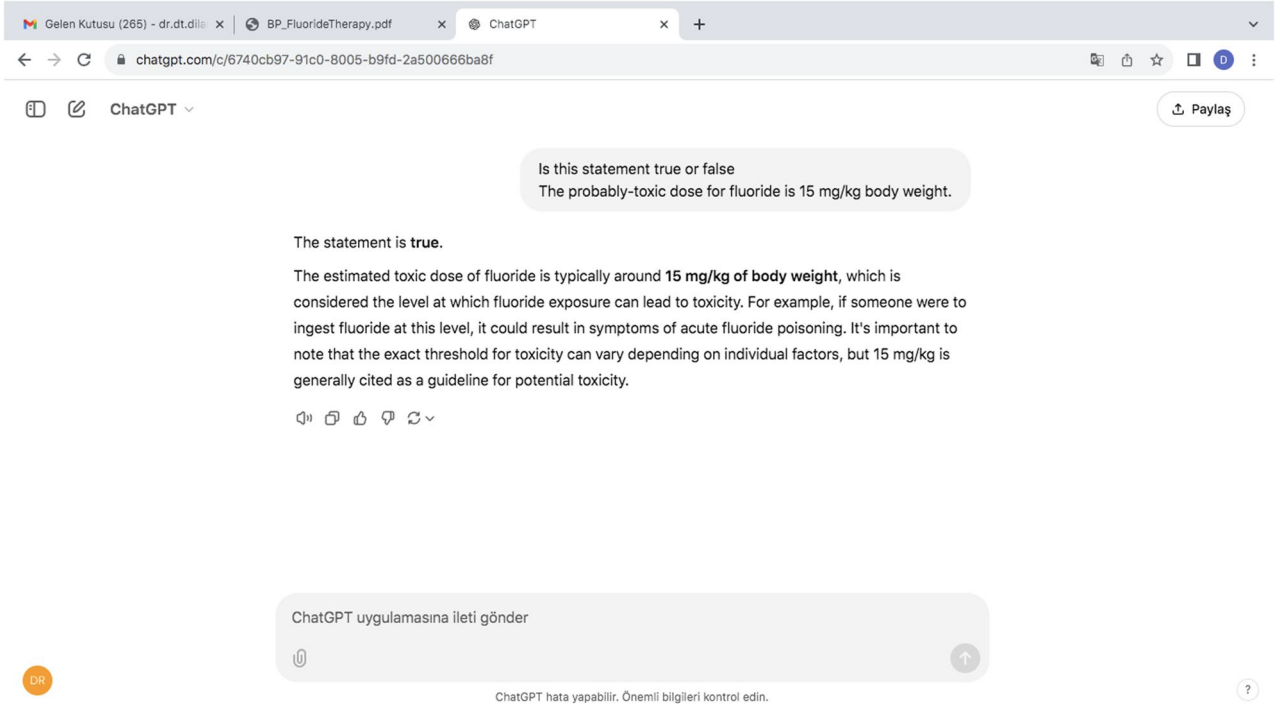

Fig. 3 ChatGPT answer to Question 8

In order to ensure the validity and reliability of the questionnaire, its content was developed based on current pediatric dentistry clinical guidelines. Content and face validity were assessed by three pediatric dentistry experts, who reviewed each item for accuracy and clarity. The questionnaire was pilot-tested on 10 dentists to evaluate comprehensibility and usability, after which minor wording modifications were made. Because the

questionnaire consisted of independent true/false knowledge items, internal consistency statistics were not applicable; therefore, reliability was assured through expert review and pilot testing. Two researchers verified data entry accuracy; however, questionnaire scoring was objective and based on predefined answer keys, so no inter-rater reliability calculation was required.

Statistical analysis

Statistical significance was determined using a threshold of $p < 0.05$ for all analyses. The analyses were performed using IBM SPSS version 27 (IBM Corp., USA). For testing the relationship between categorical variables, the Pearson Chi-Square test was used when the sample size assumption (expected value > 5) was met, and Fisher's Exact test was applied when this assumption was not met.

Results

The distribution of responses across AI types for each topic is presented in Table 1. Statistically significant differences were identified among AI systems regarding "Individual Topical Fluoride Applications," "Professional Topical Fluoride Applications," "Systemic Fluoride Applications," "Fluorosis," and total success levels ($p < 0.001$). Overall, Claude and Gemini consistently demonstrated the highest accuracy across most fluoride-related categories, whereas Copilot and ChatGPT exhibited comparatively lower performance. Notably, Claude achieved perfect accuracy in Systemic Fluoride Applications, while Gemini excelled in Fluorosis-related questions. In contrast, ChatGPT showed a marked decline in Fluorosis accuracy. Claude also yielded the highest overall correct response rate, followed closely by Gemini.

Table 2 shows the performance of dental professionals across the same topics. A total of 200 individuals were randomly invited to participate in the study via professional email groups. A total of 28 participants from each group were included in the study, resulting in an overall response rate of 14.0%. Significant differences were observed between the professional groups in all fluoride categories ($p < 0.05$). Pediatric dentists demonstrated the highest accuracy in each domain, followed by PhD students in pediatric dentistry. General dentists and dental students consistently showed lower performance, particularly in the Fluorosis category, where general dentists exhibited the lowest accuracy.

Table 3 presents the combined distribution of correct responses by both AI types and human professional groups. Statistically significant associations were found between response accuracy and both AI systems and dental professional categories across all topics ($p < 0.05$). Claude and Gemini showed the strongest performance overall, closely paralleling the accuracy levels of pediatric

Table 1 Distribution of responses given according to AI types for topics and the relationships between them

Topic		Incorrect			Correct				
	AI type	*n*	%	%C	*n*	%	%C	Test Statistics	*p*
Individual topical fluoride applications	ChatGPT	73[a]	32,6	23,6	151[a]	67,4	25,7	33,922	< 0,001*
	Gemini	76[a]	33,9	24,6	148[a]	66,1	25,2		
	Claude	51[a]	22,8	16,5	173[b]	77,2	29,5		
	Copilot	109[a]	48,7	35,3	115[b]	51,3	19,6		
Professional topical fluoride applications	ChatGPT	27[a]	12,1	36,0	197[b]	87,9	24,0	17,855	< 0,001*
	Gemini	10[a]	4,5	13,3	214[b]	95,5	26,1		
	Claude	10[a]	4,5	13,3	214[b]	95,5	26,1		
	Copilot	28[a]	12,5	37,3	196[b]	87,5	23,9		
Systemic fluoride applications	ChatGPT	8[a]	5,7	53,3	132[b]	94,3	24,2	9,087**	< 0,001*
	Gemini	4[a]	2,9	26,7	136[a]	97,1	25,0		
	Claude	0[a]	0,0	0,0	140[b]	100,0	25,7		
	Copilot	3[a]	2,1	20,0	137[a]	97,9	25,1		
Fluorosis	ChatGPT	54[a]	96,4	43,9	2[b]	3,6	2,0	62,802	< 0,001*
	Gemini	13[a]	23,2	10,6	43[b]	76,8	42,6		
	Claude	28[a]	50,0	22,8	28[a]	50,0	27,7		
	Copilot	28[a]	50,0	22,8	28[a]	50,0	27,7		
Total	ChatGPT	162[a]	25,2	31,0	482[b]	74,8	23,5	46,869	< 0,001*
	Gemini	103[a]	16,0	19,7	541[b]	84,0	26,3		
	Claude	89[a]	13,8	17,0	555[b]	86,2	27,0		
	Copilot	168[a]	26,1	32,2	476[b]	73,9	23,2		

Different superscripts small letters indicate significance *$p < 0.05$,

%: Row percentage, C%: Column percentage for responses

Different letters indicate statistically significant differences between column percentages within the same row, based on adjusted *p*-values (Bonferroni method) following the Chi-square or Fisher's exact test

Table 2 Distribution of responses given according to dental professional groups for topics and the relationships between them

Topic		Incorrect			Correct				
	Dental Professional Groups	***n***	**%**	**%C**	***n***	**%**	**%C**	**Test Statistics**	***p***
Individual Topical Fluoride Applications	Pediatric dentists	52[a]	21,7	16,0	188[b]	78,3	29,7	50,385	< 0,001*
	PhD students in pediatric dentistry	58[a]	25,0	17,8	174[b]	75,0	27,4		
	General dentists	115[a]	47,9	35,3	125[b]	52,1	19,7		
	5th grade dentistry students	101[a]	40,7	31,0	147[b]	59,3	23,2		
Professional Topical Fluoride Applications	Pediatric dentists	32[a]	13,3	16,3	208[b]	86,7	27,2	19,827	< 0,001*
	PhD students in pediatric dentistry	37[a]	15,9	18,9	195[a]	84,1	25,5		
	General dentists	64[a]	26,7	32,7	176[b]	73,3	23,0		
	5th grade dentistry students	63[a]	25,4	32,1	185[b]	74,6	24,2		
Systemic Fluoride Applications	Pediatric dentists	26[a]	17,3	16,6	124[b]	82,7	28,0	14,319	0,003*
	PhD students in pediatric dentistry	34[a]	23,4	21,7	111[a]	76,6	25,1		
	General dentists	54[a]	36,0	34,4	96[b]	64,0	21,7		
	5th grade dentistry students	43[a]	27,7	27,4	112[a]	72,3	25,3		
Fluorosis	Pediatric dentists	12[a]	20,0	12,6	48[b]	80,0	33,1	20,083	< 0,001*
	PhD students in pediatric dentistry	19[a]	32,8	20,0	39[a]	67,2	26,9		
	General dentists	34[a]	56,7	35,8	26[b]	43,3	17,9		
	5th grade dentistry students	30[a]	48,4	31,6	32[a]	51,6	22,1		
Total	Pediatric dentists	122[a]	17,7	15,8	568[b]	82,3	28,6	96,386	< 0,001*
	PhD students in pediatric dentistry	148[a]	22,2	19,1	519[b]	77,8	26,1		
	General dentists	267[a]	38,7	34,5	423[b]	61,3	21,3		
	5th grade dentistry students	237[a]	33,2	30,6	476[b]	66,8	24,0		

Different superscripts small letters indicate significance

*$p < 0.05$, %: Row percentage, C%: Column percentage for responses

Different letters indicate statistically significant differences between column percentages within the same row, based on adjusted *p*-values (Bonferroni method) following the Chi-square or Fisher's exact test

dentists and PhD students in pediatric dentistry. In contrast, general dentists and certain AI systems—particularly ChatGPT in Fluorosis-related items—demonstrated notably lower accuracy.

Overall, Claude and Gemini were the best-performing AI tools, while pediatric dentists were the most successful human group. General dentists and some AI types, including ChatGPT and Copilot, displayed significantly lower accuracy across the evaluated fluoride topics.

Discussion

Dental fluoride research has been ongoing for over a century, providing substantial evidence of its effectiveness in preventing dental caries and improving oral health [18]. In recent years, the global debate on fluoride has been divided into two groups: those who support topical fluoride applications—citing scientific evidence of their effectiveness in preventing tooth decay when used appropriately in children—and the anti-fluoride lobby [4]. Based on scientific evidence, topical fluoride applications, combined with community-based fluoride initiatives, are widely recognized as the gold standard for preventing dental caries, provided that appropriate attention is given to the risks of acute and chronic toxicity [19]. This study is one of the most detailed in dentistry to assess the accuracy of various chatbots compared to clinicians in answering fluoride-related questions.

Chatbots, powered by vast datasets drawn from diverse textual sources such as books, articles, and web content, leverage advanced deep learning techniques—including neural networks—to understand and generate responses [20]. While some studies affirm the effectiveness of chatbot systems and the relevance of their responses [21], others point out their limitations and caution against potential shortcomings, emphasizing the importance of critical evaluation [22, 23]. Additionally, the literature includes research examining the reliability and consistency of chatbots in clinical settings, including both medical and dental fields [8]. Strong et al. [24] emphasized that although ChatGPT performed satisfactorily on approximately half of the questions derived from clinical reasoning exam cases, its consistency was limited and its accuracy varied with each individual query. According to Hirosawa et al. [25], in clinical cases involving common chief complaints, ChatGPT-3.5 demonstrated a high level of diagnostic accuracy in generating differential diagnoses.

Similar to this study, some studies involved both chatbots and clinicians, presenting identical questions to both groups in order to compare the consistency and accuracy of human and chatbot responses [4, 21]. A recent study reported that the ChatGPT 3.5 chatbot provided answers to questions in the field of pediatric dentistry that were similar to those given by specialists, suggesting that it can

Table 3 Distribution of responses by AI types and dental professional groups across topics and the relationships between them

Topic		Incorrect			Correct				
	AI types and Dental Professional Groups	***n***	**%**	**%C**	***n***	**%**	**%C**	**Test Statistics**	***p***
Individual Topical Fluoride Applications	ChatGPT	73[a]	32,6	11,5	151[a]	67,4	12,4	84,312	< 0,001*
	Gemini	76[a]	33,9	12,0	148[a]	66,1	12,1		
	Claude	51[a]	22,8	8,0	173[b]	77,2	14,2		
	Copilot	109[a]	48,7	17,2	115[b]	51,3	9,4		
	Pediatric dentists	52[a]	21,7	8,2	188[b]	78,3	15,4		
	PhD students in pediatric dentistry	58[a]	25,0	9,1	174[b]	75,0	14,3		
	General dentists	115[a]	47,9	18,1	125[b]	52,1	10,2		
	5th grade dentistry students	101[a]	40,7	15,9	147[b]	59,3	12,0		
Professional Topical Fluoride Applications	ChatGPT	27[a]	12,1	10,0	197[a]	87,9	12,4	90,751	< 0,001*
	Gemini	10[a]	4,5	3,7	214[b]	95,5	13,5		
	Claude	10[a]	4,5	3,7	214[b]	95,5	13,5		
	Copilot	28[a]	12,5	10,3	196[a]	87,5	12,4		
	Pediatric dentists	32[a]	13,3	11,8	208[a]	86,7	13,1		
	PhD students in pediatric dentistry	37[a]	15,9	13,7	195[a]	84,1	12,3		
	General dentists	64[a]	26,7	23,6	176[b]	73,3	11,1		
	5th grade dentistry students	63[a]	25,4	23,2	185[b]	74,6	11,7		
Systemic Fluoride Applications	ChatGPT	8[a]	5,7	4,7	132[b]	94,3	13,4	150,292	< 0,001*
	Gemini	4[a]	2,9	2,3	136[b]	97,1	13,8		
	Claude	0[a]	0,0	0,0	140[b]	100,0	14,2		
	Copilot	3[a]	2,1	1,7	137[b]	97,9	13,9		
	Pediatric dentists	26[a]	17,3	15,1	124[a]	82,7	12,6		
	PhD students in pediatric dentistry	34[a]	23,4	19,8	111[b]	76,6	11,2		
	General dentists	54[a]	36,0	31,4	96[b]	64,0	9,7		
	5th grade dentistry students	43[a]	27,7	25,0	112[b]	72,3	11,3		
Fluorosis	ChatGPT	54[a]	96,4	24,8	2[b]	3,6	0,8	92,633	< 0,001*
	Gemini	13[a]	23,2	6,0	43[b]	76,8	17,5		
	Claude	28[a]	50,0	12,8	28[a]	50,0	11,4		
	Copilot	28[a]	50,0	12,8	28[a]	50,0	11,4		
	Pediatric dentists	12[a]	20,0	5,5	48[b]	80,0	19,5		
	PhD students in pediatric dentistry	19[a]	32,8	8,7	39[b]	67,2	15,9		
	General dentists	34[a]	56,7	15,6	26[a]	43,3	10,6		
	5th grade dentistry students	30[a]	48,4	13,8	32[a]	51,6	13,0		
Total	ChatGPT	162[a]	25,2	12,5	482[a]	74,8	11,9	190,804	< 0,001*
	Gemini	103[a]	16,0	7,9	541[b]	84,0	13,4		
	Claude	89[a]	13,8	6,9	555[b]	86,2	13,7		
	Copilot	168[a]	26,1	13,0	476[a]	73,9	11,8		
	Pediatric dentists	122[a]	17,7	9,4	568[b]	82,3	14,1		
	PhD students in pediatric dentistry	148[a]	22,2	11,4	519[a]	77,8	12,8		
	General dentists	267[a]	38,7	20,6	423[b]	61,3	10,5		
	5th grade dentistry students	237[a]	33,2	18,3	476[b]	66,8	11,8		

Different superscripts small letters indicate significance

*$p < 0.05$, %: Row percentage, C%: Column percentage for responses

Different letters indicate statistically significant differences between column percentages within the same row, based on adjusted *p*-values (Bonferroni method) following the Chi-square or Fisher's exact test

assist both dentists and patients, although improvements are still needed [21] while the another study showed that ChatGPT was found to be sufficient and comprehensive in fluoride releated questions [4]. In this study, although the percentage of correct answers given by ChatGPT was lower compared to the other chatbots included, its accuracy rate was still higher than its incorrect answer rate and was considered sufficient. The results of the current study are consistent with the findings of previous studies [4, 21]. On the other hand, Rokhshad et al. [8] posed pediatric dentistry-related questions to AI-powered chatbots and dentists with varying education levels, finding that pediatric dentists were significantly more accurate than both other clinicians and chatbots similar to our

study, and that ChatGPT demonstrated the highest accuracy among the chatbots. In the present study, ChatGPT demonstrated lower accuracy compared to Claude, Gemini, and pediatric dentists, with Claude consistently outperforming all other chatbots and general dentists and fifth-year students across most question categories. In a previous study, pediatric dentistry-related questions were presented to general dentists, dental students, and AI-supported chatbots, with the chatbots exhibiting lower performance compared to both the dentists and the students [26]. Furthermore, Mohammad-Rahimi et al. [11] have revealed considerable discrepancies in accuracy among different chatbots, as also demonstrated by our findings [11]. These variations in chatbot performance are believed to be influenced by differences in study design, regional and cultural differences, and the time periods in which the studies were conducted. Furthermore, real-world fluoride delivery is shaped by clinical workflow barriers and caregiver-related factors, which may influence how pediatric dentists interpret and apply guideline-based decisions in practice [27].

Sismanoglu and Capan [14] evaluated the academic performance of ChatGPT-4.0 and Gemini using questions from the Turkish Dental Specialty Examination (DUS), reporting that both achieved passing scores while also identifying knowledge gaps and stressing the need for caution before fully relying on chatbot outputs [13]. Furthermore, as contrasting study by Jung et al. [28] in Korea found that none of the chatbots evaluated met the minimum passing score required by the Pediatric Dentistry National Examination. In the present study, it was found that Gemini and Claude chatbots exhibited the highest levels of accuracy in answering questions related to fluoride, even surpassing the accuracy rates of general dentists and dental students in Turkey, who are expected to take similar examinations. A comparison of the results from these studies reveals that although both emphasize the ability of chatbots to deliver accurate responses within a specialized field, discrepancies in accuracy levels are evident, with certain chatbots exhibiting superior performance based on factors such as the type of exam and the geographical context. The superior performance of Claude and Gemini in this study may be attributed to differences in their underlying model architecture, training data composition, and reinforcement mechanisms.

Additionally, National and regional variations in pediatric fluoride guidelines may have contributed to performance differences between chatbots and dental professionals. While our answer key was based on international recommendations (AAPD, EAPD, ADA), local protocols regarding toothpaste concentrations and supplement use can differ, meaning that some responses marked incorrect in this study may still be appropriate within local practices [2, 3, 29].

The rapidly evolving nature of AI systems must also be considered when interpreting these findings. Most of the studies cited in the literature are recent, reflecting the accelerated development of large language models [6–8]. Because generative AI models undergo frequent updates, improvements in training data, architecture, and alignment strategies may significantly influence their performance over time. Consequently, the differences observed between chatbot models in this study may partly reflect their developmental stage at the time of data collection rather than intrinsic long-term performance characteristics. Similar observations in recent health-related AI evaluations also emphasize that temporal factors and update cycles can directly impact accuracy and reliability, underscoring the need for continuous re-assessment of AI tools in clinical contexts [9, 10].

This study has several limitations. First, all questions were presented in English, which may have advantaged AI models predominantly trained on English data, limiting the applicability of results to non-English contexts. Second, the focus on fluoride-related content restricts the generalizability of the findings to other areas of dental or medical knowledge. Third, as AI models are continuously updated, the results reflect performance at a specific point in time and may vary with future versions. Additionally, the clinical implications of these findings should be interpreted within this context; although some AI systems demonstrated higher accuracy, the variability observed across models indicates that chatbots cannot yet replace professional guidance in pediatric dentistry. Their inconsistent performance highlights the risk of misinformation for parents relying solely on AI-generated advice. Finally, the use of a binary (correct/incorrect) evaluation system may have overlooked partially accurate or contextually appropriate responses, particularly in complex clinical scenarios.

Conclusion

This study revealed substantial variability in the accuracy of AI chatbots when addressing fluoride-related questions in pediatric dentistry. Claude and Gemini demonstrated the highest reliability, whereas ChatGPT and Copilot showed notable limitations, particularly in fluorosis-related responses. Although some AI systems may serve as supportive educational tools, they cannot replace the clinical judgment of dental professionals. Continued monitoring of chatbot performance and improvements in model training are essential for ensuring safe and accurate use. Future research should explore their effectiveness across broader dental topics and multilingual contexts.

Limitations

This study is limited to fluoride-related content in pediatric dentistry, which may reduce generalizability to broader clinical areas. Although measures were taken to minimize dependency among chatbot responses, repeated outputs cannot be considered fully independent. Moreover, because AI systems and clinical guidelines evolve rapidly, the reported chatbot performance reflects a specific point in time. Finally, humans answered each item once while chatbots provided multiple outputs, and therefore comparisons between the two groups should be interpreted with caution.

Abbreviations

ADA	American Dental Association
AI	Artificial intelligence
AAPD	American Academy of Pediatric Dentistry
EAPD	European Academy of Pediatric Dentistry
IAPD	International Association of Pediatric Dentistry

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12903-026-08502-4.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

Not applicable.

Authors' contributions

DD: Conceptualization, Methodology, Investigation, Writing – original draft preparation, Supervision.SK: Data curation, Formal analysis, Writing – review & editing.SI: Investigation, Data curation, Writing – review & editing.

Funding

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

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The Scientific Research Ethical Committee of Biruni University approved this study and all procedures involving human participants, ensuring that the study was conducted in accordance with the principles of the Declaration of Helsinki (2024-BİAEK/04–14). Written informed consent was obtained from all participants.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 11 August 2025 / Accepted: 27 April 2026
Published online: 06 May 2026

References

1. Buzalaf MAR, Pessan JP, Honório HM, Ten Cate JM. Mechanisms of action of fluoride for caries control. Monogr Oral Sci. 2011;22:97–114.
2. American Academy of Pediatric Dentistry. Policy on use of fluoride. The Reference Manual of Pediatric Dentistry. Chicago, Ill.: American Academy of Pediatric Dentistry; 2023. pp. 100–2.
3. Toumba KJ, Twetman S, Splieth C, Parnell C, Van Loveren C, Lygidakis NA. Guidelines on the use of fluoride for caries prevention in children: an updated EAPD policy document. Eur Arch Paediatr Dent. 2019;20:507–16.
4. Buldur M, Sezer B. Can Artificial Intelligence effectively respond to frequently asked questions about fluoride usage and effects? A qualitative study on ChatGPT. Fluoride. 2023;56(3). https://dergipark.org.tr/tr/pub/mutftd/article/1746766.
5. Lotto M, Sá Menezes T, Zakir Hussain I, Tsao SF, Ahmad Butt Z, Morita P, et al. Characterization of false or misleading fluoride content on Instagram: infodemiology study. J Med Internet Res. 2022;24(5):e37519.
6. Schwendicke FA, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. J Dent Res. 2020;99(7):769–74.
7. Mahajan K, Kunte SS, Patil KV, Shah PP, Shah RV, Jajoo SS. Artificial intelligence in pediatric dentistry—A systematic review. J Dent Res Rev. 2023;10(1):7–12.
8. 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.
9. Hiwa DS, Abdalla SS, Muhialdeen AS, Hamasalih HM, Karim SO. Assessment of nursing skill and knowledge of ChatGPT, Gemini, Microsoft Copilot, and Llama: a comparative study. Barw Med J. 2024;2(2):3–6.
10. Fattah FH, Salih AM, Salih AM, Asaad SK, Ghafour AK, Bapir R, et al. Comparative analysis of ChatGPT and Gemini (Bard) in medical inquiry: a scoping review. Front Digit Health. 2025;7:1482712.
11. Mohammad-Rahimi H, Ourang SA, Pourhoseingholi MA, Dianat O, Dummer PMH, Nosrat A. Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. Int Endod J. 2024;57(3):305–14.
12. Wójcik D, Adamiak O, Czerepak G, Tokarczuk O, Szalewski L. A bi-linguistic comparative analysis of ChatGPT-4, Gemini, and Claude performance on Polish medical–dental final examinations. Sci Rep. 2023;15(1):33083.
13. ArılıÖztürk E, Turan Gökduman C, Çanakçi BC. Evaluation of the performance of ChatGPT-4 and ChatGPT-4o as a learning tool in endodontics. Int Endod J. 2025. https://doi.org/10.1111/iej.14217.
14. Sismanoglu S, Sirinoglu Capan B. Performance of artificial intelligence on Turkish dental specialization exam: can ChatGPT-4.0 and Gemini advanced achieve comparable results to humans? BMC Med Educ. 2025;25(1):214.
15. Sezer B, Aydoğdu T. Performance of Advanced Artificial Intelligence Models in Traumatic Dental Injuries in Primary Dentition: A Comparative Evaluation of ChatGPT-4 Omni, DeepSeek, Gemini Advanced, and Claude 3.7. Appl Sci. 2025;15(14):7778.
16. IAPD Foundational Articles and Consensus Recommendations: Use of Fluoride for Caries Prevention. 2022. http://www.iapdworld.org/2022_05_use-of-fluoride-for-caries-prevention
17. Toumba KJ, Twetman S, Splieth C, Parnell C, Van Loveren C, Lygidakis NA. Guidelines on the use of fluoride for caries prevention in children: an updated EAPD policy document. Eur Archives Pediatr Dentistry. 2019;20(6):507–16.
18. Unde MP, Patil RU, Dastoor PP. The untold story of fluoridation: revisiting the changing perspectives. Indian J Occup Environ Med. 2018;22(3):121–7.
19. Ullah R, Zafar MS, Shahani N. Potential fluoride toxicity from oral medicaments: a review. Iran J Basic Med Sci. 2017;20(8):841.
20. Huynh LM, Bonebrake BT, Schultis K, Quach A, Deibert CM. RETRACTED: New Artificial Intelligence ChatGPT Performs Poorly on the 2022 Self-assessment Study Program for Urology. Urol Pract. 2023;10(4):409–15.
21. Bayraktar Nahir C. Can ChatGPT be guide in pediatric dentistry? BMC Oral Health. 2025;25(1):9.
22. Helvacioglu-Yigit D, Demirturk H, Ali K, Tamimi D, Koenig L, Almashraqi A. Evaluating artificial intelligence chatbots for patient education in oral and maxillofacial radiology. Oral Surg Oral Med Oral Pathol Oral Radiol. 2025. Epub ahead of print. https://doi.org/10.1016/j.oooo.2025.01.001.
23. Mustuloğlu Ş, Deniz BP. Evaluation of chatbots in the emergency management of avulsion injuries. Dent Traumatol. 2025;41(4):437–44.
24. Strong E, DiGiammarino A, Weng Y, Basaviah P, Hosamani P, Kumar A, et al. Performance of ChatGPT on free-response, clinical reasoning exams. medRxiv. 2023. https://doi.org/10.1101/2023.03.02.23286751.
25. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained

transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health. 2023;20(4):3378.
26. Azadi A, Gorjinejad F, Mohammad-Rahimi H, Tabrizi R, Alam M, Golkar M. Evaluation of AI-generated responses by different artificial intelligence chatbots to the clinical decision-making case-based questions in oral and maxillofacial surgery. Oral Surg Oral Med Oral Pathol Oral Radiol. 2024;137(6):587–93.
27. Shvydko N, Jensen JL, O'Neill TR. Barriers and facilitators to delivery of fluoride varnish application in pediatric well-child visits: a post-implementation analysis. Glob Implement Res Appl. 2025;5:418–26.
28. 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. J Korean Acad Pediatr Dent. 2024;51(3):299–309.
29. Alamoudi RM, Marta FM, Alqahtani FA, Alshahri AM, Aboqraihah MN, Alharbi RE, Alaama IA, Alnasser AA, Kurdi DM, Albalawi RM. Fluoride use in pediatric dentistry: balancing benefits and risks. Int J Community Med Public Health. 2025;12(2). https://dx.doi.org/10.18203/2394-6040.ijcmph20250048.

Publisher's note