BMC Medical Education

RESEARCH Open Access

# Comparative performance evaluation of ChatGPT-4 Omni and Gemini Advanced in the Turkish Dentistry Specialization Exam

Makbule Buse Dundar Sari[1] and Berkant Sezer[1*]

## Abstract

**Background** In recent years, advancements in artificial intelligence (AI) have led to the widespread integration of large language models and their chatbot applications into various fields, including dental education. This study aimed to evaluate the accuracy of ChatGPT-4 Omni (ChatGPT-4o) and Gemini Advanced in answering multiple-choice questions from the Turkish Dentistry Specialization Exams (DUS) across various disciplines.

**Methods** A total of 1,504 multiple-choice questions from 10 years of DUS exams were analyzed to compare the accuracy of ChatGPT-4o and Gemini Advanced. The questions were categorized into Fundamental Medical Sciences ($n=514$) and Clinical Dental Sciences ($n=990$). Each question was submitted to both chatbots, resulting in 3,008 responses. Accuracy was assessed using the official answer keys. Chi-square tests and Bonferroni post-hoc analyses were used to compare accuracy across disciplines and examine year-based variations.

**Results** ChatGPT-4o achieved an overall accuracy rate of 84%, while Gemini Advanced achieved 81.8% ($p=0.110$). For the Fundamental Medical Sciences questions, no statistically significant differences were observed across sub-disciplines, with overall accuracies of 92.6% for ChatGPT-4o and 93.4% for Gemini Advanced. For the Clinical Dental Sciences questions, ChatGPT-4o outperformed Gemini Advanced in Prosthetic Dentistry ($p=0.013$) and Dentomaxillofacial Radiology ($p=0.001$), whereas Gemini Advanced showed higher accuracy in Pediatric Dentistry ($p=0.008$). Across all Clinical Dental Sciences questions, ChatGPT-4o achieved an accuracy of 79.5%, compared to 75.8% for Gemini Advanced, and this difference was statistically significant ($p=0.046$).

**Conclusions** AI-based chatbots demonstrate strong potential in answering multiple-choice dentistry questions. However, variations in performance across disciplines were observed, indicating differences in accuracy depending on the subject area. These findings highlight the potential educational implications of integrating AI into dental curricula, particularly as supplementary tools for exam preparation and knowledge reinforcement. Nevertheless, cautious integration is required to ensure that AI supports, rather than replaces, critical thinking and professional expertise.

**Keywords** Artificial intelligence, Chatbot, ChatGPT, Dental education, Dentistry, Exam, Gemini, Large language models, Natural language processing

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

© 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/.

## Background

Artificial intelligence (AI) refers to computer systems and algorithms designed to simulate human intelligence. A key subfield, Natural Language Processing (NLP), enables computers to interpret and process human language in both written and spoken forms [1]. AI technologies have rapidly expanded in healthcare for their ability to analyze large datasets, detect patterns, and support decision-making in both clinical and educational contexts. Large Language Models (LLMs), deployed through AI-based chatbots, are increasingly used to provide information efficiently and assist in learning processes, particularly within medical and dental education [2].

Among the most widely used LLMs are ChatGPT-4 Omni (ChatGPT-4o) (OpenAI, San Francisco, CA, USA) and Gemini Advanced (Google LLC, Mountain View, CA, USA). ChatGPT, introduced in 2022, initiated extensive research on the applications of LLMs in medicine and dentistry, while Gemini Advanced has emerged as an alternative LLM developed by Google [3].

AI-based chatbots have also been explored in education for delivering real-time feedback, identifying knowledge gaps, and supporting personalized learning [4]. Given the specialized nature of medical and dental education, integrating AI technologies may enhance academic performance and learning efficiency [5]. However, accuracy remains dependent on the quality of training data and underlying algorithms. A key limitation is hallucination, where AI generates factually incorrect or contextually inappropriate responses, posing risks in critical decision-making [6]. Rigorous validation of AI content is therefore essential before full integration into education and assessments [7].

In Türkiye, dental specialization requires passing the Turkish Dentistry Specialization Exam (DUS), a standardized multiple-choice test covering both fundamental medical and clinical dental sciences [8]. Multiple-choice formats are widely used in licensing exams as they objectively assess knowledge across domains. Evaluating AI performance in such exams provides insight into both benchmarking accuracy and potential educational applications [9].

Although interest in AI's role in healthcare education is growing, studies focusing specifically on dental applications remain limited [10]. Previous research has often examined single specialties, providing only partial perspectives on chatbot capabilities [11, 12]. For instance, Mahmoud et al. [13] evaluated LLMs in oral and maxillofacial surgery using a specialty-focused question bank; while informative, such approaches limit generalizability. The present study broadens the scope by analyzing 10 consecutive years of the DUS (2012–2021), encompassing both fundamental medical and clinical dental sciences. This broader and longitudinal design offers novel insights into discipline-specific variations and temporal stability.

Recently, a comprehensive narrative review provided an updated synthesis of the expanding applications of ChatGPT in dentistry, covering clinical, educational, and research domains [14]. Their review highlighted opportunities such as enhanced diagnostic support, improved patient communication, and the integration of AI into curriculum design, while also emphasizing key limitations including concerns about accuracy, contextual reliability, and ethical considerations. This broader perspective underscores that the role of LLMs in dentistry extends beyond examination performance and reflects a growing need to contextualize empirical evaluations—such as the present study—within the wider landscape of AI adoption in dental practice and education.

This study aims to compare the performance of ChatGPT-4o and Gemini Advanced on DUS exams questions between 2012 and 2021. Their accuracy in fundamental medical and clinical dental sciences was examined to assess their effectiveness in responding to domain-specific questions. The null hypothesis tested was that there would be no statistically significant difference in accuracy between ChatGPT-4o and Gemini Advanced.

## Methods

### Study design and ethical considerations

This study was designed as a comparative cross-sectional analysis and adhered to the latest guidelines of the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement, as outlined by the EQUATOR Network. Because the study did not involve human or animal participants and utilized only publicly available examination materials, it qualified for exemption from Institutional Review Board (IRB) review. An official IRB exemption document was obtained. All questions and answer choices from the 2012–2021 DUS exams were retrieved verbatim from the publicly accessible official website of the Student Selection and Placement Center (ÖSYM), a government-affiliated institution (https://www.osym.gov.tr/TR,15070/dus-cikmis-sorular.html). Data collection using ChatGPT-4o and Gemini Advanced was completed in December 2024.

### Data source, inclusion, and exclusion criteria

Graduates of the five-year undergraduate dentistry program in Türkiye who wish to specialize in a specific field must take the DUS, administered by the ÖSYM. The exam was first introduced in 2012 and was offered twice a year, in the spring and fall, until 2014. Since 2015, it has been administered once annually. The exam is held on pre-announced dates at designated test centers, and based on their scores, candidates apply for specialization in various clinical disciplines at different universities

across the country. Admissions to institutions and specialties are determined by exam scores and ranking.

The DUS comprises 120 multiple-choice questions, each with five answer options and a single correct answer. Of these, 40 questions assess fundamental medical sciences, while 80 questions focus on clinical dental sciences. The distribution of the 40 fundamental medical sciences questions is as follows: 6 anatomy, 4 histology-embryology, 6 physiology, 6 biochemistry, 6 microbiology, 4 pathology, 4 pharmacology, and 4 medical biology and genetics. The 80 clinical dental sciences questions are evenly distributed among the eight clinical dentistry disciplines, with 10 questions allocated to each: restorative dentistry, prosthetic dentistry, oral and maxillofacial surgery, dentomaxillofacial radiology, periodontology, orthodontics, endodontics, and pediatric dentistry.

The dataset used in this study comprises the questions and answers from the DUS exams conducted between 2012 and 2021, obtained through ÖSYM's official open-access platform. Questions and answers from exams held after 2021 were excluded due to restricted access. The inclusion criteria involved retrieving and recording the questions and their respective answers through the specified access link. All questions were presented in their original language, Turkish, to minimize linguistic bias. Additionally, the correct answers were determined based on the officially announced answers by ÖSYM.

The exclusion criteria for the study were as follows: (1) Questions officially annulled by ÖSYM due to errors, multiple correct answers, or the absence of a correct answer; (2) Questions containing figures, images, or graphs. The exclusion of visual content-based questions was necessary because current AI models, including ChatGPT-4o and Gemini Advanced, primarily rely on text-based processing and lack the capability to analyze complex images effectively. Previous studies have demonstrated that while LLMs perform well on text-based multiple-choice questions, their accuracy significantly declines when interpreting figures or radiographic images, which require spatial reasoning and domain-specific visual expertise [15, 16].

### Primary and secondary outcomes

The primary outcome was the accuracy rate (%), defined as the proportion of correct answers out of the total number of questions. This measure was used to compare the performance of OpenAI's ChatGPT-4o and Google's Gemini Advanced in answering the DUS exams questions across different disciplines and overall. The secondary outcomes included comparisons of accuracy based on the year of the exam and the distribution of correct and incorrect responses. Correct and incorrect answers were presented as n (number) and % (percentage). The accuracy performance of the chatbots was evaluated using the following formula: Performance (%) = (Total number of correct answers/Total number of questions) × 100.

### AI-based chatbot testing: prompt engineering and input encoding

The latest premium version of ChatGPT-4o and Gemini Advanced was used in this study. To ensure procedural consistency and minimize bias, the same investigator entered all multiple-choice questions and answer choices into both chatbots via copy-paste, and the responses were recorded in a standardized Microsoft Excel template designed for reproducibility. No additional instructions, prompts, or guiding phrases were provided (e.g., no statements such as "select one correct answer"), and each question was presented exactly as it appeared in the DUS database. To minimize potential order effects, all questions were randomized once, and the same randomized sequence was used independently for both models. Additionally, a new chat session was initiated for each question to prevent memory retention or crossover learning effects.

Although the investigator entering the questions could not be blinded due to visible model interfaces, the researcher responsible for data analysis remained fully blinded to the model assignments. Each question was submitted three times to each model independently, and the final answer was determined using a 2/3 consistency rule to ensure response stability and minimize random variation. In rare cases of persistent uncertainty (< 2% of all questions), additional regenerations were performed until at least 85% consistency was achieved. The 85% threshold was chosen to reduce random variation and ensure response stability, consistent with reproducibility criteria applied in previous LLMs-based studies [17, 18].

### Statistical analysis

Descriptive statistics were summarized using frequencies, and percentages were reported for variables. The Chi-square test was used to compare the performance of ChatGPT-4o and Gemini Advanced models. To minimize bias, the researcher responsible for data analysis was blinded to the AI model assignments. Statistical analyses were carried out using IBM SPSS Statistics (Version 26, IBM Corp., Armonk, NY, USA). A $p$-value of 0.05 was set as the threshold for statistical significance, and for all post-hoc pairwise comparisons Bonferroni correction was consistently applied. The Bonferroni-adjusted significance threshold was calculated using the formula 0.05/n, where n represents the number of pairwise comparisons performed. Unadjusted p-values are presented in the tables; however, only those meeting the Bonferroni-adjusted significance thresholds were interpreted as statistically significant. In addition, 95% confidence intervals (CIs) for accuracy rates were calculated using the Wilson

method to provide estimates of uncertainty and to enable clearer interpretation of the results.

## Results

A total of 1,560 questions were administered in the DUS exams between 2012 and 2021. After excluding 12 questions officially annulled by ÖSYM and 44 questions containing visual content, the final dataset comprised 1,504 questions.

Table 1 summarizes the results for fundamental medical sciences, showing that both models achieved consistently high accuracy rates above 90%, with no significant differences across disciplines. Gemini Advanced achieved an accuracy rate above 90% in all subfields, while ChatGPT-4o exceeded 90% in all but three areas: Anatomy, Histology-Embryology, and Physiology. Notably, ChatGPT-4o answered all Biochemistry questions correctly. When all fundamental medical sciences questions were considered, ChatGPT-4o demonstrated an accuracy rate of 92.6% (95% CI: 90–94.6), and Gemini Advanced achieved 93.4% (95% CI: 90.9–95.2). The difference between the models was not statistically significant ($p = 0.625$).

Table 2 highlights the outcomes for clinical dental sciences, where overall accuracy levels were lower than in fundamental sciences and discipline-specific differences were more pronounced. Gemini Advanced outperformed ChatGPT-4o in Endodontics and Pediatric Dentistry, while ChatGPT-4o showed superior accuracy in all remaining disciplines. Statistically significant results were found in favor of ChatGPT-4o in Prosthetic Dentistry ($p = 0.013$), where its accuracy was 10.2% points higher, and in Dentomaxillofacial Radiology ($p = 0.001$), with a 15.1% advantage. Conversely, Gemini Advanced had significantly higher accuracy in Pediatric Dentistry ($p = 0.008$), outperforming ChatGPT-4o by 12.4%. Overall, ChatGPT-4o achieved a significantly higher total accuracy in clinical dental sciences at 79.5% (95% CI: 76.9–81.9) compared to Gemini Advanced at 75.8% (95% CI: 73–78.3) ($p = 0.046$), corresponding to an absolute difference of 3.7%.

**Table 1** Correct and incorrect answers provided by ChatGPT-4o and Gemini Advanced for Turkish Dentistry Specialization Exam (DUS) questions in fundamental medical sciences and disciplines

| | Total Answers | | ChatGPT-4o Answers | | Gemini Advanced Answers | | |
|---|---|---|---|---|---|---|---|
| | *N* (%) | 95% CI | *N* (%) | 95% CI | *N* (%) | 95% CI | *p*-value* |
| Anatomy | | | | | | | |
| *Correct* | 130 (89) | 82.9–93.1 | 62 (84.9) | 75.0–91.4.0.4 | 68 (93.2) | 84.9–97 | 0.112 |
| *Incorrect* | 16 (11) | 6.9–17.1 | 11 (15.1) | 8.6–25 | 5 (6.8) | 3–15.1.1 | |
| Histology-Embryology | | | | | | | |
| *Correct* | 94 (90.4) | 83.2–94.7 | 45 (86.5) | 74.7–93.3 | 49 (94.2) | 84.4–98 | 0.183 |
| *Incorrect* | 10 (9.6) | 5.3–16.8 | 7 (13.5) | 6.7–25.3 | 3 (5.8) | 2–15.6.6 | |
| Physiology | | | | | | | |
| *Correct* | 139 (90.3) | 84.6–94 | 67 (87) | 77.7–92.8 | 72 (93.5) | 85.7–97.2 | 0.174 |
| *Incorrect* | 15 (9.7) | 6–15.4.4 | 10 (13) | 7.2–22.3 | 5 (6.5) | 2.8–14.3 | |
| Biochemistry | | | | | | | |
| *Correct* | 152 (97.4) | 93.6–99 | 78 (100) | 95.3–100 | 74 (94.5) | 87.5–98 | 0.120 |
| *Incorrect* | 4 (2.6) | 1–6.4.4 | 0 (0) | 0–4.7.7 | 4 (5.5) | 2–12.5.5 | |
| Microbiology | | | | | | | |
| *Correct* | 149 (95.5) | 91–97.8.8 | 75 (96.2) | 89.3–98.7 | 74 (94.5) | 87.5–98 | 1.000 |
| *Incorrect* | 7 (4.5) | 2.2–9.2 | 3 (3.8) | 1.3–10.7 | 4 (5.5) | 2–12.5.5 | |
| Pathology | | | | | | | |
| *Correct* | 99 (95.2) | 89.2–97.9 | 50 (96.2) | 87–98.9.9 | 49 (94.2) | 84.4–98 | 1.000 |
| *Incorrect* | 5 (4.8) | 2.1–10.8 | 2 (3.8) | 1.1–13 | 3 (5.8) | 2–15.6.6 | |
| Pharmacology | | | | | | | |
| *Correct* | 95 (91.3) | 84.4–95.4 | 48 (92.3) | 81.8–97 | 47 (90.4) | 79.4–95.8 | 1.000 |
| *Incorrect* | 9 (8.7) | 4.6–15.6 | 4 (7.7) | 3–18.2.2 | 5 (9.6) | 4.2–20.6 | |
| Medical Biology and Genetics | | | | | | | |
| *Correct* | 98 (93.3) | 88–97.3.3 | 51 (98.1) | 89–99.7.7 | 47 (90.4) | 79.4–95.8 | 0.205 |
| *Incorrect* | 6 (6.7) | 2.7–12 | 1 (1.9) | 0.3–10.1 | 5 (9.6) | 4.2–20.6 | |
| Fundamental Medical Sciences – Total | | | | | | | |
| *Correct* | 956 (93) | 91.3–94.4 | 476 (92.6) | 90–94.6.6 | 480 (93.4) | 90.9–95.2 | 0.625 |
| *Incorrect* | 72 (7) | 5.6–8.7 | 38 (7.4) | 5.4–10 | 34 (6.6) | 4.8–9.1 | |

*N* Number, *CI* Confidence Interval

*Chi-square test

**Table 2** Correct and incorrect answers provided by ChatGPT-4o and Gemini Advanced for Turkish Dentistry Specialization Exam (DUS) questions in clinical dental sciences and disciplines

| | Total Answers | | ChatGPT-4o Answers | | Gemini Advanced Answers | | |
|---|---|---|---|---|---|---|---|
| | *N* (%) | 95% CI | *N* (%) | 95% CI | *N* (%) | 95% CI | *p*-value* |
| Restorative Dentistry | | | | | | | |
| *Correct* | 229 (90.2) | 85.9–93.2 | 117 (92.1) | 86.1–95.7 | 112 (88.2) | 81.4–92.7 | 0.292 |
| *Incorrect* | 25 (9.8) | 6.8–14.1 | 10 (7.9) | 4.3–13.9 | 15 (11.8) | 7.3–18.6 | |
| Prosthetic Dentistry | | | | | | | |
| *Correct* | 159 (63.1) | 57–68.8.8 | 89 (70.7) | 62.2–77.9 | 70 (55.6) | 46.8–63.9 | **0.013** |
| *Incorrect* | 93 (36.9) | 31.2–43 | 37 (29.3) | 22.1–37.8 | 56 (44.4) | 36.1–53.2 | |
| Oral and Maxillofacial Surgery | | | | | | | |
| *Correct* | 212 (86.2) | 81.3–89.9 | 111 (90.2) | 83.7–94.3 | 101 (82.1) | 74.4–87.9 | 0.065 |
| *Incorrect* | 34 (13.8) | 10.1–18.7 | 12 (9.8) | 5.7–16.3 | 22 (17.9) | 12.1–25.6 | |
| Dentomaxillofacial Radiology | | | | | | | |
| *Correct* | 207 (84.1) | 79.1–88.2 | 113 (91.9) | 85.7–95.5 | 94 (76.4) | 68.2–83.1 | **0.001** |
| *Incorrect* | 39 (15.9) | 11.8–20.9 | 10 (9.1) | 4.5–14.3 | 29 (23.6) | 16.9–31.8 | |
| Periodontology | | | | | | | |
| *Correct* | 201 (79.1) | 73.7–83.7 | 102 (80.3) | 72.6–86.3 | 99 (78) | 70–84.3.3 | 0.643 |
| *Incorrect* | 53 (20.9) | 16.3–26.3 | 25 (19.7) | 13.7–27.4 | 28 (22) | 15.7–30 | |
| Orthodontics | | | | | | | |
| *Correct* | 169 (71.6) | 65.5–77 | 87 (73.7) | 65.1–80.8 | 82 (69.5) | 60.7–77.1 | 0.470 |
| *Incorrect* | 67 (28.4) | 23–34.5.5 | 31 (26.3) | 19.2–34.9 | 36 (30.5) | 22.9–39.3 | |
| Endodontics | | | | | | | |
| *Correct* | 163 (66.8) | 60.7–72.4 | 78 (64.5) | 55.1–71.9 | 85 (69.7) | 61–77.1.1 | 0.341 |
| *Incorrect* | 81 (33.2) | 27.6–39.3 | 44 (35.5) | 28.1–44.9 | 37 (30.3) | 22.9–39 | |
| Pediatric Dentistry | | | | | | | |
| *Correct* | 197 (79.4) | 74–84 | 90 (72.6) | 64.1–79.7 | 107 (86.3) | 79.1–91.3 | **0.008** |
| *Incorrect* | 51 (20.6) | 16–26 | 34 (27.4) | 20.3–35.9 | 17 (13.7) | 8.7–20.9 | |
| Clinical Dental Sciences – Total | | | | | | | |
| *Correct* | 1537 (77.6) | 75.7–79.4 | 787 (79.5) | 76.9–81.9 | 750 (75.8) | 73.0–78.3.0.3 | **0.046** |
| *Incorrect* | 443 (22.4) | 20.6–24.3 | 203 (20.5) | 18.1–23.1 | 240 (24.2) | 18.1–23.1 | |

Bold values indicate statistically significance

*N* Number, *CI* Confidence Interval

*Chi-square test

**Table 3** Correct and incorrect answers provided by ChatGPT-4o and Gemini Advanced for all Turkish Dentistry Specialization Exam (DUS) questions

| | Total Answers | | ChatGPT-4o Answers | | Gemini Advanced Answers | | *p*-value* |
|---|---|---|---|---|---|---|---|
| | *N* (%) | 95% CI | *N* (%) | 95% CI | *N* (%) | 95% CI | |
| All Questions | | | | | | | |
| *Correct* | 2493 (82.9) | 81.5–84.2 | 1263 (84) | 82–85.7.7 | 1230 (81.8) | 79.8–83.7 | 0.110 |
| *Incorrect* | 515 (17.1) | 15.8–18.5 | 241 (16) | 14.3–18 | 274 (18.2) | 16.3–20.2 | |

*N* Number, *CI* Confidence Interval

*Chi-square test

Table 3 provides the overall performance across all disciplines, showing that both chatbots maintained accuracy levels above 80% with no significant overall difference. When considering all the Fundamental Medical Sciences and Clinical Dental Sciences questions, the accuracy performance of both chatbots was compared across a total of 1,504 questions. No statistically significant difference was found between the accuracy performance of the two chatbots ($p = 0.110$), with ChatGPT-4o achieving 84% (95% CI: 82–85.7) and Gemini Advanced achieving 81.8% (95% CI: 79.8–83.7), both exceeding 80% accuracy.

Table 4 presents the year-based analysis, where the models showed fluctuations in performance, although none reached statistical significance. When performance was analyzed by exam year, Gemini Advanced showed higher accuracy in 2016 and 2021, whereas ChatGPT-4o outperformed in all other years. Nevertheless, no statistically significant difference was found in year-based performance comparisons (2012: $p = 0.898$; 2013: $p = 0.716$;

**Table 4** Correct and incorrect answers provided by ChatGPT-4o and Gemini Advanced for Turkish Dentistry Specialization Exam (DUS) questions in different years

| | Total Answers *N* (%) | 95% CI | ChatGPT-4o Answers *N* (%) | 95% CI | Gemini Advanced Answers *N* (%) | 95% CI | *p*-value* |
|---|---|---|---|---|---|---|---|
| 2012 | | | | | | | |
| *Correct* | 405 (85.1) | 81.6–88 | 203 (85.3) | 80.2–89.2 | 202 (84.9) | 79.8–88.9 | 0.898 |
| *Incorrect* | 71 (14.9) | 12–18.4.4 | 35 (14.7) | 10.8–19.8 | 36 (15.1) | 11.1–20.2 | |
| 2013 | | | | | | | |
| *Correct* | 381 (82.1) | 78.4–85.3 | 192 (82.8) | 77.4–87.1 | 189 (81.5) | 75.9–85.9 | 0.716 |
| *Incorrect* | 83 (17.9) | 14.7–21.6 | 40 (17.2) | 12.9–22.6 | 43 (18.5) | 14.1–24 | |
| 2014 | | | | | | | |
| *Correct* | 400 (84.4) | 80.8–87.4 | 202 (85.2) | 80.2–89.2 | 198 (83.5) | 78.3–87.7 | 0.613 |
| *Incorrect* | 74 (15.6) | 12.6–19.2 | 35 (14.8) | 10.8–19.9 | 39 (16.5) | 12.3–21.7 | |
| 2015 | | | | | | | |
| *Correct* | 201 (86.6) | 81.7–90.4 | 105 (90.5) | 83.8–94.6 | 96 (82.8) | 74.9–88.6 | 0.082 |
| *Incorrect* | 31 (13.4) | 9.6–18.3 | 11 (9.5) | 5.4–16.2 | 20 (17.2) | 11.4–25.1 | |
| 2016 | | | | | | | |
| *Correct* | 205 (85.4) | 80.4–89.3 | 102 (85) | 77.5–90.3 | 103 (85.8) | 78.5–90.9 | 0.855 |
| *Incorrect* | 35 (14.6) | 10.7–19.6 | 18 (15) | 9.7–22.5 | 17 (14.2) | 9–21.5.5 | |
| 2017 | | | | | | | |
| *Correct* | 201 (83.8) | 78.9–87.9 | 96 (85.7) | 78–91.1.1 | 105 (82) | 74.4–87.9 | 0.440 |
| *Incorrect* | 39 (16.2) | 12.1–21.1 | 16 (14.3) | 8.9–22 | 23 (18) | 12.1–25.6 | |
| 2018 | | | | | | | |
| *Correct* | 162 (75) | 68.3–80.7 | 82 (75.9) | 67.7–82.5 | 80 (74.1) | 65.9–81 | 0.753 |
| *Incorrect* | 54 (25) | 19.3–31.7 | 26 (24.1) | 17.5–32.3 | 28 (25.9) | 19–34.1.1 | |
| 2019 | | | | | | | |
| *Correct* | 168 (76.4) | 69.6–82 | 88 (80) | 71.4–86.4 | 80 (72.7) | 63.9–80 | 0.204 |
| *Incorrect* | 52 (23.6) | 18–30.4.4 | 22 (20) | 13.6–28.6 | 30 (27.2) | 20–36.1.1 | |
| 2020 | | | | | | | |
| *Correct* | 200 (84) | 78–88.7.7 | 101 (84.9) | 77–90.4.4 | 99 (83.2) | 75.2–89.1 | 0.723 |
| *Incorrect* | 38 (16) | 11.3–22 | 18 (15.1) | 9.6–23 | 20 (16.8) | 10.9–24.8 | |
| 2021 | | | | | | | |
| *Correct* | 186 (83) | 76.8–87.9 | 92 (82.1) | 73.3–88.4 | 94 (83.9) | 75.3–89.9 | 0.722 |
| *Incorrect* | 38 (17) | 12.1–23.2 | 20 (17.9) | 11.6–26.7 | 18 (16.1) | 10.1–24.7 | |

*N* Number, *CI* Confidence Interval

*Chi-square test

2014: $p = 0.613$; 2015: $p = 0.082$; 2016: $p = 0.855$; 2017: $p = 0.440$; 2018: $p = 0.753$; 2019: $p = 0.204$; 2020: $p = 0.723$; 2021: $p = 0.722$). Figure 1 illustrates year-to-year accuracy trends for ChatGPT-4o, Gemini Advanced, and combined performance.

Table 5 reports the post-hoc Bonferroni analysis, highlighting discipline-specific differences in the distribution of correct and incorrect answers. For ChatGPT-4o, the highest number of incorrect responses was observed in Endodontics, whereas no incorrect answers were recorded in Biochemistry. For Gemini Advanced, the highest proportion of incorrect responses occurred in Prosthetic Dentistry, while the lowest frequencies were noted in Histology-Embryology and Pathology.

Table 6 summarizes the distribution of correct and incorrect answers across the two main categories, showing that both models more accurately in Fundamental Medical Sciences and had a higher proportion of incorrect responses in Clinical Dental Sciences.

## Discussion

The accuracy and comprehensiveness of responses generated by LLMs are increasingly being investigated in scientific research. Studies on continuously updated AI models constitute a significant area of inquiry in this field. In this study, ChatGPT-4o and Gemini Advanced achieved overall accuracies above 80% when tested on 1,504 questions from the DUS. Discipline-specific differences were evident, with ChatGPT-4o demonstrating superior performance in Prosthetic Dentistry and Dentomaxillofacial Radiology, whereas Gemini Advanced outperformed in Pediatric Dentistry. Year-based analysis further revealed stable performance across a decade, underscoring the reliability of the models over time.

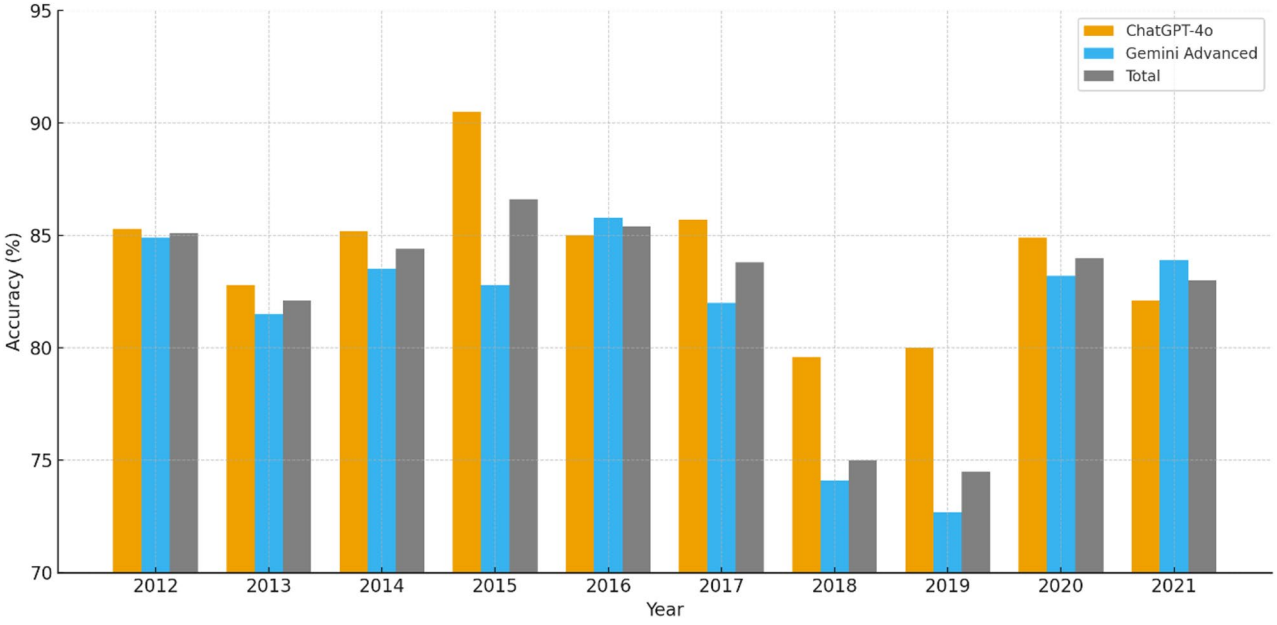



**Fig. 1** Year-based accuracy rates of ChatGPT-4o, Gemini Advanced, and overall performance in the Turkish Dentistry Specialization Exams (2012–2021) (DUS)

**Table 5** Distribution of correct and incorrect answers provided by ChatGPT-4o and gemini advanced to Turkish dentistry specialization exam (DUS) questions among different branches

| | **ChatGPT-4o Answers** | | | **Gemini Advanced Answers** | | |
|---|---|---|---|---|---|---|
| | **Incorrect** ***N*** **(%)** | **Correct** ***N*** **(%)** | ***p*-value*** | **Incorrect** ***N*** **(%)** | **Correct** ***N*** **(%)** | ***p*-value*** |
| Anatomy | 11 (4.6) | 62 (4.9) | **< 0.001** | 5 (1.8) | 68 (5.5) | **< 0.001** |
| Histology-Embryology†‡ | 7 (2.9) | 45 (3.6) | | 3 (1.1) | 49 (4) | |
| Physiology ‡ | 10 (4.1) | 67 (5.3) | | 5 (1.8) | 72 (5.9) | |
| Biochemistry | 0 (0) | 78 (6.2) | | 4 (1.5) | 74 (6) | |
| Microbiology † | 3 (1.2) | 75 (5.9) | | 4 (1.5) | 74 (6) | |
| Pathology ‡ | 2 (0.8) | 50 (4) | | 3 (1.1) | 49 (4) | |
| Pharmacology † | 4 (1.7) | 48 (3.8) | | 5 (1.8) | 47 (3.8) | |
| Medical Biology and Genetics | 1 (0.4) | 51 (4) | | 5 (1.8) | 47 (3.8) | |
| Restorative Dentistry ‡ | 10 (4.1) | 117 (9.3) | | 15 (5.5) | 112 (9.1) | |
| Prosthetic Dentistry ‡ | 37 (15.4) | 89 (7) | | 56 (20.4) | 70 (5.7) | |
| Oral and Maxillofacial Surgery | 12 (5) | 111 (8.8) | | 22 (8) | 101 (8.2) | |
| Dentomaxillofacial Radiology | 10 (4.1) | 113 (8.9) | | 29 (10.6) | 94 (7.6) | |
| Periodontology †‡ | 25 (10.4) | 102 (8.1) | | 28 (10.2) | 99 (8) | |
| Orthodontics †‡ | 31 (12.9) | 87 (6.9) | | 36 (13.1) | 82 (6.7) | |
| Endodontics † | 44 (18.3) | 78 (6.2) | | 37 (13.5) | 85 (6.9) | |
| Pediatric Dentistry | 34 (14.1) | 90 (7.1) | | 17 (6.2) | 107 (8.7) | |
| Total | 241 (100) | 1263 (100) | | 274 (100) | 1230 (100) | |

Bold values indicate statistical significance. Unadjusted *p*-values are shown; statistical significance was determined using Bonferroni-adjusted thresholds ($p < 0.001$ for 66 pairwise comparisons)

*N* Number

*Chi-square test with post-hoc Bonferroni test

†Significant after Bonferroni correction for ChatGPT-4o

‡Significant after Bonferroni correction for Gemini Advanced

Consequently, the null hypothesis tested in this study was partially rejected.

ChatGPT and Gemini differ in their AI architectures and data processing capabilities. Due to its updated version, ChatGPT can now access real-time online data through web browsing, providing more up-to-date information. While this facilitates access to scientific information in fields such as dentistry, it is crucial to carefully

**Table 6** Distribution of correct and incorrect answers provided by ChatGPT-4o and Gemini Advanced to Turkish Dentistry Specialization Exam (DUS) questions among main disciplines

| | ChatGPT-4o Answers | | | Gemini Advanced Answers | | |
|---|---|---|---|---|---|---|
| | Incorrect *N* (%) | Correct *N* (%) | *p*-value* | Incorrect *N* (%) | Correct *N* (%) | *p*-value* |
| Fundamental Medical Sciences | 38 (15.8) | 476 (37.7) | **< 0.001** | 34 (12.4) | 480 (39) | **< 0.001** |
| Clinical Dental Sciences | 203 (84.2) | 787 (62.3) | | 240 (87.6) | 750 (61) | |
| Total | 241 (100) | 1263 (100) | | 274 (100) | 1230 (100) | |

Bold values indicate statistical significance. Unadjusted *p*-values are shown; since only two groups were compared, the Bonferroni adjustment was not required and the significance threshold remained at $p < 0.05$

*N* Number

*Chi-square test with post-hoc Bonferroni test

evaluate both models in terms of accuracy and reliability [19]. The findings of this study highlight the potential applications of AI-based chatbots in dentistry, as well as their strengths and limitations.

AI-based LLMs such as ChatGPT and Gemini were not originally designed for medical or dental problem-solving, and this may limit their ability to interpret complex clinical relationships [20]. Nevertheless, several studies have shown that ChatGPT models can achieve high accuracy in medical and dental examinations [21]. For example, Mackey et al. reported an accuracy of 89% for ChatGPT-4o across multiple U.S. clinical subject exams, although performance varied by specialty [22]. Uehara et al. similarly demonstrated that ChatGPT-3.5 and ChatGPT-4o performed well in the Japanese National Dental Examination, particularly in basic medical sciences such as pathology, pharmacology, and microbiology [23]. Chau et al. further confirmed this pattern, showing that ChatGPT-4o reached passing thresholds in both United States and United Kingdom dental licensing exams, while ChatGPT-3.5 did not [24]. Other studies, however, have highlighted variability between models and contexts. Lin et al. compared ChatGPT-3.5, Gemini, and Claude2 in the Taiwanese National Dental Licensing Examination and reported moderate accuracy, with Claude2 outperforming other models at 54.8% [25]. In Türkiye, an dentomaxillofacial radiology–focused study found ChatGPT-4o to be the most accurate model (86.1%), followed by Bard (61.8%), ChatGPT-3.5 (43.9%), and Copilot (41.5%) [11]. Further cross-national evaluations, including the United States Medical Licensing Examination (USMLE), Professional and Linguistic Assessments Board (PLAB), Hong Kong, and Chinese medical licensing exams, consistently showed GPT-4o outperforming other LLMs [26]. Conversely, Avşar et al. found that while ChatGPT-3.5 and Gemini displayed comparable knowledge in prosthodontics-related Turkish exams, their accuracy remained limited [27].

Recent literature also highlights the advantages of newer ChatGPT versions. Danesh et al. reported that ChatGPT-4o produced more accurate dental content than ChatGPT-3.5 in the Integrated National Board Dental Examination [28]. Similarly, Meyer et al. found that GPT-4 achieved an average success rate of 85% in German medical licensing exams, whereas GPT-3.5 passed only one of three tests [29]. In the European Association for Osseointegration Implant Dentistry Certification Examination, ChatGPT-4o not only passed the exam but also outperformed both ChatGPT-3.5 and licensed dentists [30]. Collectively, these findings indicate that recent LLM versions benefit from improved language processing capabilities and expanded training datasets. This rationale supported our selection of ChatGPT-4o for the present study, where it achieved an accuracy rate of 84%. Accordingly, the choice of model should be viewed not merely as a comparison with earlier versions but as an assessment of the academic applicability of the most advanced AI-based systems available.

In the study by Sismanoglu and Capan, ChatGPT-4o outperformed Gemini Advanced in DUS exams from 2020 to 2021 [31]. In our study, which examined all DUS exams from 2012 to 2021, the models' performances remained generally stable over time, although notable differences emerged across specific disciplines. Such discipline-based variations may reflect the diverse nature of dental specialties, which range from protocol-driven clinical procedures to complex decision-making and developmental or preventive frameworks. These differences likely influence how effectively LLMs process and reproduce accurate information. Additionally, the representation of specialty-specific terminology and concepts within training corpora may contribute to performance variability across fields [32].

Recent Turkish studies have also evaluated LLM performance in specific DUS disciplines. Gokkurt Yilmaz et al. [33] reported that LLM accuracy varied considerably in head and neck anatomy questions, with ChatGPT-o1 outperforming Gemini 1.5 and several other models. In a separate oral pathology analysis, the authors found marked performance differences, with ChatGPT-o1 achieving the highest accuracy and Copilot the lowest [34]. These findings reinforce that LLM performance can fluctuate substantially across dental specialties. The present study aligns with this pattern by demonstrating

discipline-specific differences between ChatGPT-4o and Gemini Advanced, while extending previous research through its broader, decade-long evaluation across all DUS disciplines.

For example, Gemini Advanced's higher performance in Pediatric Dentistry may reflect broader representation of preventive and developmental pediatric topics in its training data, whereas ChatGPT-4o's strengths in Prosthetic Dentistry and Dentomaxillofacial Radiology may be linked to better optimization for structured reasoning and clinically oriented content. Although some of these differences reached statistical significance, their absolute magnitudes were modest, limiting their practical relevance in most educational or clinical contexts. Nevertheless, even small differences may have implications in high-stakes examination scenarios, particularly for examinees near the pass–fail threshold. Overall, the observed disparities likely stem from uneven distribution of specialty-specific knowledge and varying levels of reasoning complexity across disciplines, rather than the inherent superiority of one model over another. The stable year-to-year performance further suggests that model accuracy does not decline over time, which may relate to fluctuations in exam difficulty as well as ongoing model updates and learning processes.

The findings of this study have important implications for dental education and clinical practice. The high accuracy of advanced LLMs suggests that they may serve as supplementary tools for exam preparation by providing rapid, structured information. However, their higher error rates in clinically oriented subjects highlight the need for cautious and supervised use, particularly in areas requiring complex reasoning and contextual interpretation.

From a clinical perspective, the results emphasize that AI-based chatbots may assist practitioners in efficiently retrieving knowledge and summarizing clinical guidelines, but their limitations necessitate that clinical judgment and evidence-based decision-making remain central. Thus, while LLMs hold promise as supportive resources, they should be regarded as complementary to—rather than replacements for—formal education and professional expertise.

The stable performance of ChatGPT-4o and Gemini Advanced over a decade also indicates potential value in curriculum design, helping students identify knowledge gaps and strengthening self-assessment. Nonetheless, discipline-specific weaknesses emphasize the importance of integrating AI literacy and critical appraisal skills into dental education.

The integration of AI into dental education and assessments raises interconnected concerns related to ethics, equity, and academic integrity. While incorporating AI literacy into dental curricula may help students use these tools responsibly and balance efficiency with critical thinking, unresolved risks remain. Excessive reliance on chatbots can hinder the development of independent problem-solving skills, and unequal access to advanced AI systems may deepen educational disparities across institutions. Recent literature highlights these issues: Schwendicke et al. emphasized the need for governance, accountability, and explainability in AI-supported learning [35]; McGrath et al. noted risks linked to algorithmic bias, reduced human interaction, and the digital divide [36]; and Uribe et al. reported widespread concerns among educators about unclear implementation frameworks and potential over-dependence on AI-based decision support [37]. These concerns extend directly to high-stakes examinations, where the increasing availability of AI tools poses risks to exam security, as students may exploit chatbots during assessments, potentially compromising the validity and fairness of results. Together, these considerations underscore the need for carefully designed guidelines that safeguard equity and academic integrity while supporting responsible AI integration in dental education.

Beyond individual misuse, institutions may also face challenges in ensuring that assessment formats remain resistant to unauthorized AI support. Such risks highlight the necessity of developing robust exam protocols and integrity safeguards alongside the pedagogical integration of AI. Without clear guidelines and monitoring strategies, there is a danger that the benefits of these technologies could be overshadowed by compromised fairness and credibility in dental education.

Unlike international licensing examinations such as the Integrated National Board Dental Examination (INBDE) or the USMLE, the DUS does not operate on a pass/fail basis; instead, candidates are ranked by their scores and residency positions are allocated according to available quotas. This structural difference makes direct benchmarking challenging. Nevertheless, comparison with international examinations provides useful context. Data from the INBDE show that overall failure rates declined from 24.5% in 2020 to 8.7% in 2023, before rising again to 16.1% in 2024 following changes in exam standards [38]. Similarly, the USMLE reports average pass rates of approximately 88–90% for Step 1, 86–97% for Step 2, and 85–97% for Step 3, depending on the examinee group and year [39].

When viewed against these human benchmarks, the accuracy rates observed in our study—84% for ChatGPT-4o and 82% for Gemini Advanced—suggest that AI models perform at levels comparable to or even exceeding the minimum passing performance seen in major international licensing exams. However, it is crucial to emphasize that the qualitative nature of AI errors, particularly in clinically oriented specialties, limits their suitability as

replacements for human decision-making despite these comparable scores.

The unique contribution of the present study lies in its comprehensive analysis of a large-scale dataset covering 10 consecutive years of the DUS exams (2012–2021), encompassing 1,504 multiple-choice questions across both fundamental medical sciences and clinical dental sciences. Unlike prior studies that examined AI performance in isolated national exams or limited subject domains, our work provides a longitudinal, discipline-specific evaluation of chatbot accuracy over a decade. This robust dataset enables a more reliable assessment of model stability and highlights variations across clinical specialties that may not be evident in shorter-term or single-exam studies.

One of the main strengths of this study is the ability to track the models' performances over a prolonged period, which allows for a more nuanced evaluation of temporal stability. By integrating both fundamental medical sciences and clinical dental sciences, the study also provides a broad perspective on discipline-specific differences that could otherwise be overlooked. Moreover, the large sample size reduces the likelihood of bias stemming from single-year fluctuations or limited question pools, thereby strengthening the reliability of the findings.

A key limitation of this study is that it only examines Turkish text-based questions, which may restrict the generalizability of the findings. AI models can demonstrate varying accuracy depending on the language of the input, and the style and structure of national dental examinations may also differ across countries, potentially influencing model performance. Furthermore, the exclusion of image-based questions—particularly radiology-heavy content such as dentomaxillofacial radiology—represents a major limitation that substantially reduces the real-world applicability of the findings.

Since visual interpretation is integral to dental education and clinical examinations, this omission limits the extent to which our results can be extrapolated to actual practice. Future work should explore the use of multimodal AI systems, such as GPT-4o with integrated vision capabilities, which may allow the inclusion of figure- and radiology-based questions and enable a more comprehensive evaluation of AI performance. Another limitation is that the performance of AI models was not directly compared with that of human test-takers, which would have provided a clearer benchmark for contextualizing their relative accuracy. Recent literature has also highlighted the importance of incorporating human benchmarks and multimodal evaluation when assessing AI's role in dental examinations [40, 41].

Additionally, AI models are continuously updated, and the performance observed in this study may not fully represent future versions. Future research should focus on evaluating AI models in case-based clinical assessments and exploring their integration into dental education. Furthermore, assessing AI models' performances in different languages could enhance the generalizability of results. Discipline-specific improvements in AI training and optimization could help mitigate the disparities observed across different areas of dentistry. Finally, balancing the distribution of training data across disciplines could contribute to the development of more consistent AI models for dental education, and these insights collectively emphasize the need for continuous evaluation and responsible integration of AI in dental education.

## Conclusion

This study demonstrated that both ChatGPT-4o and Gemini Advanced provide highly accurate responses to multiple-choice dentistry questions, with discipline-specific variations. While no significant difference was observed between the two models in fundamental medical sciences, their performance varied in clinical dental sciences. ChatGPT-4o outperformed Gemini Advanced in Prosthetic Dentistry and Dentomaxillofacial Radiology, whereas Gemini Advanced performed better in Pediatric Dentistry. These findings highlight the need for careful consideration when applying AI models in clinically oriented subjects, where accurate decision-making is crucial. AI-supported educational tools should be optimized to ensure improved performance across all dental disciplines, particularly in complex clinical scenarios. These findings underscore the importance of integrating AI into dental education and assessments through discipline-specific validation, continuous performance monitoring, and the development of tailored training datasets to maximize reliability and educational value, and future studies should continue longitudinal evaluations to guide responsible and equitable AI integration into dental education.

**Acknowledgements**
Not applicable.

**Clinical trial number**
Not applicable.

**Authors' contributions**
Conceptualization: M.B.D.S. and B.S. Methodology: M.B.D.S. and B.S. Formal analysis: B.S. Writing – original draft preparation: M.B.D.S. and B.S. Writing – review and editing: B.S. Supervision and integrity of the data and analysis: B.S.

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

**Data availability**
The data that support the findings of this study are available from the corresponding author (B.S.) upon reasonable requests.

## Declarations

**Ethical approval and consent to participate**
This study did not involve human or animal subjects, and therefore, ethical review and approval were not required.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

Received: 12 May 2025 / Accepted: 12 January 2026
Published online: 17 January 2026

**References**

1. Acharya S, Godhi BS, Saxena V, Assiry AA, Alessa NA, Dawasaz AA, Alqarni A, Karobari MI. Role of artificial intelligence in behavior management of pediatric dental patients-a mini review. J Clin Pediatr Dent. 2024;48:24–30. https://doi.org/10.22514/jocpd.2024.055.
2. Sezer B, Aydoğdu T. Performance of advanced artificial intelligence models in pulp therapy for immature permanent teeth: a comparison of ChatGPT-4 Omni, DeepSeek, and Gemini Advanced in accuracy, completeness, response time, and readability. J Endod. 2025;51:1675–84. https://doi.org/10.1016/j.joen.2025.08.011.
3. Sezer B, Aydoğdu T. Performance of advanced artificial intelligence models in traumatic dental injuries in primary dentition: a comparative evaluation of ChatGPT-4 Omni, DeepSeek, Gemini Advanced, and Claude 3.7 in terms of accuracy, completeness, response time, and readability. Appl Sci. 2025;15:7778. https://doi.org/10.3390/app15147778.
4. Fang Q, Reynaldi R, Araminta AS, Kamal I, Saini P, Afshari FS, Tan SC, Yuan JC, Qomariyah NN, Sukotjo C. Artificial intelligence (AI)-driven dental education: exploring the role of chatbots in a clinical learning environment. J Prosthet Dent. 2025;134:1296–303. https://doi.org/10.1016/j.prosdent.2024.03.038.
5. Naseer MA, Saeed S, Afzal A, Ali S, Malik MGR. Navigating the integration of artificial intelligence in the medical education curriculum: a mixed-methods study exploring the perspectives of medical students and faculty in Pakistan. BMC Med Educ. 2025;25:273. https://doi.org/10.1186/s12909-024-06552-2.
6. Künzle P, Paris S. Performance of large language artificial intelligence models on solving restorative dentistry and endodontics student assessments. Clin Oral Investig. 2024;28:575. https://doi.org/10.1007/s00784-024-05968-w.
7. 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. https://doi.org/10.2196/52346.
8. Kandemir M, Sarıbaş EE. Comparative performance of large language models in answering periodontology questions from the Turkish Dental Specialty Examination: a cross-sectional study on accuracy and coverage. BMC Oral Health. 2025;25:1804. https://doi.org/10.1186/s12903-025-07241-2.
9. Law AK, So J, Lui CT, Choi YF, Cheung KH, Kei-Ching Hung K, et al. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. BMC Med Educ. 2025;25:208. https://doi.org/10.1186/s12909-025-06796-6.
10. Jin HK, Lee HE, Kim E. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. BMC Med Educ. 2024;24:1013. https://doi.org/10.1186/s12909-024-05944-8.
11. Çekïç EC, Tavşan O. Evaluating large language models using national endodontic specialty examination questions: are they ready for real-world dentistry? BMC Med Educ. 2025;25:1308. https://doi.org/10.1186/s12909-025-07896-z.
12. Tosun B, Yilmaz ZS. Comparison of artificial intelligence systems in answering prosthodontics questions from the dental specialty exam in Turkey. J Dent Sci. 2025;20:1454–9. https://doi.org/10.1016/j.jds.2025.01.025.
13. Mahmoud R, Shuster A, Kleinman S, Arbel S, Ianculovici C, Peleg O. Evaluating artificial intelligence chatbots in oral and maxillofacial surgery board exams: performance and potential. J Oral Maxillofac Surg. 2025;83:382–9. https://doi.org/10.1016/j.joms.2024.11.007.
14. Puleio F, Lo Giudice G, Bellocchio AM, Boschetti CE, Lo Giudice R. Clinical, research, and educational applications of ChatGPT in dentistry: a narrative review. Appl Sci. 2024;14:10802. https://doi.org/10.3390/app142310802.
15. Jeong H, Han SS, Yu Y, Kim S, Jeon KJ. How well do large language model-based chatbots perform in oral and maxillofacial radiology? Dentomaxillofac Radiol. 2024;53:390–5. https://doi.org/10.1093/dmfr/twae021.
16. Güneş YC, Cesur T, Çamur E, Günbey Karabekmez L. Evaluating text and visual diagnostic capabilities of large language models on questions related to the Breast Imaging Reporting and Data System atlas 5th edition. Diagn Interv Radiol. 2025;31:111–29. https://doi.org/10.4274/dir.2024.242876.
17. El-Hajjami A, Camille S. How good are synthetic requirements? Evaluating LLM-Generated datasets for AI4RE. ArXiv Preprint arXiv. 2025;2506:21138. https://doi.org/10.48550/arXiv.2506.21138.
18. Camlet A, Kusiak A, Ossowska A, Świetlik D. Advances in periodontal diagnostics: application of multimodal language models in visual interpretation of panoramic radiographs. Diagnostics. 2025;15:1851. https://doi.org/10.3390/diagnostics15151851.
19. Sezer B, Okutan AE. Evaluation of ChatGPT-4's performance on pediatric dentistry questions: accuracy and completeness analysis. BMC Oral Health. 2025;25:1427. https://doi.org/10.1186/s12903-025-06791-9.
20. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst. 2023;47:33. https://doi.org/10.1007/s10916-023-01925-4.
21. Sabry Abdel-Messih M, Kamel Boulos MN. ChatGPT in clinical toxicology. JMIR Med Educ. 2023;9:e46876. https://doi.org/10.2196/46876.
22. Mackey BP, Garabet R, Maule L, Tadesse A, Cross J, Weingarten M. Evaluating ChatGPT-4 in medical education: an assessment of subject exam performance reveals limitations in clinical curriculum support for students. Discover Artificial Intelligence. 2024;4:38. https://doi.org/10.1007/s44163-024-00135-2.
23. Uehara O, Morikawa T, Harada F, Sugiyama N, Matsuki Y, Hiraki D, Sakurai H, Kado T, Yoshida K, Murata Y, Matsuoka H, Nagasawa T, Furuichi Y, Abiko Y, Miura H. Performance of ChatGPT-3.5 and ChatGPT-4o in the Japanese National dental examination. J Dent Educ. 2025;89:459–66. https://doi.org/10.1002/jdd.13766.
24. Chau RCW, Thu KM, Yu OY, Hsung RT, Lo ECM, Lam WYH. Performance of generative artificial intelligence in dental licensing examinations. Int Dent J. 2024;74:616–21. https://doi.org/10.1016/j.identj.2023.12.007.
25. Lin CC, Sun J, Chang C, Chang Y, Chang JZ. Performance of artificial intelligence chatbots in national dental licensing examination. J Dent Sci. 2025;20:2307–14. https://doi.org/10.1016/j.jds.2025.05.012.
26. Chen Y, Huang X, Yang F, Lin H, Lin H, Zheng Z, et al. Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: a comparison study. BMC Med Educ. 2024;24:1372. https://doi.org/10.1186/s12909-024-06309-x.
27. Bilgin Avşar D, 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. Turk Klin J Dent Sci. 2024;30:668–73. https://doi.org/10.5336/dentalsci.2024-104610.
28. Danesh A, Pazouki H, Danesh K, Danesh F, Danesh A. The performance of artificial intelligence language models in board-style dental knowledge assessment: a preliminary study on ChatGPT. J Am Dent Assoc. 2023;154:970–4. https://doi.org/10.1016/j.adaj.2023.07.016.
29. Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination: observational study. JMIR Med Educ. 2024;10:e50965. https://doi.org/10.2196/50965.
30. Revilla-León M, Barmak BA, Sailer I, Kois JC, Att W. Performance of an artificial intelligence-based chatbot (ChatGPT) answering the European certification in implant dentistry exam. Int J Prosthodont. 2024;37:221–4. https://doi.org/10.11607/ijp.8852.
31. 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;25:214. https://doi.org/10.1186/s12909-024-06389-9.
32. 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. https://doi.org/10.1111/eje.13073.
33. Gokkurt Yilmaz BN, Ozbey F, Yilmaz BE. Evaluation of the performance of different large language models on head and neck anatomy questions in the dentistry specialization exam in Turkey. Surg Radiol Anat. 2025;47(1):211. https://doi.org/10.1007/s00276-025-03723-8.
34. 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. https://doi.org/10.1186/s12903-025-05926-2.
35. Schwendicke F, Chaurasia A, Wiegand T, Uribe SE, Fontana M, Akota I, et al. Artificial intelligence for oral and dental healthcare: core education curriculum. J Dent. 2023;128:104363. https://doi.org/10.1016/j.jdent.2022.104363.
36. McGrath C, Chau CWR, Molina GF. Monitoring oral health remotely: ethical considerations when using AI among vulnerable populations. Front Oral Health. 2025;6:1587630. https://doi.org/10.3389/froh.2025.1587630.
37. Uribe SE, Maldupa I, Kavadella A, El Tantawi M, Chaurasia A, Fontana M, et al. Artificial intelligence chatbots and large language models in dental education: Worldwide survey of educators. Eur J Dent Educ. 2024;28:865–76. https://doi.org/10.1111/eje.13009.
38. Integrated National Board Dental Examination (INBDE) Technical Report. 2024. https://jcnde.ada.org/-/media/project/ada-organization/ada/jcnde/files/inbde_technical_report.pdf. Accessed 30 Sep 2025.
39. United States Medical Licensing Examination Performance Data. 2024. https://www.usmle.org/performance-data. Accessed 30 Sep 2025.
40. Daungsupawong H, Wiwanitkit V. Generative artificial intelligence in dental licensing examinations: comment. Int Dent J. 2024;74:361. https://doi.org/10.1016/j.identj.2024.01.021.
41. Chau RCW, Thu KM, Yu OY, Lo ECM, Hsung RT, Lam WYH. Response to generative AI in dental licensing examinations: comment. Int Dent J. 2024;74:897–8. https://doi.org/10.1016/j.identj.2024.02.002.

## Publisher's Note

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