BMC Oral Health

RESEARCH

Open Access

Comparative evaluation of the performance of different AI chatbots in answering pediatric dentistry questions in the Turkish dental specialization exam

Sena Sakin Ulubay[1*]

Abstract

Background Artificial intelligence-powered chatbots, whose use is becoming increasingly widespread, are frequently used in medicine and dentistry for tasks such as generating academic texts, student education, clinical diagnosis, and patient information. Given their inherent limitations, the accuracy of chatbot responses in the healthcare field is a crucial issue. Evaluating the performance of artificial intelligence in national examinations can provide objective data regarding its use as an educational support tool. This study aims to comparatively evaluate the performance of various artificial intelligence chatbots on questions from the pediatric dentistry section of the DUS (Dental Specialization Exam).

Methods A total of 127 questions asked in the DUS in the field of pediatric dentistry between 2012 and 2021 were included in the study. The questions were divided into nine subject-area categories, two exam years, and two question types: clinical case-based and theoretical knowledge-based. The ChatGPT-5, ChatGPT-4o, Gemini 2.5 Pro, Gemini 2.5 Flash, Claude Opus 4, and Claude Sonnet 4 models were evaluated in the study. Descriptive data are presented as numbers and percentages. Statistical analyses were performed using Cochran's Q test ($p < 0.05$); pairwise comparisons were made using the McNemar test and Bonferroni correction.

Results Gemini 2.5 Pro answered 93.7% of the questions correctly; ChatGPT-5 and Claude Opus 4 answered 88.9%, ChatGPT 4 and Gemini 2.5 Flash answered 87.4%, and Claude Sonnet 4 answered 83.4%. Gemini 2.5 Pro's performance in correctly answering questions was significantly higher than that of Claude Sonnet 4 ($p = 0.001$). No statistically significant differences were found in the chatbots' correct-answer performance when comparing by subject, year, and question type (clinical case and theoretical knowledge) ($p > 0.05$).

Conclusions This study demonstrates that chatbots perform well at answering exam questions in pediatric dentistry and may serve aspromising supplementary educational tool.

Keywords Chatbots, ChatGPT, Claude, Dental specialization exam, Exam, Gemini, Pediatric dentistry

*Correspondence:
Sena Sakin Ulubay
senasakn@gmail.com
[1]Department of Pediatric Dentistry, Faculty of Dentistry, Recep Tayyip Erdogan University, Rize, Turkey

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

Artificial intelligence (AI) based chatbots are powered by Large Language Models, trained on large datasets to perform complex language understanding and generation tasks [1]. Large Language Models, which use deep learning algorithms, learn language structure from large amounts of text such as books, articles, and websites, and produce human-like responses [2]. The use of Large Language Models, which have developed rapidly in recent years, is becoming increasingly widespread in health, education, and scientific communication [2]. The first generative Pre-Trained Transformer model based on the Large Language Model technology, released by OpenAI in 2018 (ChatGPT, OpenAI, LLC, San Francisco, CA, USA), marked a significant milestone in artificial intelligence [3]. ChatGPT, a large language model that uses artificial intelligence and machine learning to converse with users, provides meaningful, contextually relevant responses to natural-language inputs based on a comprehensive knowledge base [3]. Another AI application offered as an alternative to ChatGPT is Gemini, an AI-based chatbot that answers user questions, participates in conversations, and generates creative content (Google, Mountain View, CA). Gemini provides advanced capabilities in text-based interactions, offering users quick and effective solutions to their requests [4]. Claude (Anthropic, California, USA) is another popular chatbot that can provide human-like responses [4].

The performance of AI chatbots can be affected by many factors, including the frequency of model updates, the complexity of questions, and the intelligibility of commands [5, 6]. AI-powered chatbots are frequently used in medicine and dentistry for tasks such as generating academic texts, student training, clinical diagnosis, and patient information [7, 8]. However, given the general limitations of AI's response generation, the accuracy of chatbot responses in healthcare is a critical issue [9]. Evaluating the performance of artificial intelligence in national medical and dental exams can provide objective data on its use as an educational support tool [10]. For this purpose, many studies have been conducted to measure the performance of various chatbots in different medical and dental exams [11–13]. These studies have reported that models like ChatGPT-4 achieve high accuracy on the United States Medical Licensing Exam and surpass previous best performances [11]. Similarly, in dentistry, chatbots have demonstrated significant proficiency in answering questions from the Integrated National Board of Dentistry Exam, with models such as ChatGPT-4 achieving remarkable accuracy [12]. These findings suggest that chatbots can serve as valuable resources for medical and dental students, providing explanations, test simulations, and personalized learning pathways. However, while their performance on standardized tests is impressive, integrating these models into real-world educational settings requires careful consideration to ensure they complement traditional teaching methods and fill gaps in education.

Recently, the demand for specialist training in dentistry, as in medicine, has been steadily increasing. In Türkiye, the Dental Specialization Exam (DUS) is a mandatory national examination for dentists seeking admission to specialist training programs. Organized by the Student Selection and Placement Center (ÖSYM), the exam consists of multiple-choice questions that measure candidates' basic dental knowledge and clinical competencies. Multiple-choice questions are the most commonly used objective method for measuring student learning because they allow assessment of a wide range of topics [13]. AI-powered chatbots are known to be highly effective at answering multiple-choice questions [14].

Dental studies have frequently focused on ChatGPT, whereas research on other large language models, such as Gemini and Claude, remains limited [15–18]. This study aims to evaluate the performance of three AI chatbots (ChatGPT-5, ChatGPT 4, Gemini 2.5 Pro, Gemini 2.5 Flash, Claude Opus 4, and Claude Sonnet 4), available in both paid and free versions, on questions asked in the pediatric dentistry section of the DUS, and to examine them comparatively based on subject matter, year, and question category.

Methods

Ethical approval

Since the study did not involve human or animal subjects, ethical committee approval is not required.

Study design

The dataset consists of pediatric dentistry questions from the DUS exams (13 exams) between 2012 and 2021, obtained from an open-source question bank (https://www.osym.gov.tr/TR,15070/dus-cikmissorular.html). Because ÖSYM ceased publishing questions from 2022 onwards, questions from exams after that year, as well as 3 questions canceled by ÖSYM, were excluded from the study. The analysis was conducted on 127 questions. Three of these questions contained images (Question 71 from the 2019 exam; question 72 from the 2019 exam; question 79 from the 2018 exam). All questions included in the study had five options and only one correct answer. The questions were categorized into nine subject-matter categories (Table 2), two exam years (Table 3), and two question types: clinical case-based and theoretical knowledge-based (Table 4).

The study evaluated ChatGPT-5 and ChatGPT-4o (OpenAI, San Francisco, CA, USA), Gemini 2.5 Pro and Gemini 2.5 Flash (Google DeepMind, Mountain View, CA, USA), and Claude Opus 4 and Claude Sonnet 4

(Anthropic, San Francisco, CA, USA). All evaluations were performed in October 2025 using the official web interfaces of the respective providers. At the time of evaluation, ChatGPT-5, Gemini 2.5 Pro, and Claude Opus 4 were accessible through paid subscriptions, whereas ChatGPT-4o, Gemini 2.5 Flash, and Claude Sonnet 4 were available free of charge. The evaluated AI chatbots (OpenAI ChatGPT, Google Gemini, and Anthropic Claude) were selected because they were among the most widely used and publicly accessible proprietary large language models at the time of the study [19]. Developed by leading AI companies, these models have been extensively investigated in the medical and dental literature for applications including clinical reasoning, medical education, and healthcare-related question answering [16, 18, 20–22]. They were also chosen because they can be accessed directly through standard web interfaces without requiring specialized technical expertise or dedicated computing infrastructure, making them readily available to dental students, educators, and clinicians in routine educational and clinical practice. Open-source models were not included because they generally require additional technical setup for deployment. Other commercial AI assistants were excluded because they typically integrate multiple underlying language models rather than representing independent proprietary foundation models. Restricting the comparison to ChatGPT, Gemini, and Claude therefore enabled a direct comparison of three widely accessible, representative proprietary foundation models.

A total of 127 multiple-choice questions were posed to all chatbots by a single researcher on October 4, 2025. New user accounts were created for each chatbot. A separate new chat session was initiated for each question to prevent responses from being influenced by the previous conversation. All DUS questions were manually entered into each AI chatbot in their original Turkish language exactly as published by ÖSYM. No role assignment (e.g., "You are an expert in pediatric dentistry") or additional contextual instructions were provided. Each question was entered into a new conversation together with the standardized instruction (translated here for clarity), "Read the following question and select only the correct option." No follow-up prompts or clarifications were used. For multimodal questions containing clinical photographs or radiographs, the respective image was uploaded directly alongside the text prompt without any additional accompanying descriptions, explanations, or contextual hints. The standardized prompt template is provided in Supplementary File 1.

Table 1 Distribution and comparison of the answers given by chatbots to the Pediatric Dentistry questions asked in the DUS exam

Model	Correct	Incorrect	Percentage	*p* value
ChatGPT-5	113	14	88.9[ab]	***p* = 0.012***
ChatGPT-4o	111	16	87.4[ab]	
Gemini 2.5 Pro	119	8	93.7[a]	
Gemini 2.5 Flash	111	16	87.4[ab]	
Claude Opus 4	113	14	88.9[ab]	
Claude Sonnet 4	106	21	83.4[b]	

Cochran's Q test. * There is a statistically significant difference. ($p < 0.05$). McNemar test was used for pairwise comparisons. (adjusted Bonferroni $p < 0.0033$). Different superscript letters indicate statistically significant differences between groups

Statistical analysis

Statistical analyses were performed using IBM SPSS 23 (IBM SPSS, Chicago, IL, USA). Descriptive statistics of the data were presented as numbers and percentages. Cochran's Q test was used to compare the performance of chatbots in answering questions correctly. If a significant difference was found, pairwise comparisons were performed using the McNemar test with a Bonferroni correction. The significance level was accepted as $p < 0.05$.

A post-hoc power analysis was performed for the Cochran's Q tests. Because the six AI models answered the same questions, the data were treated as paired binary observations. Using five degrees of freedom, a two-sided Type I error level of 0.05, the observed Cochran's Q statistics, and the corresponding noncentral chi-square distributions, achieved power $(1 - \beta)$ was calculated. The achieved power was found 99% for all questions ($n = 127$). In the subgroups, 86.9% for the clinical case questions ($n = 21$) and 81.5% for the questions from the 2016–2021 period ($n = 59$). Both subgroups values exceeded the conventionally accepted 80% threshold, indicating that the available numbers of questions were sufficient to detect the observed overall performance differences among the six models in these subgroups. Additionally, an empirical later power estimate could not be obtained for the local anesthesia subgroup in which all questions were answered correctly by all six models, as no response variability or mismatched pairs were found.

Results

Chatbots were asked 127 multiple-choice questions about pediatric dentistry. The distribution and comparison of the answers given by the chatbots to the pediatric dentistry questions asked in the DUS are shown in Table 1. Accordingly, Gemini 2.5 Pro answered 93.7% of the questions correctly; ChatGPT-5 and Claude Opus 4 answered 88.9%, ChatGPT 4 and Gemini 2.5 Flash answered 87.4%, and Claude Sonnet 4 answered 83.4% correctly. A statistically significant difference was observed in the chatbots' performance on correctly answering the 127 pediatric dentistry questions ($p = 0.012$). In pairwise comparisons to determine which groups differed significantly, it was found that the significant difference originated between

Claude Sonnet 4 and Gemini 2.5 Pro ($p = 0.001$; adjusted Bonferroni $p < 0.0033$). Accordingly, it was determined that Gemini 2.5 Pro's performance in correctly answering questions was significantly higher than Claude Sonnet 4's.

The 127 questions asked of the chatbots were grouped into 9 subject-matter categories. Gemini 2.5 Pro correctly answered all questions in the categories of "Endodontic treatments in primary and young permanent teeth", "Traumatic dental injuries in primary and permanent teeth", and "Restorative materials and pit and fissure sealants used in children"; ChatGPT-5, ChatGPT-4o, and Claude Sonnet 4 correctly answered all questions in the category of "Psychological approach to child patients, behavioral guidance techniques, sedation, and general anesthesia"; and Gemini 2.5 Flash and Claude Opus 4 correctly answered all questions in the category of "Endodontic treatments in primary and young permanent teeth". Furthermore, all chatbots correctly answered the questions in the "Local anesthesia and drug use in pediatric dentistry" category. No statistically significant difference was found in the chatbots' correct-answer performance across any subject matter ($p > 0.05$) (Table 2).

The questions asked of the chatbots were divided into two categories by year. While there was no statistically significant difference in the performance of chatbots in the exams conducted in 2015 and earlier, a difference was detected in the performance of chatbots in the exams conducted in 2016 and later ($p = 0.021$); however, pairwise comparisons revealed that the difference was not statistically significant (adjusted Bonferroni $p < 0.0033$) (Table 3).

The questions asked of the chatbots were divided into two categories: clinical case questions and theoretical

Table 2 Analysis of the performance of chatbots in correctly answering Pediatric Dentistry questions asked in the DUS exam, categorized by subject

	ChatGPT-5	ChatGPT-4o	Gemini 2.5 Pro	Gemini 2.5 Flash	Claude Opus 4	Claude Sonnet 4	*p* value
	Correct/Total	Correct/Total	Correct/Total	Correct/Total	Correct/Total	Correct/Total	
Endodontic treatments in primary and young permanent teeth	10/11	9/11	11/11	11/11	11/11	10/11	0.283
The embryology of primary and permanent teeth, the morphological and histological characteristics of primary teeth, the developmental anomalies of teeth, and the eruption of primary and permanent teeth	18/19	17/19	18/19	16/19	15/19	15/19	0.224
Psychological approach to the pediatric patient, behavioral guidance techniques, sedation, and general anesthesia.	12/12	12/12	11/12	11/12	11/12	12/12	0.416
Childhood diseases, oral manifestations of systemic diseases in children, oral lesions, developmental anomalies, and periodontal diseases in children.	24/26	23/26	24/26	23/26	24/26	22/26	0.352
Traumatic dental injuries in primary and permanent teeth	14/15	13/15	15/15	13/15	13/15	12/15	0.487
Restorative materials used in children and pit and fissure sealants	12/14	12/14	14/14	13/14	13/14	13/14	0.416
Examination and treatment planning in pediatric dentistry, occlusal relationships in children, occlusal guidance, and space maintainers	9/12	10/12	10/12	9/12	9/12	7/12	0.480
Local anesthesia and drug use in pediatric dentistry.	7/7	7/7	7/7	7/7	7/7	7/7	-
Etiology, spread, classification, and diagnosis of caries in primary dentition, methods for determining caries risk and caries activity, and fluoride in caries prophylaxis	7/11	8/11	9/11	8/11	10/11	7/11	0.286

Cochran's Q test

Statistical significance was set at $p < 0.05$

Table 3 Comparison of the responses of chatbots to the Pediatric Dentistry questions asked in the DUS exam, by year

	ChatGPT-5	ChatGPT-4o	Gemini 2.5 Pro	Gemini 2.5 Flash	Claude Opus 4	Claude Sonnet 4
	Correct/Total	Correct/Total	Correct/Total	Correct/Total	Correct/Total	Correct/Total
2012–2015	60/68	61/68	64/68	60/68	59/68	55/68
p value	0.300					
2016–2021	53/59	50/59	55/59	52/59	54/59	50/59
p value	**0.021***					

Cochran's Q test, * Although group comparisons using the Cochran's Q test was significant ($p < 0.05$), no significant differences were detected in pairwise comparisons (adjusted Bonferroni $p > 0.0033$)

Table 4 Comparison of the answers given by chatbots to clinical case and theory-based questions

	ChatGPT-5	ChatGPT-4o	Gemini 2.5 Pro	Gemini 2.5 Flash	Claude Opus 4	Claude Sonnet 4
	Correct/Total	Correct/Total	Correct/Total	Correct/Total	Correct/Total	Correct/Total
Clinical case questions	20/21	19/21	20/21	16/21	20/21	16/21
p value	**0.010***					
Theoretical knowledge questions	93/106	92/106	99/106	95/106	93/106	89/106
p value	0.083					

Cochran's Q test, * Although group comparisons using the Cochran's Q test was significant ($p < 0.05$), no significant differences were detected in pairwise comparisons (adjusted Bonferroni $p > 0.0033$)

questions. While there was no statistically significant difference in the chatbots' correct-answering performance on the theoretical questions ($p > 0.050$), a statistically significant difference was found in their correct-answering performance on the clinical case questions ($p = 0.010$). Although ChatGPT-5, Gemini 2.5 Pro, and Claude Opus 4 answered more questions correctly, no statistically significant difference was found in pairwise comparisons (adjusted Bonferroni $p > 0.0033$) (Table 4).

Discussion

Technological advancements and human feedback continually update AI-based chatbots and are widely used across fields, including healthcare [23]. However, the accuracy of responses produced by these robots is increasingly being examined in scientific research. Although the primary goal in designing these robots is not to answer questions in medicine and dentistry [2], studies have shown they perform well in these areas [5, 24]. In addition, many studies examine the performance of chatbots on national centralized examinations in medicine and dentistry [25–27]. To our knowledge, studies examining the performance of chatbots in pediatric dentistry on the Turkish Dental Specialization Exam are scarce in the literature. Moreover, as technology advances, comprehensive evaluations of different models and new versions are needed. Therefore, in this study, the performance of ChatGPT-5, ChatGPT-4o, Gemini 2.5 Pro, Gemini 2.5 Flash, Claude Opus 4, and Claude Sonnet 4 AI chatbots in answering questions in pediatric dentistry on the Turkish Dental Specialization Exam was comparatively evaluated. According to the results of this study, Gemini 2.5 Pro showed the highest performance, followed by ChatGPT-5 and Claude Opus 4. In contrast, ChatGPT-4o and Gemini 2.5 Flash showed lower performance, and Claude Sonnet 4 had the lowest accuracy. A significant difference was found between Gemini 2.5 Pro and Claude Sonnet 4, but the difference between the other chatbots was not significant. All chatbots achieved an accuracy level above 80%. The findings show that AI chatbots perform well at answering exam questions in pediatric dentistry.

Numerous studies have examined the DUS performance of AI chatbots, either in general or within a specific area of expertise. One study showed that ChatGPT-4o and Gemini Advanced achieved over 80% overall accuracy on all questions asked between 2012 and 2021 [28]. In the DUS questions in the field of endodontics, ChatGPT-4o, Gemini Advanced, and Claude achieved more than 80% accuracy [17]; in the field of prosthetic dentistry, ChatGPT 4 reached 75.8%, ChatGPT 3.5 74.21%, Claude AI 62.5%, Gemini Advanced 54.68%, and Gemini AI 46.1% accuracy [29]. However, a study using earlier versions of chatbots that reported lower accuracy (54.3%) is also reported in the literature [30]. In the current study, Gemini 2.5 Pro answered 93.7% of the questions correctly; ChatGPT-5 and Claude Opus 4 answered 88.9%, ChatGPT-4o and Gemini 2.5 Flash answered 87.4%, and Claude Sonnet answered 83.4%. It is believed that high accuracy was achieved in this study due to the use of the most up-to-date chatbot versions. When the results obtained are evaluated alongside the literature, it is evident that different artificial intelligence applications exhibit varying levels of success depending on the field of expertise. These differences between fields may stem from the heterogeneity of dental specialties or from the degree of representation of specialty-specific terminology and concepts in the robots' training datasets [31]. Accordingly, the findings suggest that a single artificial

intelligence program may not be equally effective across all specialties.

The literature contains studies reporting that chatbots can outperform one another in answering DUS questions [15, 28, 29] or exhibit similar performance [17, 32, 33]. ChatGPT 01 showed significantly higher accuracy than Gemini 1.5, Gemini 2, ChatGPT-4o, and ChatGPT 4 in answering DUS questions in the field of oral pathology [15]; while in the field of anatomy, ChatGPT-4o has higher accuracy than Gemini and Claude [34]. There are also studies reporting no significant difference between ChatGPT-4o, Gemini Advanced, and Claude in answering DUS questions in the field of endodontics [17]; between ChatGPT-5 and Gemini Pro in the field of oral, dental, and maxillofacial surgery [32], and between ChatGPT-4o Plus and Gemini in the field of pediatric dentistry [33]. Sismanoglu and Capan reported that ChatGPT 4.0 and Gemini Advanced showed similar performance in answering pediatric dentistry questions in the DUS exam in 2020 and 2021. Still, Gemini Advanced performed better when comparing performance across all questions [35]. In their study, Dündar Sarı and Sezer reported that Gemini Advanced showed higher accuracy than ChatGPT-4o in pediatric dentistry DUS questions. Still, ChatGPT-4o showed superior performance in prosthetic dentistry and maxillofacial radiology, and they found discipline-specific differences in the chatbots' performance [28]. In the present study, Gemini 2.5 Pro showed significantly higher accuracy than Claude Sonnet 4 on DUS questions in pediatric dentistry, but there were no significant differences among the other chatbots. The fact that chatbots excel in different areas of expertise can be explained by the methods they use to access, structure, and reason about information, which provide different advantages depending on the area of expertise.

The performance of chatbots on specific subjects within pediatric dentistry has been investigated in several studies. In the Korean National Dental Board Examination, ChatGPT 3.5 achieved the highest accuracy on the subjects of "pulp treatments" and "trauma" among pediatric dentistry questions. In contrast, Gemini showed higher performance on the subjects "tooth development and morphology" and "occlusion/orthodontics" than on other subjects [25]. Sezer and Okutan reported that ChatGPT 4 achieved the highest performance on curriculum questions in pediatric dentistry on the topic of "fissure sealants." At the same time, it showed the lowest performance on "fluoride" and "pulp treatments" [36]. In Meriç's study, ChatGPT-4o Plus achieved the highest accuracy on DUS questions in pediatric dentistry, specifically the topics "histological and morphological characteristics of dental tissues" and "behavioral guidance." In contrast, Gemini showed the highest accuracy for the subjects "histological and morphological characteristics of dental tissues" and "drug use in pedodontics". ChatGPT-4o Plus was least successful on the subject of "fluorides," while Gemini was least successful on the subject of "local anesthesia" [33]. No significant difference was found in the performance of ChatGPT-4o Plus and Gemini across subjects [33]. In the present study, Gemini 2.5 Pro correctly answered all questions in the areas of "Endodontic treatments in primary and young permanent teeth", "Traumatic dental injuries in primary and permanent teeth", and "Restorative materials and pit and fissure sealants used in children"; ChatGPT-5, ChatGPT-4o, and Claude Sonnet 4 correctly answered all questions in the area of "Psychological approach to the child patient, behavioral guidance techniques, sedation, and general anesthesia"; and Gemini 2.5 Flash and Claude Opus 4 correctly answered all questions in the area of "Endodontic treatments in primary and young permanent teeth". Furthermore, all chatbots correctly answered the questions in the area of "Local anesthesia and drug use in pediatric dentistry". However, no significant difference was found in the accuracy performance of the chatbots across different subjects in this study. This result can be explained by the fact that the overall success rates of the chatbots were high and similar. The models having similarly large knowledge bases and the questions mostly being based on standard guide information may have contributed to the lack of significant differences across subjects.

A study comparing the DUS performance of chatbots year by year found no significant difference between ChatGPT-4o and Gemini Advanced in answering all DUS questions [28]. In the field of oral, dental, and maxillofacial radiology, only in the 2013 DUS questions was Gemini reported to exhibit significantly lower accuracy than ChatGPT-4o and ChatGPT-4o mini [37]. Icen and Temur found that ChatGPT-5.2 performed significantly better than ChatGPT 4.0 in both the early (2012–2018) and late (2019–2025) periods of the DUS exam in oral, maxillofacial, and facial radiology [38]. In the present study, the performance of chatbots was found to be similar in exams from 2015 and earlier; although a performance difference was observed between the robots in exams from 2016 onwards, pairwise comparisons determined that the difference was not statistically significant. The general similarity in difficulty level and question structure of the exams administered in different years may have limited the emergence of a significant difference in the robots' performance. In addition, exams conducted from 2022 onwards could not be included in the study due to the fact that the questions were not published by OSYM. Therefore, the content characteristics of more recent exams and the performance of new generation chatbots in these exams could not be evaluated within the scope of this study.

Previous studies have indicated that case-based questions, which require higher-level cognitive skills, contextual understanding, and clinical judgment compared to theoretical knowledge questions, pose greater challenges for AI-powered chatbots [29, 39]. In DUS questions measuring theoretical knowledge in oral pathology, ChatGPT o1 performed better than Gemini 1.5, Gemini 2, Claude, ChatGPT-4o, and ChatGPT 4; while in case questions, ChatGPT o1 and Claude performed better than ChatGPT 4. Therefore, it is emphasized that these models may possess more advanced skills in interpreting clinical scenarios and reasoning [15]. In the field of prosthetic dentistry, ChatGPT 4 achieved high performance on theoretical knowledge questions in the DUS exam. At the same time, Gemini AI performed poorly on theoretical knowledge questions, and both Gemini AI and Gemini Advanced performed poorly on case questions [29]. In a study examining the performance of chatbots (ChatGPT 3.5, ChatGPT-4o Free, ChatGPT-4o Plus, Claude Sonnet, Gemini 1.5, Gemini Advanced) in the field of restorative dentistry on the DUS, Gemini Advanced achieved the highest accuracy in theoretical knowledge questions. In contrast, ChatGPT-4o Plus, Gemini 1.5, and Gemini Advanced achieved excellent accuracy on case-based questions, demonstrating a superior ability to handle complex clinical scenarios and case analyses [40]. In the present study, the chatbots showed similar performance in correctly answering theoretical questions. However, in clinical case questions, although ChatGPT 5, Gemini 2.5 Pro, and Claude Opus 4 achieved higher correct answers, the difference was not statistically significant. The fact that the paid versions had higher correct answers in clinical case questions but the difference was not statistically significant may be related to the fact that all current versions of chatbots generally show similar and high success rates. The fact that no significant differences were found between models in questions based on theoretical knowledge indicates, as also stated by Kung et al., that comprehensive and database-driven training processes for chatbots can provide a significant advantage in answering questions based on theoretical knowledge [41].

When the results of the current study are evaluated alongside the literature, the success of chatbots in solving exam questions in medicine and dentistry [5, 21] demonstrates the potential of artificial intelligence for integration into educational processes in these fields [41]. Chatbots can improve case-based learning in dental education by creating medical notes, compiling patient information, and supporting diagnosis and treatment planning [42]. These capabilities create a hands-on learning environment by presenting students with virtual case simulations and a range of clinical scenarios. Thus, they can be considered an important tool in preparing students for real clinical challenges. However, it is important to note that the answers generated by these chatbots can change depending on the question asked and previous conversations [43]. This phenomenon, described as a kind of "hallucination phenomenon," has led to discussions about how chatbots can give persuasive, repetitive, but incorrect answers [25]. This raises concerns about the risk of artificial intelligence providing incorrect medical advice in healthcare. Therefore, AI systems should be used as supplementary tools in education and clinical practice, rather than as primary methods.

The sample size of 127 questions yielded a post-hoc power of 99%, sufficient for primary analysis. Although the clinical case subgroup included only 21 questions, the post-hoc power of the omnibus Cochran's Q test was 86.9%, exceeding the conventional 80% threshold. Similarly, the analysis of questions from the 2016–2021 period achieved a power of 81.5%. Thus, the available sample sizes were adequate to detect the observed overall differences among the six AI models in these two subgroups. However, a limitation of this study is that the six-model design required 15 pairwise McNemar comparisons, and the Bonferroni adjustment reduced the pairwise significance threshold to $\alpha = 0.0033$. Consequently, the statistical power of the pairwise comparisons was lower than that of the omnibus tests, and the absence of significant pairwise findings should not be interpreted as evidence of equivalent performance. The local anaesthesia category comprised only seven questions, and all six models answered all seven questions correctly. Because no response variability or discordant pairs were present, an empirical post-hoc power estimate could not be obtained for this category; these findings should therefore be interpreted descriptively and with caution. Another limitation of this study is that only publicly available DUS questions from 2012 to 2021 could be included because more recent examinations are not accessible. Therefore, the findings may not fully reflect the performance of current chatbot models on contemporary examination content. Furthermore, since these questions have been publicly available for several years, the high accuracy observed in this study should be interpreted with caution due to the possibility of data leakage. Since only Turkish DUS questions aligned with the Turkish dental curriculum were used in this study, the findings may vary across different education systems and languages. Finally, this study evaluated artificial intelligence models accessible as of October 2025. Given the constantly evolving nature of artificial intelligence systems, long-term, iterative evaluation studies are needed to monitor performance changes over time.

Despite these limitations, this study's comparative examination of multiple chatbots' performance in pediatric dentistry contributes to a comprehensive understanding of the relative success rates of current artificial

intelligence systems. Furthermore, the inclusion of both paid and free versions of the chatbots enabled comparisons between models and the evaluation of the academic performance of versions at different access levels. A significant strength of the DUS questions used in the study is their multiple-choice and structured assessment format, as well as their preparation based on standard curricula and current guidelines. Although the dataset consisted primarily of text-based multiple-choice questions, it also included a limited number of image-based questions. While these questions represented only a small proportion of the dataset, their inclusion helped the evaluation more faithfully reflect the original examination content. Analyzing the questions by year, subject matter, and clinical case- and theoretical knowledge-based categories enabled a detailed evaluation of the chatbots' performance across different content areas and cognitive levels. In these respects, the study makes a unique and multifaceted contribution to both the pediatric dentistry literature and the potential applications of artificial intelligence in dental education.

Conclusions

In conclusion, this study demonstrates that chatbots can answer pediatric dentistry questions with high accuracy and serve as an effective educational tool in this field. However, these systems should not be considered an absolute source of information; their responses should be critically evaluated and used under human supervision. While they may help provide quick access to information and summarize guidelines in clinical practice, they should be viewed as complementary tools that support clinical judgment and evidence-based decision-making, rather than replacing them.

Abbreviations

AI	Artificial intelligence
DUS	Dental Specialization Exam
ÖSYM	Student Selection and Placement Center

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12903-026-09371-7.

Supplementary Material 1.

Acknowledgements

Not applicable.

Author's contributions

S. SAKIN ULUBAY conceptualized the study, developed the methodology, conducted the investigation, performed formal analyses and drafted and reviewed the manuscript.

Funding

This study has been supported by the Recep Tayyip Erdoğan University Development Foundation (Grant number: 020260070130482).

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 8 June 2026 / Accepted: 16 July 2026
Published online: 21 July 2026

References

1. Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. Int J Oral Sci. 2023;15(1):29.
2. 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(1):22.
3. 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(6):269–74.
4. Raiaan MAK, Mukta MSH, Fatema K, Fahad NM, Sakib S, Mim MMJ, et al. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. IEEE access. 2024;12:26839–74.
5. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312.
6. Krishna S, Bhambra N, Bleakney R, Bhayana R. Evaluation of reliability, repeatability, robustness, and confidence of GPT-3.5 and GPT-4 on a radiology board–style examination. Radiology. 2024;311(2):e232715.
7. 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(7):1098–102.
8. Brozović J, Mikulić B, Tomas M, Juzbašić M, Blašković M. Assessing the performance of Bing Chat artificial intelligence: dental exams, clinical guidelines, and patients' frequent questions. J Dent. 2024;144:104927.
9. 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.
10. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ. 2023;9(1):e48002.
11. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80.
12. Tussie C, Starosta A. Comparing the dental knowledge of large language models. Br Dent J. 2024:1–3. https://doi.org/10.1038/s41415-024-8015-2.
13. Zaidi NLB, Grob KL, Monrad SM, Kurtz JB, Tai A, Ahmed AZ, et al. Pushing Critical Thinking Skills With Multiple-Choice Questions: Does Bloom's Taxonomy Work? Acad Med. 2018;93(6):856–9.
14. Gonsalves C. On ChatGPT: what promise remains for multiple choice assessment? J Learn Dev High Educ. 2023;(27). https://doi.org/10.47408/jldhe.vi27.1009.
15. Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F. Artificial intelligence performance in answering multiple-choice oral pathology questions: a comparative analysis. BMC Oral Health. 2025;25(1):573.
16. Karatekin AO. Comparison and review of different versions of openai chatgpt, anthropic claude and google gemini large language models' performance on endodontics questions in the turkish dentistry specialization exam. Eur J Dent Educ. 2026. https://doi.org/10.1111/eje.70121.
17. Turan Gökduman C, Arılı Öztürk E, Aktaş Ş, Çanakçi BC. Comparison of chatbots' accuracy in endodontics questions in dentistry specialization exam

in Türkiye: ChatGPT-4o, Gemini Advanced, Copilot, and Claude. BMC Oral Health. 2025;26(1):28.
18. 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.
19. Similarweb. Top AI tools: Most used Gen-AI in August 2025. 2025.
20. Akitomo T, Hamada M, Hamaguchi S, Kusaka S, Nomura R. Generative Artificial Intelligence and Large Language Models in Paediatric Dentistry: A Scoping Review. Int Dent J. 2026;76(2):109401.
21. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne). 2024;11:1477898.
22. Younis HA, Eisa TAE, Nasser M, Sahib TM, Noor AA, Alyasiri OM et al. A Systematic Review and Meta-Analysis of Artificial Intelligence Tools in Medicine and Healthcare: Applications, Considerations, Limitations, Motivation and Challenges. Diagnostics (Basel). 2024;14(1):109.
23. Arsenijevic U, Jovic M, editors. Artificial intelligence marketing: chatbots. 2019 international conference on artificial intelligence: applications and innovations (IC-AIAI); IEEE. 2019.
24. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination? a descriptive study. J Educ Eval Health Prof. 2023;20:1.
25. Jung YS, Chae YK, Kim MS, Lee H-S, 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.
26. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, Arteaga-Cisneros KF, Chalco XCB, Ordoñez MAB et al. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian national licensing medical examination: a cross-sectional study. J Educational Evaluation Health Professions. 2023;20:30.
27. Uehara O, Morikawa T, Harada F, Sugiyama N, Matsuki Y, Hiraki D, et al. Performance of ChatGPT-3.5 and ChatGPT-4o in the Japanese national dental examination. J Dent Educ. 2025;89(4):459–66.
28. Dundar Sari MB, Sezer B. Comparative performance evaluation of ChatGPT-4 Omni and Gemini Advanced in the Turkish Dentistry Specialization Exam. BMC Med Educ. 2026;26(1):251.
29. Tosun B, Yilmaz ZS. Comparison of artificial intelligence systems in answering prosthodontics questions from the dental specialty exam in Turkey. J Dent Sci. 2025;20(3):1454–9.
30. Aşık A, Kuru E. Diş Hekimliğinde Uzmanlık Eğitim Giriş Sınavında Sorulan Çocuk Diş Hekimliği Sorularına ChatGPT'nin Verdiği Cevapların Analizi: Kesitsel Araştırma. Turkiye Klinikleri J Dent Sci. 2025;31(3):401–6.
31. Xiong YT, Zhan ZZ, Zhong CL, Zeng W, Guo JX, Tang W, et al. Evaluating the performance of large language models (LLMs) in answering and analysing the Chinese dental licensing examination. Eur J Dent Educ. 2025;29(2):332–40.
32. Çetiner EY. Comparative Evaluation of ChatGPT-5 and Gemini 2.5 Pro in Answering Oral and Maxillofacial Surgery Questions from Dentistry Specialization Exams: A Cross-Sectional Study. EurAsian J Oral Maxillofacial Surg. 2025;4(3):59–65.
33. Meriç E. Diş Hekimliğinde Uzmanlık Sınavında Sorulan Çocuk Diş Hekimliği Sorularını Cevaplamada ChatGPT-4o Plus ve Gemini'nin Karşılaştırmalı Analizi: Kesitsel Çalışma. Turkiye Klinikleri J Dent Sci. 2025;31(4):401–7.
34. Tassoker M. Who Knows Anatomy Best? A Comparative Study of ChatGPT-4o, DeepSeek, Gemini, and Claude. Clin Anat. 2026;39(1):25–9.
35. Sismanoglu S, Capan BS. Performance of artificial intelligence on Turkish dental specialization exam: can ChatGPT-4.0 and gemini advanced achieve comparable results to humans? BMC Med Educ. 2025;25(1):214.
36. Sezer B, Okutan AE. Evaluation of ChatGPT-4's performance on pediatric dentistry questions: accuracy and completeness analysis. BMC Oral Health. 2025;25(1):1427.
37. Peker RB, ChatGPT CGPT, Plus. Gemini ve Microsoft Copilot Büyük Dil Modellerinin Türkiye'deki Diş Hekimliği Uzmanlık Eğitimi Giriş Sınavı'nda Sorulan Ağız, Diş ve Çene Radyolojisi Sorularını Cevaplama Performanslarının Değerlendirilmesi. 7tepe Klinik Dergisi. 2024;21(3):130–5.
38. Icen M, Temur KT. Comparative diagnostic accuracy of multiple large language models in oral and maxillofacial radiology specialty examinations: a 13-year analysis of performance and topic trends. BMC Oral Health. 2026;26(1):822. https://doi.org/10.1186/s12903-026-08236-3.
39. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J stomatology oral maxillofacial Surg. 2023;124(5):101471.
40. Haberal M, Hançerlioğulları D. Can artificial intelligence chatbots think like dentists? A comparative analysis based on dental specialty examination questions in restorative dentistry. BMC Oral Health. 2026;26(1):231.
41. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.
42. Alser M, Waisberg E. Concerns with the usage of ChatGPT in academia and medicine: a viewpoint. Am J Med Open. 2023;9:100036.
43. Avşar DB, Ertan AA. Diş hekimliğinde uzmanlık sınavında sorulan protetik diş tedavisi sorularının ChatGPT-3.5 ve Gemini tarafından cevaplanma performanslarının karşılaştırmalı olarak incelenmesi: Kesitsel araştırma. Turkiye Klinikleri Dishekimligi Bilimleri Dergisi. 2024;30(4):668–73.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.