

Evaluating the accuracy of ChatGPT and Gemini models in dental specialty exams: A comparative study of pediatric and restorative dentistry

Ayşenur Altuğ Yıldırım¹, Menekşe Alim²

¹ Çankırı Karatekin University, Faculty of Dentistry, Department of Restorative Dentistry, Çankırı, Türkiye

² Private Practice, Ankara, Türkiye

Received: 8 May 2026

Revised: 10 June 2026

Accepted: 20 June 2026

Published: 28 June 2026

Correspondence:

Dr. Ayşenur ALTUĞ YILDIRIM

Çankırı Karatekin University, Faculty of
Dentistry, Department of Restorative
Dentistry, Çankırı, Türkiye.

E-mail: aysenuraltug@karatekin.edu.tr



How to cite this article:

Altuğ Yıldırım A, Alim M. Evaluating the accuracy of ChatGPT and Gemini models in dental specialty exams: A comparative study of pediatric and restorative dentistry. *Int Dent Res*. 2026;16(S1):e260673. <https://doi.org/10.5577/intdentres.673>

Abstract

Aim: The integration of large language models (LLMs) such as ChatGPT and Gemini into healthcare education presents novel opportunities for knowledge dissemination and assessment. Despite their growing prominence, empirical evidence regarding their reliability and domain-specific accuracy in dental specialty examinations remains limited. This study aimed to evaluate and compare the performance of four AI-driven LLMs—ChatGPT-3.5, ChatGPT-4o, Gemini 2.0 Flash, and Gemini 2.0 Advanced—by employing official past questions from the Turkish Dental Specialty Examination (DUS) in Pediatric Dentistry and Restorative Dentistry.

Methodology: A total of 178 multiple-choice questions (90 in Restorative Dentistry and 88 in Pediatric Dentistry) were obtained from DUS examinations conducted between 2012 and 2017. All items were presented in Turkish to each model via standardized prompts in isolated sessions to prevent cross-interaction and memory bias. Response accuracy was determined by evaluating the correctness of the chosen answers. Comparative analyses between models, domains, and annual trends were conducted using non-parametric statistical tests, including the Mann-Whitney U and Friedman tests.

Results: ChatGPT-4o achieved the highest overall accuracy across both Pediatric (90.88%) and Restorative Dentistry (93.02%), whereas ChatGPT-3.5 exhibited the lowest performance in Pediatric Dentistry (70.71%). Statistically significant inter-model differences were observed in Pediatric Dentistry ($p = 0.002$), but not in Restorative Dentistry ($p = 0.254$). Temporal analyses highlighted the consistent performance of ChatGPT-4o, while Gemini 2.0 Flash displayed marked fluctuations in accuracy across certain years and subdomains.

Conclusion: Among the evaluated LLMs, ChatGPT-4o demonstrated superior and stable performance in addressing dental specialty exam questions, particularly in content requiring contextual comprehension. These findings underscore the potential of advanced AI tools to enhance dental education and examination readiness. Nonetheless, appropriate model selection should be tailored to the complexity of the task, domain specificity, and pedagogical objectives.

Keywords: Large language models, dental specialty examination, ChatGPT, Gemini, dental education, artificial intelligence

Introduction

The integration of artificial intelligence (AI) into healthcare has precipitated significant advancements across various disciplines, including dentistry (1-3). Recent investigations into large language models (LLMs) have underscored their transformative potential in dental education, clinical reasoning, and scientific inquiry (4-8). Trained on extensive, diverse textual corpora, these models can enhance contextual comprehension and have exhibited promising performance, particularly in the medical and dental domains (9). In dentistry, AI systems contribute meaningfully to diagnostic decision-making, treatment planning, radiographic interpretation, outcome prediction, and administrative optimization through the management of digital patient records and clinical workflows (10).

In dental education, LLMs have been explored as interactive learning tools that can enrich educational experiences and foster active student engagement. They have been shown to support students' understanding of complex concepts, facilitate learning through simulations, and serve as virtual teaching assistants (11-13). Moreover, they offer considerable utility in patient education by delivering accessible explanations regarding oral health conditions, preventive measures, and treatment protocols (14).

However, LLMs may occasionally generate inaccurate or contextually inappropriate information, which poses a risk to reliability—particularly in healthcare settings (8, 15). To ensure clinical relevance and safety, it is therefore imperative that such models be fine-tuned on domain-specific resources, including dental curricula, clinical guidelines, and peer-reviewed scientific literature (16).

Among the current generation of AI-driven LLMs, OpenAI's ChatGPT and Google's Gemini have emerged as prominent tools. ChatGPT-3.5, built on the GPT-3 architecture, is optimized for prompt and coherent responses to fact-based queries (OpenAI, 2023) (17-19). ChatGPT-4o ("omni"), released in 2024, integrates multimodal capabilities—processing text, image, and audio inputs—and demonstrates superior performance on tasks requiring high-level cognitive processing (OpenAI, 2024). In parallel, Google DeepMind's Gemini 2.0 Flash is engineered for real-time responsiveness, while its more advanced counterpart, Gemini 2.0 Advanced, is designed for computationally intensive tasks necessitating deeper linguistic and contextual understanding (Google DeepMind, 2024a; 2024b). While ChatGPT-3.5 and Gemini 2.0 Flash are publicly accessible, ChatGPT-4o and Gemini 2.0 Advanced are subscription-based.

These models have been evaluated across various medical and dental examinations and have delivered promising results in knowledge-based assessments (20, 21). For instance, ChatGPT-4 successfully passed the written components of dental licensing examinations in the United States and the United Kingdom, whereas ChatGPT-3.5 fell short of the passing threshold (22). In

the U.S. Integrated National Board Dental Examination (INBDE), both models demonstrated high accuracy, correctly answering approximately 80% of knowledge-based items (23). Similarly, in the 2023 Japanese National Dental Examination, ChatGPT-4 attained the highest composite score, followed by Bard and ChatGPT-3.5 (24).

The present study aims to evaluate and compare the response accuracy and reliability of four large language models—ChatGPT-3.5, ChatGPT-4o, Gemini 2.0 Flash, and Gemini 2.0 Advanced—when presented with official questions from the Turkish Dental Specialty Examination (DUS) in Restorative and Pediatric Dentistry.

- H₁: There is a statistically significant difference in the success rates of each AI model between the Pediatric Dentistry and Restorative Dentistry domains.
- H₂: There is a statistically significant difference in the success rates of AI models, specifically in Pediatric Dentistry.
- H₃: There is a statistically significant difference in the success rates of AI models, specifically in Restorative Dentistry.

Materials and Methods

The Dental Specialty Examination (DUS) is a national standardized examination administered in Türkiye for graduates of dental faculties seeking admission to postgraduate specialty training programs. Organized by the Assessment, Selection, and Placement Center (ÖSYM), the DUS assesses candidates' theoretical knowledge through a multiple-choice format and comprises two main sections: basic sciences and clinical sciences. Due to its emphasis on evidence-based knowledge, clinical reasoning, and familiarity with current literature, the DUS serves as an appropriate framework for evaluating the performance of large language models (LLMs).

The examination is held either once or twice annually, depending on the year, typically during the spring and fall terms. Within the clinical sciences section, ten questions are allocated to each dental specialty, including Restorative Dentistry, Endodontics, Pediatric Dentistry, Periodontology, Prosthodontics, Oral and Maxillofacial Surgery, Oral and Maxillofacial Radiology, and Orthodontics. All items are structured as five-option multiple-choice questions.

For this study, questions from the Restorative Dentistry and Pediatric Dentistry sections were extracted from publicly available DUS examinations conducted between 2012 and 2017. A total of 90 questions in Restorative Dentistry and 88 in Pediatric Dentistry were included, with two Pediatric Dentistry questions excluded due to official cancellations. All questions were obtained from publicly available DUS (Dental Specialty Examination) booklets published by ÖSYM (Assessment, Selection, and Placement Center) and were presented to

the LLMs in their original Turkish format without modification (25).

The evaluation involved four large language models: ChatGPT-3.5 and Gemini 2.0 Flash (free-access), as well as ChatGPT-4o and Gemini 2.0 Advanced (subscription-based). All questions were submitted in Turkish using standardized prompts by a single operator. Each question was submitted using the following standardized Turkish prompt: “Aşağıdaki sorunun doğru cevabı hangi seçenekte verilmiştir?” followed by the original multiple-choice question text and answer options. The majority of the evaluated items were text-based. However, two questions in the Restorative Dentistry domain included visual content (e.g., clinical photographs). These image-based items were digitally presented to all four models and were included in the analysis. ChatGPT-3.5 and Gemini 2.0 Flash were tested from March 5th to 10th, 2025, while ChatGPT-4o and Gemini 2.0 Advanced were tested from March 12th to 15th, 2025.

To minimize potential memory-related bias, the chat session was refreshed after each question, ensuring isolated query conditions. Each item was submitted only once to prevent adaptive learning or retrospective model optimization. All model responses were documented for subsequent analysis.

Statistical Analysis

The accuracy of responses generated by the four AI models—ChatGPT-3.5, ChatGPT-4o, Gemini 2.0 Flash, and Gemini 2.0 Advanced—was statistically evaluated using SPSS version 24.0 (IBM Corp., Armonk, NY, US).

The distribution of the data was initially assessed using the Kolmogorov-Smirnov and Shapiro-Wilk tests to assess conformity with the normality assumptions. Because the data did not meet the criteria for normal distribution, nonparametric statistical methods were applied. The Mann-Whitney U test was used to compare success rates between two independent groups,

specifically to evaluate model performance across the Pediatric and Restorative Dentistry domains. A total of 178 questions (90 in Restorative Dentistry and 88 in Pediatric Dentistry) were evaluated and grouped into 39 topic categories (21 in Pediatric Dentistry and 18 in Restorative Dentistry). For each topic category, the percentage of correctly answered questions was calculated per model. All statistical analyses were conducted on these topic-level success rates as the unit of analysis.

To compare success rates across all four AI models, the Friedman test—a nonparametric alternative for repeated-measures designs—was conducted. In cases where significant differences were detected, post hoc pairwise comparisons were performed to identify the source of variation.

Descriptive statistics, including mean, standard deviation, and median, were calculated to assess temporal trends in model performance across examination years. A p-value of less than 0.05 was considered indicative of statistical significance.

Results

The average success rate of ChatGPT-3.5 in Pediatric Dentistry was 70.71%, while in Restorative Dentistry it was 85.91%. However, this difference was not statistically significant ($U = 151$, $p = 0.259$). Similarly, ChatGPT-4o achieved average success rates of 90.88% and 93.02% in Pediatric and Restorative Dentistry, respectively, with no significant difference between the two domains ($U = 183$, $p = 0.836$).

Gemini 2.0 Flash yielded success rates of 71.86% in Pediatric Dentistry and 86.30% in Restorative Dentistry, although the difference did not reach statistical significance ($U = 135.5$, $p = 0.115$). Gemini 2.0 Advanced displayed nearly equivalent performance in both domains, with success rates of 85.68% and 85.73%, respectively ($U = 184.5$, $p = 0.885$).

Table 1. Comparison of the success rates (%) of ChatGPT and Gemini models in Pediatric and Restorative Dentistry domains of the Turkish Dental Specialty Exam.

	Pediatric Dentistry			Restorative Dentistry			Mann-Whitney U Test	
	%	SD	M	%	SD	M	U	p
ChatGPT-3.5 (success rate %)	70.71	36.08	80.00	85.91	17.35	92.86	151.000	0.259
ChatGPT-4o (success rate %)	90.88	22.67	100.00	93.02	13.11	100.00	183.000	0.836
Gemini 2.0 Flash (success rate %)	71.86	29.87	75.00	86.30	19.29	100.00	135.500	0.115
Gemini 2.0 Advanced (success rate %)	85.68	25.27	100.00	85.73	22.98	100.00	184.500	0.885

SD: Standard deviation, M: Median

The success rates of the four AI models in Pediatric Dentistry and Restorative Dentistry were compared, and overall differences across both domains were analyzed using the Friedman test (Table 2).

The Friedman test was conducted to compare the overall performance of the four AI models across both domains (Table 2). In Pediatric Dentistry, ChatGPT-4o achieved the highest accuracy (90.88%), followed by

Gemini 2.0 Advanced (85.68%), Gemini 2.0 Flash (71.86%), and ChatGPT-3.5 (70.71%). In Restorative Dentistry, ChatGPT-4o again recorded the highest score (93.02%), while the other models demonstrated relatively similar performance: Gemini 2.0 Advanced (85.73%), Gemini 2.0 Flash (86.30%), and ChatGPT-3.5 (85.91%).

According to the Friedman test, the differences among AI models were statistically significant in Pediatric Dentistry ($\chi^2 = 15.176$, $p = 0.002$). Post hoc analysis indicated that ChatGPT-3.5 and Gemini 2.0 Flash performed significantly worse compared to ChatGPT-4o and Gemini 2.0 Advanced. Conversely, no statistically

significant difference was observed among the models in Restorative Dentistry ($\chi^2 = 4.066$, $p = 0.254$), indicating comparable performance in this domain. The comparative success rates of all four models across both domains are illustrated in Figure 1.

Yearly performance trends

To evaluate temporal performance trends, yearly success rates for each AI model were analyzed in both Pediatric and Restorative Dentistry (Table 3).

Table 2. Comparison of the success rates (%) of ChatGPT and Gemini models in Pediatric and Restorative Dentistry domains according to the Friedman test.

	Pediatric Dentistry			Restorative Dentistry		
	%	SD	M	%	SD	M
ChatGPT-3.5 (success rate %)	70.71	36.08	80.00	85.91	17.35	92.86
ChatGPT-4o (success rate %)	90.88	22.67	100.00	93.02	13.11	100.00
Gemini 2.0 Flash (success rate %)	71.86	29.87	75.00	86.30	19.29	100.00
Gemini 2.0 Advanced (success rate %)	85.68	25.27	100.00	85.73	22.98	100.00
Friedman Test	χ^2	15.176		4.066		
	p	0.002		0.254		
	Post-hoc	1-2, 1-4, 2-3, 3-4				

SD: Standard deviation, M: Median

Table 3. Comparison of the success rates (%) of ChatGPT and Gemini models by examination term (2012-2017) in Pediatric and Restorative Dentistry domains of the Turkish Dental Specialty Exam.

		ChatGPT-3.5 (success rate %)	ChatGPT-4o (success rate %)	Gemini 2.0 Flash (success rate %)	Gemini 2.0 Advanced (success rate %)
Pediatric Dentistry	2012.1	90.00	100.00	80.00	90.00
	2012.2	70.00	100.00	90.00	90.00
	2013.1	66.67	100.00	77.78	88.89
	2013.2	88.89	100.00	100.00	100.00
	2014.1	90.00	100.00	70.00	100.00
	2014.2	90.00	90.00	60.00	80.00
	2015	80.00	80.00	70.00	90.00
	2016	70.00	90.00	80.00	70.00
	2017	60.00	80.00	70.00	70.00
Restorative Dentistry	2012.1	80.00	80.00	80.00	90.00
	2012.2	80.00	80.00	80.00	80.00
	2013.1	90.00	100.00	90.00	80.00
	2013.2	90.00	100.00	70.00	80.00
	2014.1	90.00	100.00	100.00	100.00
	2014.2	70.00	90.00	60.00	90.00
	2015	90.00	100.00	90.00	80.00
	2016	90.00	90.00	80.00	70.00
	2017	100.00	100.00	100.00	90.00

Pediatric Dentistry

ChatGPT-4o consistently outperformed other models, achieving 100% success in five exam periods (2012.1, 2012.2, 2013.1, 2013.2, and 2014.1). In contrast, ChatGPT-3.5 showed notable variability, with its lowest performance in 2013.1 (66.67%) and peaks in 2013.2 (88.89%) and 2014.1 (90.00%). Gemini 2.0 Flash also demonstrated fluctuating results—while scoring 90.00% in 2012.2, its accuracy dropped to 60.00% in 2014.2 and hovered around 70.00% in 2015 and 2017. Gemini 2.0 Advanced maintained high scores (>90%) in the early years but declined in 2016 and 2017, reaching approximately 70.00%.

Overall, ChatGPT-4o was the most stable and accurate model in Pediatric Dentistry, while the other models displayed greater year-to-year variability. All models achieved 100% accuracy on topics such as "sedation and general anesthesia," "dental care in children with disabilities," "dentoalveolar infections," and "assessment of caries activity" (Fig. 2), indicating that these subjects likely reflect fundamental, standardized content commonly found in training datasets. In contrast, ChatGPT-3.5 scored 0% on more specialized topics such as "embryology of primary and permanent teeth" and "oral manifestations of systemic diseases," while the other models maintained higher accuracy in these areas.

Restorative Dentistry

Compared to Pediatric Dentistry, differences among models were less pronounced in Restorative Dentistry. All models generally achieved success rates above 80.00%. ChatGPT-4o achieved 100% accuracy in 2013.1, 2013.2, 2014.1, 2015, and 2017, but recorded lower performance in 2012.1 and 2012.2 (80.00%). ChatGPT-3.5 demonstrated a gradual improvement over time, culminating in 100% accuracy in 2017.

Gemini 2.0 Flash exhibited stable performance in most years but saw significant drops in 2013.2 (70.00%) and 2014.2 (60.00%). Gemini 2.0 Advanced maintained high accuracy levels, reaching 100% in 2014.1, yet its performance also declined in 2016 (70.00%). All models performed well in fundamental areas of Restorative Dentistry, such as "glass ionomer-based materials," "dentin hypersensitivity," and "composite resins." However, in topics that involve both technical aspects and ongoing clinical debates—such as "cavity preparations," "dental amalgams," and "tooth discolorations"—some models achieved accuracy rates below 50% (Fig. 3).

In summary, although all models performed more consistently in Restorative Dentistry compared to Pediatric Dentistry, ChatGPT-4o remained the most accurate and reliable model across both domains.

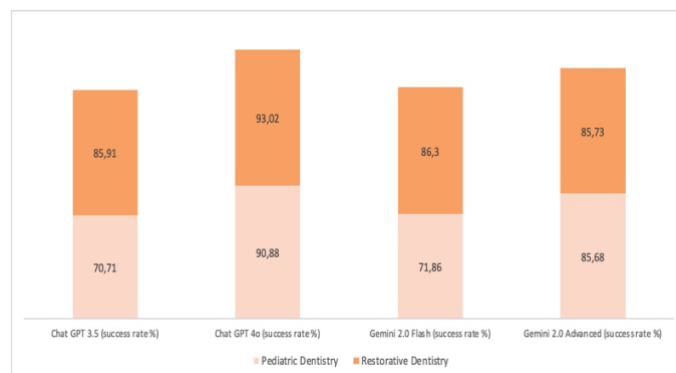


Figure 1. Comparative success rates (%) of ChatGPT and Gemini models in Pediatric and Restorative Dentistry domains of the Turkish Dental Specialty Exam (DUS).

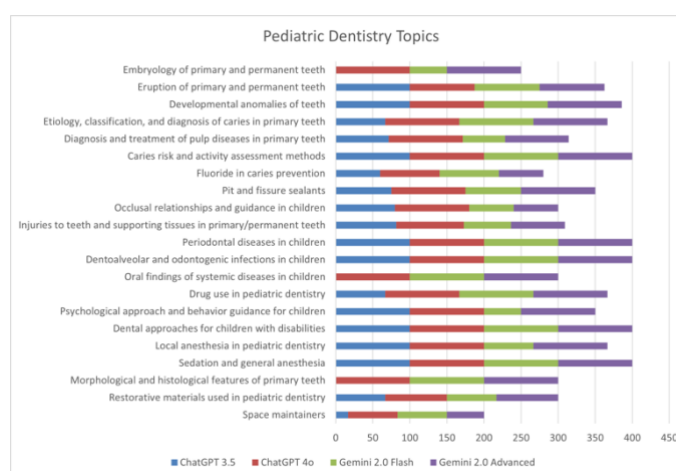


Figure 2. Distribution of correct answers by ChatGPT-3.5, ChatGPT-4o, Gemini 2.0 Flash, and Gemini 2.0 Advanced across different Pediatric Dentistry topics from the Turkish Dental Specialty Exam (DUS).

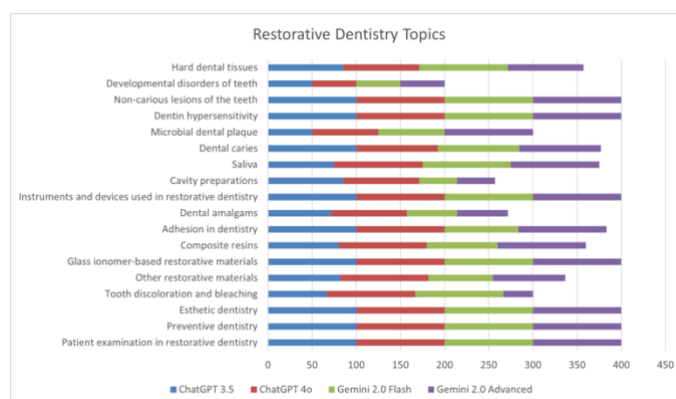


Figure 3. Distribution of correct answers by ChatGPT-3.5, ChatGPT-4o, Gemini 2.0 Flash, and Gemini 2.0 Advanced across different Restorative Dentistry topics from the Turkish Dental Specialty Exam (DUS).

Discussion

In recent years, large language models (LLMs) have emerged as transformative tools in medical and dental education, offering significant potential in knowledge-based assessment and the facilitation of clinical reasoning (26-28). With continuous advancements in natural language processing and information retrieval capabilities, models such as ChatGPT and Gemini have increasingly been utilized in the evaluation of performance in high-stakes, domain-specific examinations. Owing to the structured and conceptually demanding nature of such assessments, LLMs provide a novel framework for exploring the integration of artificial intelligence into healthcare education. Simultaneously, their use enables a critical evaluation of their capabilities and limitations in interpreting and generating content within complex, discipline-specific knowledge domains.

A recent comparative analysis of three major U.S. dental examinations showed that both ChatGPT-3.5 and ChatGPT-4 were capable of effectively answering knowledge-based, case-based, and comprehension-oriented questions, achieving commendable scores (23). These findings suggest that ChatGPT-4 possesses sufficient competence to pass dental licensing examinations and can be effectively employed in domain-specific applications. More broadly, this supports the growing role of LLMs as supportive tools in knowledge-based assessment in the health sciences.

This study comparatively evaluated the response accuracy of four AI-based LLMs—ChatGPT-3.5, ChatGPT-4o, Gemini 2.0 Flash, and Gemini 2.0 Advanced—using questions from the Turkish Dental Specialty Examination (DUS) in Pediatric and Restorative Dentistry. Based on the statistical analyses, the proposed hypotheses can be evaluated as follows: H1 was not supported, as none of the four models showed a statistically significant difference in performance between Pediatric and Restorative Dentistry. H2 was confirmed, with certain models performing significantly lower in Pediatric Dentistry. H3 was not supported, indicating comparable intermodel performance in Restorative Dentistry. Additionally, descriptive temporal analyses revealed year-to-year variability, particularly for Gemini 2.0 Flash and ChatGPT-3.5. These findings may indicate that the conceptual structure and contextual depth of each domain influence model performance, although this was not directly tested in the present study.

Pediatric Dentistry encompasses topics that require a more complex and context-sensitive understanding, such as behavior management, the characteristics of pediatric patients, and developmental dental conditions. Therefore, performance in this domain is closely tied to an LLM's ability to comprehend context and synthesize information. Notably, ChatGPT-4o demonstrated the highest accuracy in Pediatric Dentistry, while ChatGPT-3.5 showed significantly lower performance. This suggests that the enhanced natural language processing

capabilities of the GPT-4 architecture confer a clear advantage in contextual reasoning.

The topic-based analysis in Pediatric Dentistry revealed that ChatGPT-3.5 exhibited notable limitations in interpreting context-specific clinical information, particularly in specialized areas such as embryology and oral manifestations of systemic diseases. This result underscores the model's limited ability to synthesize detailed content and offer effective decision-making support in complex scenarios, suggesting that the GPT-4 architecture's enhanced natural language processing capabilities confer a clear advantage in contextual reasoning.

In the field of Restorative Dentistry, no statistically significant performance differences were observed among the artificial intelligence models, likely due to the domain's more systematic and standardized knowledge structure, which is governed by well-established clinical protocols. All models achieved high accuracy rates, consistently exceeding 80%. This finding is consistent with previous literature indicating that ChatGPT and similar models are capable of accurately responding to knowledge-based questions in structured subject areas.

In Restorative Dentistry, the lower performance of Gemini 2.0 Flash and ChatGPT-3.5 in clinically debated topics such as cavity preparations and dental amalgams suggests potential limitations in the diversity and currency of the training data used in earlier-generation models. By contrast, the high accuracy demonstrated by Gemini 2.0 Advanced in handling technically complex topics supports the notion that multi-layered deep learning architectures and training on contemporary literature enhance the quality of knowledge representation.

Year-by-year analyses revealed that the ChatGPT-4o model demonstrated the most consistent and highest performance across both Pediatric Dentistry and Restorative Dentistry domains. The notable declines in accuracy observed in certain years for the Gemini 2.0 Advanced—and particularly for the Gemini 2.0 Flash—models imply that the training datasets for these models may not have adequately captured temporal variations. This finding indicates that the performance of large language models (LLMs) is influenced not only by architectural complexity but also by the breadth and currency of the training data.

The findings of our study are in line with an expanding body of literature demonstrating the growing potential and accuracy of large language models in medical and dental education and assessment. Recent studies have shown that GPT-4 achieved over 80% accuracy on the United States Medical Licensing Examination (USMLE), in some cases outperforming medical students (29, 30). Similarly, Kung et al. (20) found that ChatGPT demonstrated a moderate level of accuracy approaching the passing threshold on the USMLE. A comprehensive review of studies conducted across eight countries revealed that GPT-3.5 achieved an overall accuracy of 54%, falling short of the passing threshold on all evaluated dental licensing exams. In

contrast, GPT-4 achieved a notably higher accuracy of 72% and passed the majority of these examinations (1). Similarly, in a recent comparative analysis using the Korean Dental Licensing Examination as a benchmark, ChatGPT-4 and Claude3-Opus surpassed the passing threshold across all evaluated years, while ChatGPT-3.5 consistently failed to meet the required scores. Notably, Claude3-Opus slightly outperformed ChatGPT-4 overall, yet both models demonstrated strong performance in knowledge-based subjects, whereas all LLMs underperformed in areas requiring clinical reasoning (21). These findings collectively indicate that ChatGPT demonstrates more consistent performance across various national examination systems.

Consistent with our study, among the evaluated large language models, ChatGPT-4o demonstrated the highest accuracy in responding to assessment questions in restorative dentistry and endodontics, particularly on topics such as caries and direct restorations, indicating its potential value as a supportive tool in dental education (19). When evaluating the reliability of large language models in delivering health-related information, particularly in the context of dental materials, it is essential to compare its outputs with established and authoritative sources. In this regard, the United States Food and Drug Administration (FDA) serves as a globally recognized reference, especially on issues such as the safety and use of dental amalgam. In a recent study by Buldur and Sezer, ChatGPT-4's responses to the FDA's frequently asked questions on dental amalgam were evaluated, and the authors reported that the AI-generated content was largely consistent with FDA information in terms of clarity, completeness, and agreement with reputable health organizations, although some discrepancies were observed in specific technical details and emphasis (16).

Supporting these findings, a recent comparative evaluation involving four different LLMs concluded that ChatGPT-4.0 generated the most accurate, comprehensive, and clinically relevant responses to evidence-based periodontal questions, significantly outperforming Google Gemini in this domain (18). Approaching the topic from a different perspective, Fanelli et al. developed a domain-specific variant of ChatGPT, referred to as "PerioGPT," which was trained exclusively on high-quality periodontal literature. Their study demonstrated that PerioGPT not only outperformed general-purpose LLMs in terms of accuracy and complexity but also delivered more nuanced and clinically applicable content. These results underscore the advantages of fine-tuning large language models using domain-specific corpora, highlighting that aligning AI systems with targeted disciplinary knowledge can substantially enhance their performance in terms of precision, depth, and clinical relevance (9).

Although Gemini models offer certain advantages in terms of speed, accessibility, and operational efficiency, they exhibit notable limitations in contexts that require complex clinical reasoning. Chatzopoulos et al. (18) reported that while these models performed adequately on fundamental questions in periodontology, they were

insufficient for complex, patient-specific clinical decision-making tasks. In a similar vein, Fanelli et al. (9) noted that the Gemini model was less effective than more advanced systems, such as GPT-4o, at generating coherent, contextually appropriate explanatory responses.

Nonetheless, existing literature also points to domains in which Gemini models have yielded comparatively favorable results. In a comparative evaluation employing IADT-based questions on dental avulsion, Gemini achieved statistically superior accuracy scores relative to ChatGPT-3.5 across multiple formats, including multiple-choice, binary (true/false), and explanatory question types (17). Such findings indicate that while the model can perform competently within narrowly defined subject areas, caution should be exercised when considering its application in broader clinical content assessments, particularly those requiring nuanced interpretation and domain-specific expertise.

In our study, the performance of Gemini models in answering DUS examination questions in Pediatric and Restorative Dentistry demonstrated variability across the two disciplines. The Gemini 2.0 Advanced model achieved higher accuracy rates than ChatGPT-3.5 in certain sections, indicating modest improvements in specific subdomains. However, its overall performance remained lower than that of GPT-4o. In contrast, the Gemini 2.0 Flash model consistently yielded the lowest accuracy rates among all models assessed, demonstrating limited reliability across both specialties. These findings suggest that, despite some progress, Gemini models may still exhibit inconsistencies in their ability to provide accurate responses to discipline-specific exam questions.

Collectively, these findings provide valuable insights into the applicability of large language models in educational assessment. It is essential to evaluate not only the accuracy of these models in answering exam questions but also how they can contribute to learning processes. While models like ChatGPT-4o may help dental students develop conceptual understanding in preclinical subjects, models such as Gemini 2.0 Flash may offer advantages in quick review and basic knowledge retrieval. However, the educational use of these models should be structured in accordance with pedagogical principles and ethical guidelines (11, 13). Perspectives from dental educators regarding AI-driven chatbots, such as ChatGPT, highlight the potential for significant transformations within dental education. The incorporation of such technology provides opportunities for aligning dental curricula with broader healthcare educational paradigms, fostering patient-centered care, and enhancing critical thinking competencies. It remains essential to rigorously assess the impacts of AI chatbot technology and formulate clear, evidence-based guidelines for its implementation. Ultimately, the adoption of AI in dental education extends beyond technological novelty, aiming instead to adequately equip future dental practitioners for a continuously evolving, increasingly digitalized healthcare environment (31).

Moreover, although the risk of “hallucination” (the generation of inaccurate or fabricated information), which is frequently emphasized in the literature, was not systematically measured in this study, the presence of incorrect responses in certain years and by specific models implies this risk. In particular, the performance fluctuations observed with the Gemini 2.0 Flash model may be associated with the generation of low-reliability content that could potentially mislead users. These findings suggest that AI systems should be evaluated not only in terms of accuracy but also in terms of reliability and content consistency (8, 32). Future studies should incorporate a structured hallucination assessment framework to systematically evaluate the nature, frequency, and clinical implications of AI-generated errors, including whether models tend to generate confidently stated but inaccurate information in domain-specific contexts.

Although this study was conducted within an educational framework, it provides valuable insights into the potential future integration of large language models into clinical decision-support systems. If models such as ChatGPT-4o are effectively adapted to patient-centered contexts, they may assist clinicians in diagnostic reasoning, treatment planning, and patient communication within the scope of dental practice. However, ethical, legal, and safety considerations must be thoroughly addressed before transitioning to clinical applications (10, 14). Furthermore, given that differences among models can directly impact users, careful planning is required when selecting a model and determining its purpose in educational settings.

One of the key strengths of this study lies in the comparative evaluation of four distinct large language models comprising both free-access and subscription-based versions through domain-specific performance assessments in Pediatric and Restorative Dentistry. Additionally, the application of rigorous statistical methods and the inclusion of a longitudinal analysis across multiple examination years enhance the study's overall scientific contribution.

However, certain limitations must be acknowledged. The evaluation focused exclusively on the correctness of responses (i.e., dichotomous right-or-wrong scoring), without considering the explanatory depth or clinical reasoning embedded in the answers. Furthermore, as the study was limited to two dental specialties, the findings may not be readily generalizable to other areas of dental practice. Future research should aim to include standardized scoring frameworks to assess the quality and reasoning behind AI-generated responses, broaden the disciplinary scope, and incorporate case-based scenarios that challenge higher-order clinical decision-making and reasoning competencies. An additional limitation of this study is that all examination questions were obtained from publicly available ÖSYM sources. Therefore, it cannot be excluded that some items, or similar explanations of these items, may have been included in model training corpora or indexed web resources. Accordingly, model performance may reflect not only domain-specific reasoning ability, but also prior

exposure to publicly available examination materials. Future studies should use newly developed or unpublished questions to better isolate genuine reasoning performance.

Conclusion

In conclusion, this study conducted a comparative evaluation of four large language models in their ability to respond to knowledge-based questions from the fields of Pediatric Dentistry and Restorative Dentistry, using items derived from the Turkish Dental Specialty Examination (DUS). The analysis revealed key insights into the strengths and limitations of each model. Among them, ChatGPT-4o demonstrated superior overall accuracy and consistency, particularly in complex and context-rich domains, whereas Gemini 2.0 Flash exhibited comparatively weaker performance, especially in tasks requiring contextual interpretation.

One out of three proposed hypotheses was supported (H_2), while H_1 and H_3 were not confirmed. These findings suggest that large language models hold potential as supplementary tools in dental education, although model selection should be tailored to the complexity and contextual demands of the specific domain.

Year-by-year analyses further highlighted the importance of model architecture, training data breadth, and the currency of information in shaping LLM performance. Within this framework, large language models hold potential as supplementary tools in dental education and assessment.

Disclosures

Acknowledgments: This study has been presented at the 28th Turkish Dental Association International Congress in Diyarbakır, Türkiye, held between September 18-21, 2025.

Peer-review: Externally peer-reviewed.

Author Contributions: Conception – A.A.Y.; Design – A.A.Y., M.A.; Supervision – M.A.; Materials – M.A., A.A.Y.; Data Collection and/or Processing – M.A.; Analysis and/or Interpretation – A.A.Y., M.A.; Literature Review – M.A.; Writer – A.A.Y.; Critical Review – A.A.Y., M.A.

Conflict of Interest: No conflict of interest was declared by the authors.

Funding: The authors declared that this study has received no financial support.

References

- Liu M, Okuhara T, Huang W, et al. Large language models in dental licensing examinations: systematic review and meta-analysis. *Int Dent J*. 2025;75(1):213-222. <https://doi.org/10.1016/j.identj.2024.10.014>
- Zhang Q, Wu Z, Song J, Luo S, Chai Z. Comprehensiveness of large language models in patient queries on gingival and endodontic health. *Int Dent J*. 2025;75(1):151-157. <https://doi.org/10.1016/j.identj.2024.06.022>
- 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. <https://doi.org/10.1016/j.jdent.2024.104927>
- Uribe SE, Maldupa I. Estimating the use of ChatGPT in dental research publications. *J Dent*. 2024;149:105275. <https://doi.org/10.1016/j.jdent.2024.105275>
- Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. *J Biomed Inform*. 2024;151:104620. <https://doi.org/10.1016/j.jbi.2024.104620>
- Alhaidry HM, Fatani B, Alrayes JO, Almana AM, Alhaed NK. ChatGPT in dentistry: a comprehensive review. *Cureus*. 2023;15(4):e38317. <https://doi.org/10.7759/cureus.38317>
- Kalaimani G, Sivapathasundharam B, Chockalingam RM, Karthick P. Evaluation of knowledge, attitude, and practice (KAP) of artificial intelligence among dentists and dental students: a cross-sectional online survey. *Cureus*. 2023;15(9):e44656. <https://doi.org/10.7759/cureus.44656>
- 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):1098-1102. <https://doi.org/10.1111/jerd.13046>
- Fanelli F, Saleh M, Santamaria P, Zhurakivska K, Nibali L, Troiano G. Development and comparative evaluation of a re-instructed GPT-4o model specialized in periodontology. *J Clin Periodontol*. 2025;52(5):707-716. <https://doi.org/10.1111/jcpe.14101>
- Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. *J Dent Res*. 2020;99(7):769-774. <https://doi.org/10.1177/0022034520915714>
- Lee H. The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ*. 2024;17(5):926-931. <https://doi.org/10.1002/ase.2270>
- Jamal A, Solaiman M, Alhasan K, Temsah M-H, Sayed G, Soliman M. Integrating ChatGPT in medical education: adapting curricula to cultivate competent physicians for the AI era. *Cureus*. 2023;15(8):e43036. <https://doi.org/10.7759/cureus.43036>
- Kavadella A, Da Silva MAD, Kaklamanos EG, Stamatopoulos V, Giannakopoulos K. Evaluation of ChatGPT's real-life implementation in undergraduate dental education: mixed methods study. *JMIR Med Educ*. 2024;10(1):e51344. <https://doi.org/10.2196/51344>
- Nia MF, Ahmadi M, Irankhah E. Transforming dental diagnostics with artificial intelligence: advanced integration of ChatGPT and large language models for patient care. *Front Dent Med*. 2025;5:1456208. <https://doi.org/10.3389/fdmed.2024.1456208>
- Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. *Lancet Digit Health*. 2023;5(6):e333-e335. [https://doi.org/10.1016/S2589-7500\(23\)00083-3](https://doi.org/10.1016/S2589-7500(23)00083-3)
- Buldur M, Sezer B. Evaluating the accuracy of Chat Generative Pre-trained Transformer version 4 (ChatGPT-4) responses to United States Food and Drug Administration (FDA) frequently asked questions about dental amalgam. *BMC Oral Health*. 2024;24(1):605. <https://doi.org/10.1186/s12903-024-04358-8>
- 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):178-186. <https://doi.org/10.1111/edt.12999>
- Chatzopoulos GS, Koidou VP, Tsalikis L, Kaklamanos EG. Large language models in periodontology: assessing their performance in clinically relevant questions. *J Prosthet Dent*. 2025;134(6):2328-2336. <https://doi.org/10.1016/j.prosdent.2024.10.020>
- Künzle P, Paris S. Performance of large language artificial intelligence models on solving restorative dentistry and endodontics student assessments. *Clin Oral Investig*. 2024;28(11):575. <https://doi.org/10.1007/s00784-024-05968-w>
- 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. <https://doi.org/10.1371/journal.pdig.0000198>
- Kim W, Kim BC, Yeom H-G. Performance of large language models on the Korean dental licensing examination: a comparative study. *Int Dent J*. 2025;75(1):176-184. <https://doi.org/10.1016/j.identj.2024.09.002>
- Chau RCW, Thu KM, Yu OY, Hsung RT-C, Lo ECM, Lam WYH. Performance of generative artificial intelligence in dental licensing examinations. *Int Dent J*. 2024;74(3):616-621. <https://doi.org/10.1016/j.identj.2023.12.007>
- Dashti M, Ghasemi S, Ghadimi N, et al. Performance of ChatGPT 3.5 and 4 on US dental examinations: the INBDE, ADAT, and DAT. *Imaging Sci Dent*. 2024;54(3):271-275. <https://doi.org/10.5624/isd.20240037>
- Ohta K, Ohta S. The performance of GPT-3.5, GPT-4, and Bard on the Japanese national dentist examination: a comparison study. *Cureus*. 2023;15(12):e50369. <https://doi.org/10.7759/cureus.50369>
- ÖSYM. Assessment, Selection, and Placement Center. 2025. <https://www.osym.gov.tr> Accessed 12 June, 2025.
- Rewthamrongsris P, Burapachee J, Trachoo V, Pornaveetus T. Accuracy of large language models for infective endocarditis prophylaxis in dental procedures. *Int Dent J*. 2025;75(1):206-212. <https://doi.org/10.1016/j.identj.2024.09.033>
- 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(1):108-113. <https://doi.org/10.1111/iej.13985>
- Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. *Health Care Sci*. 2023;2(4):255-263. <https://doi.org/10.1002/hcs2.61>
- Knoedler L, Alfertshofer M, Knoedler S, Hoch CC, Funk PF, Cotofana S, et al. Pure wisdom or Potemkin villages? A comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE step 3 style questions: quantitative analysis. *JMIR Med Educ*. 2024;10(1):e51148. <https://doi.org/10.2196/51148>
- Gilson A, Safranek CW, Huang T, 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(1):e45312. <https://doi.org/10.2196/45312>
- Uribe SE, Maldupa I, Kavadella A, et al. Artificial intelligence chatbots and large language models in dental education: worldwide survey of educators. *Eur J Dent Educ*. 2024;28(4):865-876. <https://doi.org/10.1111/eje.13009>
- Ning Y, Teixayavong S, Shang Y, et al. Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. *Lancet Digit Health*. 2024;6(11):e848-e856. [https://doi.org/10.1016/S2589-7500\(24\)00143-2](https://doi.org/10.1016/S2589-7500(24)00143-2)