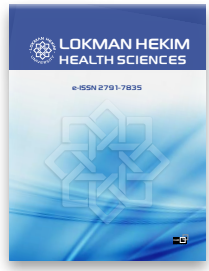
LOKMAN HEKIM HEALTH SCIENCES

DOI: 10.14744/lhhs.2025.42941
Lokman Hekim Health Sci 2026;6(1):126–133



lokmanhekimhs.com

ORIGINAL ARTICLE

Comparative Evaluation of ChatGPT-4o, Gemini 2.5 Pro and Grok-4 in Answering Orthodontics Questions from the Dentistry Specialty Examination

Samet Özden[1], Hande Erener[2]

[1]Department of Orthodontics, İnönü University Faculty of Dentistry, Malatya, Türkiye
[2]Department of Orthodontics, Tekirdağ Namık Kemal University Faculty of Dentistry, Tekirdağ, Türkiye

Abstract

Introduction: This study aimed to compare the accuracy of three advanced Large Language Models (LLMs) in answering orthodontics-related questions from the Turkish Dentistry Specialty Examination (DUS) and to assess their performance across different examination periods.

Methods: A total of 129 orthodontic questions that were publicly available from 13 DUS sessions conducted between 2012 and 2021 were included. All questions were presented in their original Turkish format, simultaneously, and under identical default settings (i.e., without fine-tuning or additional prompt engineering) by the same operator to eliminate procedural variability. Each model's responses were recorded and scored as correct (1) or incorrect (0). Accuracy comparisons among LLMs were performed using Chi-square and Fisher's exact tests with Monte Carlo correction. Statistical significance was set at $p<0.05$.

Results: No statistically significant differences were observed among the three LLMs within individual examination periods ($p>0.05$). Grok-4 achieved the highest cumulative accuracy (112/129; 86.8%), followed by Gemini (107/129; 82.9%) and ChatGPT-4o (101/129; 78.3%). The 2018 DUS yielded the lowest accuracy for all models (30%, 30%, and 50%, respectively). All three LLMs performed significantly better on text-based than on figure-based questions ($p<0.05$), with figure-based accuracy dropping to 45.5% for ChatGPT and Gemini, and 63.6% for Grok. No significant inter-model differences were found within each question type ($p>0.05$).

Discussion and Conclusion: All three LLMs demonstrated high but not flawless accuracy in orthodontics-related DUS questions, with consistent challenges in visual question interpretation. While their integration into examination preparation and dental education holds promise, further refinement in visual reasoning and domain-specific adaptation is needed before clinical or high-stakes implementation.

Keywords: Artificial intelligence; ChatGPT; Dentistry specialty examination; Gemini; Grok; Large language model

Artificial intelligence (AI) is defined as a technology that enables machines to simulate cognitive processes typically associated with human intelligence, including learning, reasoning, and problem-solving.[1,2] Among the subfields are machine learning (ML), which allows systems to improve their performance based on data, and deep

Cite this article as: *Özden S, Erener H. Comparative Evaluation of ChatGPT-4o, Gemini 2.5 Pro and Grok-4 in Answering Orthodontics Questions from the Dentistry Specialty Examination. Lokman Hekim Health Sci 2026;6(1):126–133.*

Correspondence: Samet Özden, M.D. İnönü Üniversitesi Diş Hekimliği Fakültesi, Ortodonti Anabilim Dalı, Malatya, Türkiye
E-mail: drsametozden@gmail.com **Submitted:** 18.08.2025 **Revised:** 12.09.2025 **Accepted:** 19.11.2025 **Available Online:** 16.03.2026

learning, which includes multilayer neural networks.[1] Large Language Models (LLMs) represent an advanced form of these systems, designed to process and generate natural language based on user input.[3] Representative models in this category include ChatGPT (Chat Generative Pretrained Transformer; OpenAI, San Francisco, USA), Gemini (Google DeepMind, California, USA), and Grok (xAI, San Francisco, USA).[4,5]

State-of-the-art conversational AI models, such as ChatGPT-4o and Gemini 2.5 Pro, represent the latest advancements in LLM-based systems. These models are multimodal AI systems distinguished by their capacity to simultaneously comprehend and analyze various data types, including text, audio, images and video.[6–8] Grok is an AI model developed by Elon Musk's company xAI, integrated with the X platform (formerly Twitter), and distinguished by its emphasis on freedom of expression and the promotion of more dynamic social interactions.[5]

In recent years, the use of AI models in healthcare has steadily increased, driven by advancements in ML and image analysis.[9] In the field of dentistry, these models can serve as complementary tools to support various clinical applications, such as reducing workload, facilitating diagnosis and decision-making, enhancing treatment planning, predicting treatment outcomes, and improving the accuracy and reliability of prognosis.[9,10]

AI models are capable of responding to academic and clinical questions across various scientific disciplines through their ability to comprehend and generate human language.[3,4,11] Moreover, recent studies have shown that GPT-based models can also analyze visual data, further extending their potential beyond text-based applications.[12] Thus, in addition to supporting individual information access for patients, these technologies can serve as valuable educational tools for students and professionals, particularly in the context of examination preparation.[4,7] Consequently, these models are also utilized at various stages of education. The integration of virtual reality and AI applications into educational processes paves the way for the digitalization of education. This is particularly beneficial in fields such as dentistry, where the educational process is intensive and comprehensive.[13]

In Türkiye, in order for dentistry students to undertake specialization training in various departments after completing their undergraduate studies, a Dentistry Specialty Examination (DUS), administered by the Assessment, Selection and Placement Center (ÖSYM). Each examination consists of 40 basic science questions and 80 clinical science questions. The clinical science section includes 10 questions from each major dental specialty.[13] During the preparation process for this examination, candidates can use artificial intelligence models for support. Therefore, the reliability of the responses provided by AI models is of critical importance for ensuring access to accurate information.

Various studies have been conducted to evaluate the contribution of AI models to education, utilizing questions asked on the DUS (such as anatomy, pedodontics, and prosthetic dentistry).[7,13,14] However, a review of the literature reveals that no studies have utilized DUS orthodontics questions.

The aim of the study is to assess the performance of three different AI models (ChatGPT-4o, Gemini 2.5 Pro, and Grok) in answering orthodontic questions previously asked in the DUS, and to evaluate the differences in their knowledge levels. The findings are intended to contribute to the usability of AI in examination preparation and educational processes in dentistry.

Materials and Methods

Study Place and Design

This study was conducted in a virtual environment using publicly available resources and language model interfaces. The study aimed to assess the answer accuracy of three LLMs—ChatGPT-4o (OpenAI, USA), Gemini 2.5 Pro (Google, USA), and Grok-4 (xAI, USA)—in responding to orthodontics-related questions from the DUS. At the time of data collection (July 2025), these LLMs represented the most advanced publicly accessible versions of their respective platforms.

The DUS is a nationally standardized examination administered by the ÖSYM for admission to dental specialty training programs. Each examination consists of 120 multiple-choice questions, comprising 40 basic science and 80 clinical science items, with five answer options and a single correct answer. The questions are publicly released on the official ÖSYM website following each examination.

Research Type

This was a descriptive, cross-sectional, and comparative in silico study, designed to evaluate the accuracy of LLMs in answering multiple-choice orthodontic questions. The study employed quantitative analysis using a binary scoring system to compare model performances.

Data Collection

A total of 129 orthodontics-related questions from DUS examinations conducted between 2012 and 2021 were included in the study. While 130 orthodontic questions were available during this period, one question from the 2021 DUS was officially annulled by ÖSYM and excluded. All questions were obtained directly from the official ÖSYM website (https://www.osym.gov.tr/TR,15070/dus-cikmis-sorular.html) and were preserved in their original Turkish-language format.

Due to a change in ÖSYM's publication policy, official question booklets have not been released since 2021. Consequently, DUS questions from 2022 to 2025 were not publicly available and thus could not be included in this study.

The dataset comprised both text-based theoretical questions and visual questions. Specifically, seven questions from the 2018 DUS and three questions from the 2019 DUS included visual content. These visual items were presented to the LLMs using unaltered screenshots taken directly from the official examination booklets to ensure standardization and authenticity. All images were used without modification, preserving their original resolution, color, and size, and answers with partial or ambiguous elements were scored as incorrect to ensure objective evaluation.

All text-based questions were extracted from the official ÖSYM PDF files and converted into plain text format without alteration. These were subsequently presented to each LLM in an identical format, ensuring uniformity in input structure. In line with the DUS question structure, the included orthodontic questions were predominantly clinical. The distinction between theoretical and clinical content was noted during data organization to provide a clearer basis for interpreting model performance.

Data Collection Tools

The three LLMs—ChatGPT-4o, Gemini 2.5 Pro, and Grok-4—were accessed via their official online platforms. Model responses were collected in real-time and stored digitally for analysis.

All questions were submitted individually and simultaneously to each LLM by the same operator (S.Ö) using the same computer system to maintain procedural consistency. No model was provided with any pretraining cues, additional context, or leading prompts that could result in overfitting, underfitting, or biased performance. Each question was posed in a neutral and standardized manner, replicating a realistic user query scenario without influencing the model's natural response behavior.

Each model's performance was evaluated using a binary scoring system, where a correct answer was coded as "1", and an incorrect or no answer as "0"; all evaluations were performed by the same researcher (S.Ö.) using the official answer key to ensure consistency. No partial credit was awarded. Based on the scoring results, accuracy rates were calculated for each model by dividing the total number of correct responses by the total number of questions.

Ethical Consideration

This study did not involve human participants, animal subjects, or patient data. All DUS questions obtained from a publicly accessible ÖSYM website. Therefore, ethical approval was not required, and the principles of the Declaration of Helsinki were not applicable.

Statistical Analysis

Data were analyzed using IBM SPSS v23 (IBM Corp., Armonk, NY, USA). The association between categorical variables was examined using the Monte Carlo-corrected Fisher's Exact Test and Pearson's Chi-square test. Multiple comparisons were performed using the Bonferroni-corrected z-test. Descriptive statistics for categorical variables were presented as n and %. The significance level was set at $p<0.05$.

Results

A total of 129 orthodontic questions from 13 DUS conducted between 2012 and 2021 were included in this study. Each of the three LLMs was evaluated on the same set of questions, all of which were administered in their original Turkish format under identical conditions. Among the 129 items, 11 were visual (figure-based) and 118 were textual (text-based).

Table 1 presents the answer accuracy of each LLM across individual examination periods. No statistically significant differences were observed between the three LLM models within each examination period ($p>0.05$). However, Grok and Gemini consistently provided more accurate responses than ChatGPT across most examination sessions, with both models achieving perfect accuracy in the 2015 and 2016 DUS examinations. Despite the lack of statistical significance, it was notable that Grok-4 produced the highest cumulative accuracy (112 correct; 86.8%), followed by Gemini (107 correct; 82.9%) and ChatGPT-4o (101 correct; 78.3%) across all examinations. Figure 1 illustrates the year-by-year variation in accuracy rates for

Table 1. Performance evaluation of three different LLMs in the Turkish Dental Specialty Examinations across exam sessions (2012–2021).

	ChatGPT-4o n (%)	Gemini 2.5 Pro n (%)	Grok-4 n (%)	Test statistic	p
2012-Spring	7 (70)	7 (70)	9 (90)	1.491	0.475[y]
2012-Autumn	8 (80)	7 (70)	8 (80)	0.511	1.000[x]
2013-Spring	9 (90)	10 (100)	9 (90)	1.312	1.000[x]
2013-Autumn	7 (70)	9 (90)	9 (90)	1.715	0.570[x]
2014-Spring	8 (80)	8 (80)	8 (80)	0.231	1.000[x]
2014-Autumn	7 (70)	9 (90)	8 (80)	1.278	0.843[x]
2015	9 (90)	10 (100)	10 (100)	1.885	1.000[x]
2016	9 (90)	10 (100)	10 (100)	1.885	1.000[x]
2017	10 (100)	9 (90)	10 (100)	1.885	1.000[x]
2018	3 (30)	3 (30)	5 (50)	1.161	0.714[x]
2019	7 (70)	8 (80)	8 (80)	0.511	1.000[x]
2020	9 (90)	9 (90)	10 (100)	1.312	1.000[x]
2021	8 (88.9)	8 (88.9)	8 (88.9)	0.437	1.000[x]
Test statistic	18.37	22.77	15.289		
p	0.057[x]	0.004[x]	0.082[x]		

x: Fisher's exact test with Monte Carlo correction; y: Pearson's Chi-square test; LLM: Large language model.

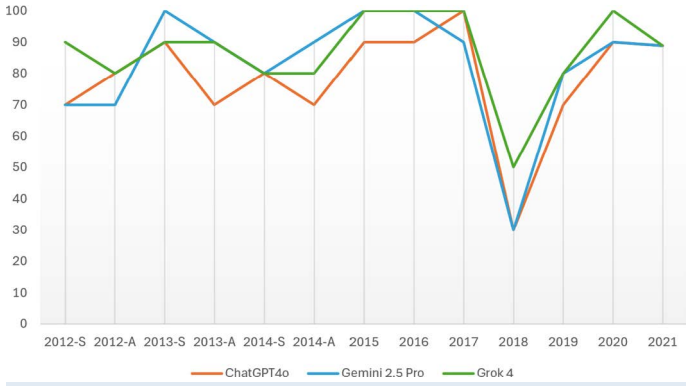


Figure 1. Performance trends of ChatGPT-4o, Gemini 2.5 Pro, and Grok-4 across 13 Turkish Dental Specialty Examinations (2012–2021).

the three models, highlighting both their generally parallel performance trends and the sharper fluctuations seen in certain periods.

Table 1 also summarizes the models' performance relative to the examination periods. While no statistically significant differences were observed among the majority of examination periods ($p>0.05$), the 2018 DUS drew attention as the session with the lowest overall accuracy across all three LLMs—only 30%, 30%, and 50% correct responses from ChatGPT, Gemini, and Grok, respectively. Although not statistically significant in the broader year-to-year context, this marked drop suggests that the 2018 examination may have posed comparatively greater challenges to all models.

Regarding the comparison between text-based and figure-based questions (Table 2), all three LLMs demonstrated a significant drop in performance for visual items ($p<0.05$). While text-based items were answered correctly at rates of 81.4% (ChatGPT), 86.4% (Gemini), and 89.0% (Grok), these figures dropped substantially for figure-based items—down to 45.5% for both ChatGPT and Gemini, and 63.6% for Grok. This difference was statistically significant for each model ($p<0.05$), indicating a consistent difficulty in processing visual information. However, when comparing the performance of the three LLMs on text and figure-based questions separately, no statistically significant differences were observed among the models ($p>0.05$), suggesting that all three exhibited comparable performance within each question type. Figure 2 compares the models' accuracy percentages for text-based, figure-based, and overall questions, clearly showing the largest performance gap in the figure-based category.

Discussion

The present-day AI, especially in advanced LLMs like ChatGPT-4o, Gemini 2.5 Pro, and Grok-4, is capable of providing meaningful engagement in specialized professional fields, i.e., dentistry. In recent years, generative AI has transformed dental education by enabling automated question answering, case simulation, and educational content generation.[15] LLMs have been employed with great success in a variety of dental education

Table 2. Performance comparison of three different LLMs in text-based and figure-based questions from the Turkish Dental Specialty Examinations (2012–2021).

	ChatGPT-4o n (%)	Gemini 2.5 Pro n (%)	Grok-4 n (%)	Test statistic	p
Text-Based Questions				2.803	0.259[x]
Incorrect	22 (18.6)	16 (13.6)	13 (11)		
Correct	96 (81.4)	102 (86.4)	105 (89)		
Figure-Based Questions				1.016	0.749[x]
Incorrect	6 (54.5)	6 (54.5)	4 (36.4)		
Correct	5 (45.5)	5 (45.5)	7 (63.6)		
p[y]	**0.013**	**0.003**	**0.039**		
General				3.251	0.217[x]
Incorrect	28 (21.7)	22 (17.1)	17 (13.2)		
Correct	101 (78.3)	107 (82.9)	112 (86.8)		

x: Fisher's exact test with Monte Carlo correction; y: Pearson's Chi-square test; LLM: Large language model.

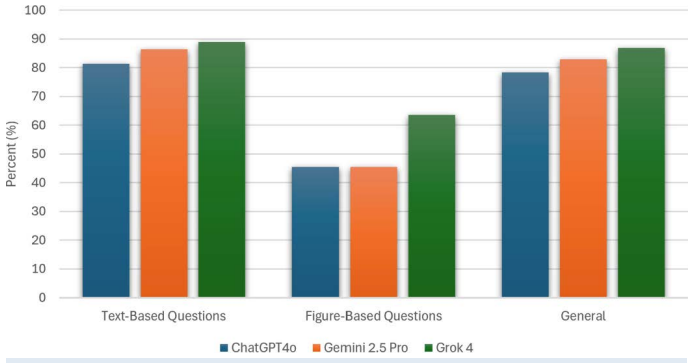


Figure 2. Comparative accuracy rates of ChatGPT-4o, Gemini 2.5 Pro, and Grok-4 for text-based questions, figure-based questions, and overall performance.

settings, including the instruction of periodontal charting. In addition, LLMs have been utilised to assist with the organisation and summarisation of dental clinical content for students. These applications demonstrate the potential of LLMs to support students' conceptual understanding by providing personalised learning experiences. However, numerous challenges persist, including model bias, data privacy concerns, and the presence of inaccurate or outdated information.[16]

The progressive development of LLM architecture has resulted in significant advancements across multiple domains, including, but not limited to, inference capability, contextual understanding and multimodality. The enhancement of visual interpretation and inference capabilities was achieved through the utilisation of the Gemini 2.5 Pro model. The utilisation of Grok-4 and ChatGPT-4o has been demonstrated to facilitate the enhancement of long-context comprehension and structured problem-solving capabilities.[17,18] In light of this context, the present study utilised the most recent and advanced versions of all three platforms: ChatGPT-4o (OpenAI) and Gemini 2.5 Pro (Google) and Grok-4 (xAI), with the objective of attaining the utmost levels of model capability, consistency, and quality in the evaluation process.

In the literature, studies in dentistry commonly use the ChatGPT and Gemini AI models, whereas research using the Grok model is limited.[3,4,7,13] Several studies have shown that ChatGPT can be used in medical and dental education, as well as in clinical decision-making.[9–11,17,19,20] In addition, some studies comparing ChatGPT and Gemini reported that Gemini could also be applied in a similar way and, in some cases, produced more scientific and evidence-based results.[21] Dave et al.[3] conducted a study similar to the present one, where three language models (GPT-4o, Grok2, and Gemini) were tested, and found that Grok2 and GPT-4o performed better than Gemini in answering dental assessment questions. Therefore, the inclusion of all three models in this study has allowed for a comprehensive evaluation not only of their contemporary technological capacities but also of the impact of different AI systems on orthodontic examination performance.

Within dental education, multiple systematic reviews and narrative studies emphasize the feasibility of integrating LLMs as virtual tutors capable of generating exam-style questions, offering immediate feedback, and supporting individualized learning experiences.[22] However, the literature also points to several caveats, including the risk of hallucinated information, limitations in updating capabilities, and the potential for overreliance if not used critically.[23,24]

The DUS is a highly standardized, nationally competitive exam consisting of 40 basic science items and 80 clinical science items; with regard to the DUS—there are many ways AI can be applied. Specifically, in terms of the orthodontics portion of the DUS, there will be a mix of text-based and visual items that will require the use of reasoning skills beyond simple recall. Although some LLMs have been shown to develop early multimodal capabilities, none of the LLMs developed to date have been able to consistently and accurately interpret dental radiographs and intra-oral images. However, incorporating LLMs into DUS preparation processes may provide a number of educational advantages such as adaptive tutorial systems for learners, automatic assessment and performance tracking for educators and simulated examination experiences for learners.

Our study leverages the most advanced versions of LLMs available as of mid-2025—ChatGPT-4o (OpenAI), Gemini 2.5 Pro (Google), and Grok-4 (xAI)—and applies a standardized binary scoring system (1 = correct, 0 = incorrect/no response) across 129 orthodontics questions. Visual items have been included in DUS examinations, with 1 item in 2015, 7 items in 2018 and 3 items in 2019, whereas the majority of the exam has consisted of text-based, theory-driven questions. Accordingly, this study evaluated the LLMs' performance not only on conventional theoretical knowledge-based items but also on visual content, allowing for a more comprehensive assessment of their capabilities. All questions were administered under uniform conditions to each model. To our knowledge, this represents the first study in the literature to evaluate the performance of LLMs specifically on orthodontic questions from the DUS examination, thereby addressing a notable gap in the current body of research.

Empirical evidence supports the growing capacity of LLMs to handle domain-specific content in standardized assessments. In a recent study, ChatGPT-4 achieved approximately 75.9% accuracy on the Integrated National Board Dental Examination (INBDE), demonstrating robust performance under exam-like conditions.[25] Similarly, it reached a mean accuracy of 79.6% when evaluated using American Academy of Periodontology (AAP) in-service examination questions.[20] Another study focusing on the AAP examination reported accuracy rates of 57.9% and 73.6% for ChatGPT-3.5 and ChatGPT-4, respectively, further emphasizing the superiority of GPT-4 in the domain of periodontology.[26] In the Swiss Federal Licensing Examination in Dental Medicine, ChatGPT-4 outperformed other models such as GPT-3.5 and Claude, showing a high degree of adaptability across examination frameworks.[27] Furthermore, a recent systematic review[28] evaluating the performance of LLMs in dental licensing examinations analyzed 11 eligible studies and reported integrated accuracy rates of 54% for GPT-3.5, 72% for GPT-4, and 56% for Bard. Despite the observed advancements, the authors concluded that the overall performance of current LLMs remains insufficient, stating that these models are not yet suitable for use in dental education or clinical diagnostics.

In light of these findings, our study provides further evidence supporting the growing competence of LLMs in domain-specific, standardized dental examinations. When examined across 13 different DUS conducted between 2012 and 2021, no statistically significant differences were observed among ChatGPT-4o, Gemini 2.5 Pro, and Grok-4 in terms of answer accuracy within each individual examination period. This consistency likely reflects the standardized delivery protocol—same operator, device, timing, and format without prior tuning or prompting—which minimized procedural bias. Moreover, the absence of significant differences among LLMs across examination periods may be due to their similar transformer-based architectures and multilingual training, leading to convergent performance in standardized assessments.

Despite the absence of significant year-specific differences, cumulative performance revealed a ranking among the models: Grok-4 achieved the highest total accuracy (112 correct; 86.8%), followed by Gemini (107 correct; 82.9%) and ChatGPT-4o (101 correct; 78.3%). Although these differences were statistically insignificant when examined by periods, possible explanations for this trend include architectural or update cycle differences between models (e.g., Grok's strong performance on both text and visual items). The trend is consistent with the notion that all three models have the ability to navigate domain-specific examination content at a high level of accuracy. However, as previously mentioned, incremental improvements to either the algorithmic reasoning of each model or how well the model aligns with its dataset may impact cumulative performance.

Another key finding emerged from the analysis of question formats. While all models demonstrated relatively high accuracy on text-based items—ranging from 81.4% (ChatGPT) to 89.0% (Grok)—their performance declined considerably for figure-based questions, with accuracy dropping to 45.5% for both ChatGPT and Gemini, and to 63.6% for Grok. This discrepancy for each model highlights a shared limitation in processing visual data. Notably, no

significant differences were found among the three LLMs when comparing their performance on text versus figure questions, suggesting that the challenge of interpreting visual content is common across models regardless of architecture.

Finally, when comparing performance across examination periods, the 2018 DUS stood out as a statistically significant outlier, with all three LLMs scoring markedly lower (30% for both ChatGPT and Gemini; 50% for Grok) compared to other periods. This suggests that either the content or structure of the 2018 examination introduced additional complexity for AI-based models. In contrast, no other examination period demonstrated significant differences, implying that the difficulty level of most DUS examinations remained relatively stable over time.

Nevertheless, several limitations of this study should be acknowledged. First, the analysis was limited to multiple-choice orthodontic questions from past DUS, which may not fully capture the breadth of clinical reasoning or decision-making required in real-world orthodontic practice. Second, the sample size of figure-based questions (n=11) was relatively small compared to text-based items, which may limit the generalizability of findings related to visual comprehension. Third, while care was taken to standardize question delivery, LLMs operate as probabilistic models and may exhibit slight output variability depending on timing or backend updates beyond user control. Finally, the study focused solely on response accuracy and did not assess other critical dimensions such as explanation quality, consistency, or interpretability, which are important in educational and diagnostic contexts.

Conclusions

This research has established a standard against which to evaluate LLMs in orthodontic specialty examinations. Despite no differences being observed between the performance of three LLMs at various time points, or with respect to question type, a 30% decrease in overall accuracy on image-based questions was noted, in comparison with those based on text. The findings of this study suggest that while LLMs show promise for use in the field of dental education, current limitations must be taken into consideration. As these systems are continually refined, with appropriate use and oversight by dental educators, there is a possibility that they can enhance students' ability to prepare for and take specialty exams, particularly when they are utilised in conjunction with relevant visual materials.

Ethics Committee Approval: This study did not involve human participants, animal subjects, or patient data. All DUS questions were obtained from a publicly accessible ÖSYM website. Therefore, ethical approval was not required for his study.

Conflict of Interest: None declared.

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

Use of AI for Writing Assistance: No AI tool was used for statistical analysis, writing, or drafting. The reference to AI was included solely to reflect the study's thematic focus, as the research itself investigates AI-based models.

Authorship Contributions: Concept: SÖ, HE; Design: SÖ, HE; Supervision: SÖ, HE; Resource: SÖ, HE; Materials: SÖ, HE; Data Collection SÖ, HE; Analysis: SÖ, HE; Literature Search: SÖ, HE; Writing and Editing: SÖ, HE; Critical Reviews: SÖ, HE.

Peer-review: Double blind peer-reviewed.

References

1. Dave M. EBD spotlight: Artificial intelligence and dental panoramic radiography. BDJ Team 2024;11(6):244-5. [CrossRef]
2. Khanagar SB, Al-Ehaideb A, Vishwanathaiah S, Maganur PC, Patil S, Naik S, et al. Scope and performance of artificial intelligence technology in orthodontic diagnosis, treatment planning, and clinical decision-making - A systematic review. J Dent Sci 2021;16(1):482-92. [CrossRef]
3. Dave M, Tattar R, Alafaleg R, Barry S, Ariyaratnam S, Roudsari RV, Patel N. Performance of large language models (ChatGPT4-0, Grok2 and Gemini) in UK dentistry and dental hygiene and therapy assessments. Br Dent J. 2025 Jun 20. doi: 10.1038/s41415-025-8383-2. Epub ahead of print. PMID: 40542155. [CrossRef]
4. Aziz AAA, Abdelrahman HH, Hassan MG. The use of ChatGPT and Google Gemini in responding to orthognathic surgery-related questions: A comparative study. J World Fed Orthod 2025;14(1):20-6. [CrossRef]
5. De Carvalho Souza ME, Weigang L. Grok. Gemini, ChatGPT and DeepSeek: Comparison and applications in conversational artificial intelligence. Intel Artif 2025;2(1):1-7.
6. Shahriar S, Lund BD, Mannuru NR, Arshad MA, Hayawi K, Bevara RVK, et al. Putting GPT-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. Appl Sci 2024;14(17):7782. [CrossRef]
7. Bilgin AD, Ertan AA. A comparative study of ChatGPT-3.5 and Gemini's performance of answering the prosthetic dentistry questions in Dentistry Specialty Exam: Cross-sectional study. Turkiye Klinikleri J Dental Sci 2024;30(4):668-73. [In Turkish] [CrossRef]
8. Gemini. An overview of the Gemini app. Available at: https://gemini.google/overview/?hl=tr. Accessed August 1, 2025.
9. Alhaidry HM, Fatani B, Alrayes JO, Almana AM, Alfhaed NK. ChatGPT in dentistry: A comprehensive review. Cureus 2023;15(4):e38317. [CrossRef]
10. Martins Lima NG, Costa L, Bittencourt Santos P. ChatGPT in orthodontics: Limitations and possibilities. Australas Orthod J 2024;40(2):19-21. [CrossRef]

11. Tanaka OM, Gasparello GG, Hartmann GC, Casagrande FA, Pithon MM. Assessing the reliability of ChatGPT: a content analysis of self-generated and self-answered questions on clear aligners, TADs and digital imaging. Dental Press J Orthod 2023;28(5):e2323183. [CrossRef]
12. Yıldırım A, Cicek O, Genç YS. Can AI-Based ChatGPT models accurately analyze hand-wrist radiographs? A comparative study. Diagnostics (Basel) 2025;15(12):1513. [CrossRef]
13. Aşık A, Kuru E. Analysis of ChatGPT's answers to pedodontics questions asked in the Dentistry Specialization Training Entrance Exam: Cross-sectional study. Turkiye Klinikleri J Dental Sci 2025;31(3):401-6. [In Turkish] [CrossRef]
14. Keskin A, Aygun T. A Performance of generative pre-trained transformers (GPT) in answering questions on anatomy in the Turkish dentistry specialization exam. JITSI 2024;5(4):188-92. [CrossRef]
15. Aura-Tormos JI, Llacer-Martinez M, Torres-Osca I. Educational applications of ChatGPT in university-based dental education. A systematic review. Eur J Dent Educ. 2025 Jul 3. doi: 10.1111/eje.70011. Epub ahead of print. PMID: 40609986. [CrossRef]
16. Claman D, Sezgin E. Artificial intelligence in dental education: opportunities and challenges of large language models and multimodal foundation models. JMIR Med Educ 2024;10:e52346. [CrossRef]
17. 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. [CrossRef]
18. Comanici G, Bieber E, Schaekermann M, Pasupat I, Sachdeva N, Dhillon I, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.2025. arXiv preprint arXiv:2507.06261.
19. Shrivastava PK, Rai A, Injety RJ, Singh S, Jain A, Mahuli AV, et al. Performance of ChatGPT in dentistry: A cross-sectional, multi-specialty and multi-centric study. Braz J Oral Sci 2025;24(1):e254954. [CrossRef]
20. Sabri H, Saleh MHA, Hazrati P, Merchant K, Misch J, Kumar PS, et al. Performance of three artificial intelligence (AI)-based large language models in standardized testing; implications for AI-assisted dental education. J Periodontal Res 2025;60(2):121-33. [CrossRef]
21. Tokgöz Kaplan T, Cankar M. Evidence-based potential of generative artificial intelligence large language models on dental avulsion: ChatGPT versus Gemini. Dent Traumatol 2025;41(2):178-86. [CrossRef]
22. Elnagar MH, Yadav S, Venugopalan SR, Lee MK, Oubaidin M, Rampa S, et al. ChatGPT and dental education: Opportunities and challenges. Semin Orthod 2024;30(4):401-4. [CrossRef]
23. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transact Inform Syst 2025;43(2):1-55. [CrossRef]
24. Spatharioti SE, Rothschild D, Goldstein DG, Hofman JM. Effects of LLM-based search on decision making: Speed, accuracy, and overreliance. In: CHI '25. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; 2025 April 26-May 01; Yokohama, Japan. 2025. pp.1-15. [CrossRef]
25. Memon MA. How well do large language models know dentistry? AI takes the test. Br Dent J 2025;238(1):33. [CrossRef]
26. Danesh A, Pazouki H, Danesh F, Danesh A, Vardar-Sengul S. Artificial intelligence in dental education: ChatGPT's performance on the periodontic in-service examination. J Periodontol 2024;95(7):682-7. [CrossRef]
27. Fuchs A, Trachsel T, Weiger R, Eggmann F. ChatGPT's performance in dentistry and allergyimmunology assessments: a comparative study. Swiss Dent J 2023;134(2):1-17. [CrossRef]
28. Liu M, Okuhara T, Huang W, Ogihara A, Nagao HS, Okada H, et al. Large language models in dental licensing examinations: systematic review and meta-analysis. Int Dent J 2025;75(1):213-22. [CrossRef]