RESEARCH

Open Access

# Performance of artificial intelligence on Turkish dental specialization exam: can ChatGPT-4.0 and gemini advanced achieve comparable results to humans?

Soner Sismanoglu[1] and Belen Sirinoglu Capan[2*]

## Abstract

**Background** AI-powered chatbots have spread to various fields including dental education and clinical assistance to treatment planning. The aim of this study is to assess and compare leading AI-powered chatbot performances in dental specialization exam (DUS) administered in Turkey and compare it with the best performer of that year.

**Methods** DUS questions for 2020 and 2021 were directed to ChatGPT-4.0 and Gemini Advanced individually. DUS questions were manually entered into AI-powered chatbot in their original form, in Turkish. The results obtained were compared with each other and the year's best performers. Candidates who score at least 45 points on this centralized exam are deemed to have passed and are eligible to select their preferred department and institution. The data was statistically analyzed using Pearson's chi-squared test ($p < 0.05$).

**Results** ChatGPT-4.0 received 83.3% correct response rate on the 2020 exam, while Gemini Advanced received 65% correct response rate. On the 2021 exam, ChatGPT-4.0 received 80.5% correct response rate, whereas Gemini Advanced received 60.2% correct response rate. ChatGPT-4.0 outperformed Gemini Advanced in both exams ($p < 0.05$). AI-powered chatbots performed worse in overall score (for 2020: ChatGPT-4.0, 65,5 and Gemini Advanced, 50.1; for 2021: ChatGPT-4.0, 65,6 and Gemini Advanced, 48.6) when compared to overall scores of the best performer of that year (68.5 points for year 2020 and 72.3 points for year 2021). This poor performance also includes the basic sciences and clinical sciences sections ($p < 0.001$). Additionally, periodontology was the clinical specialty in which both AI-powered chatbots achieved the best results, the lowest performance was determined in the endodontics and orthodontics.

**Conclusion** AI-powered chatbots, namely ChatGPT-4.0 and Gemini Advanced, passed the DUS by exceeding the threshold score of 45. However, they still lagged behind the top performers of that year, particularly in basic sciences, clinical sciences, and overall score. Additionally, they exhibited lower performance in some clinical specialties such as endodontics and orthodontics.

**Keywords** AI, Artificial Intelligence, ChatGPT, Dentistry, Gemini, Large Language models

*Correspondence:
Belen Sirinoglu Capan
belens90@hotmail.com

[1] Faculty of Dentistry, Department of Restorative Dentistry, Istanbul University-Cerrahpasa, Istanbul, Turkey
[2] Faculty of Dentistry, Department of Pediatric Dentistry, Istanbul University-Cerrahpasa, Istanbul, Turkey

## Introduction

Artificial intelligence (AI) describes technology designed to develop computer systems and algorithms that are capable of performing tasks that normally call for human intelligence [1]. The application of AI, which includes deep learning and neural networks, has significantly changed various industries, including healthcare sector. These AI systems are capable of processing vast amounts of data, analyzing connections between concepts, and producing well-reasoned answers to questions [2]. Large language models (LLMs) are multilayer deep neural networks that are pretrained and use contextual interactions between words and phrases to predict the likelihood of the next word sequence. They can also respond to multilingual questions in English, Turkish, and other languages on the internet. LLMs are particularly popular with end-users in the form of AI-powered chatbots. These capabilites have led to a rapid expansion of LLM applicability across various fields, including dental education and clinical practice [3].

The utilization of these models in dental education and clinical practice has been studied recently [4]. The most commonly used and accessible LLM models are Chat Generative Pretrained Transformer (ChatGPT; OpenAI, San Francisco, CA, USA) and Google Gemini (Google LLC., Mountain View, CA, USA), formerly known as Google Bard. In November 2022, OpenAI, an AI research and deployment company, developed and released ChatGPT, an AI-powered chatbot. Several researchers have assessed its performance for medical and dental education since its release [2, 5–7]. ChatGPT can offer comprehensive answers on a variety of subjects, including both medical and non-medical ones [1, 8]. On March 14, 2023, OpenAI released ChatGPT-4.0, an upgraded LLM, which was trained using a similar methodology to its predecessor. Additionally, ChatGPT-4.0 significantly added multimodal features, like image input [9]. ChatGPT-4.0 exhibited significantly better accuracy on practice step exams and on questions about specialized topics in comparison to ChatGPT-3.5 [10]. Likewise, Google launched their AI-powered chatbot, Google Bard, on March 21, 2023. It simulates human-like conversation with machine learning and natural language processing. Google recently unveiled the updated version of its new AI-based platform, Gemini, on February 8, 2024 [4]. It includes optimized features and improves multimodal analysis. Gemini's capabilities reflect its capacity to analyze large, complex data sets, including charts and images, representing a significant advancement on its predecessor Bard. Applications in the medical and dental fields, where data is frequently through visual representations such as medical photographs or scans, are especially pertinent to this capacity. Gemini has the potential to be a useful diagnostic and educational tool for medical practitioners by analyzing these images [11]. Gemini Advanced is available as the latest version of Google's Gemini AI-powered chatbot on the date of the study (March 2024). Google claims that Gemini Advanced represents a significant improvement compared to previous versions.

AI-powered chatbots are frequently used in areas such as academic writing, education, queries from medical and dental students, diagnostic support for doctors and information about patients' diseases [12]. Comparing the performance of AI-powered chatbots in dental education and practice to that of human professionals is a significant indicator for verifying the possible usage of AI in dentistry [3, 13]. So far, a limited number of studies evaluated the performance of AI-powered chatbots in answering inquiries about dentistry. Their application in general dental board exams, which includes questions from all specialties, is underexplored. Recently, specialization in dentistry, as in medicine, has become more and more in demand; thus, dentists in Turkey who have completed a five-year undergraduate dental program are beginning to see specialization as a career goal. The only way to receive specialization training in Turkey is to succeed in the dental specialization exam (DUS), which is a central exam based on competence and competition principles and regulated by Turkish state authorities. It questions the professional knowledge of dental graduates in clinical dentistry and basic sciences and consists of multiple-choice questions (MCQs) [14]. MCQs provide educators with achievement data, evaluate a broad range of topics, grade and provide feedback effectively, and cover a huge number of domains; thus, still the most commonly utilized objective method to measure student learning. Furthermore, answering MCQs helps students identify their knowledge gaps quickly, which is helpful for planning future learning [5, 15]. Studies have indicated that AI-powered chatbots possess a high capability for answering MCQs [16].

ChatGPT and Gemini have been used in medical education, and specialties such as urology and ophthalmology. Nevertheless, research on AI-powered chatbots' knowledge of dental field is very limited [1, 2, 6]. Although there are few studies in the literature that investigate the performance of AI-powered chatbots on specific dental specialties, there is no study that examined their performance in general dental board exams, that contain questions from all specialties together.

The utilization of AI-powered chatbots can be included in dental education and clinical practice, but only if they provide adequate answers on all specialties. Therefore, the aim of this study is to assess ChatGPT-4.0 and Gemini Advanced AI-powered chatbot performance in the DUS tests administered in 2020 and 2021 and compare it with the performance of the dentists participating in these exams. This study aims to measure the performance of commonly used AI-powered chatbots across all dental specialties, allowing for a comparison of the chatbots' knowledge and information processing capabilities with human performance. Furthermore, this is the first study to assess Gemini Advanced's -the most recent version of Gemini- performance in the field of dentistry. The first null hypothesis was that both AI-powered chatbots are capable of passing the DUS, the second null hypothesis was that there is no significant difference between the ChatGPT-4.0 and Gemini Advanced performances.

## Materials and methods

In this study, the most recent versions of the two most popular AI-powered chatbots were used. Exam questions were directed to ChatGPT-4.0 (OpenAI, San Francisco, CA, USA) and Gemini Advanced (Google LLC., Mountain View, CA, USA).

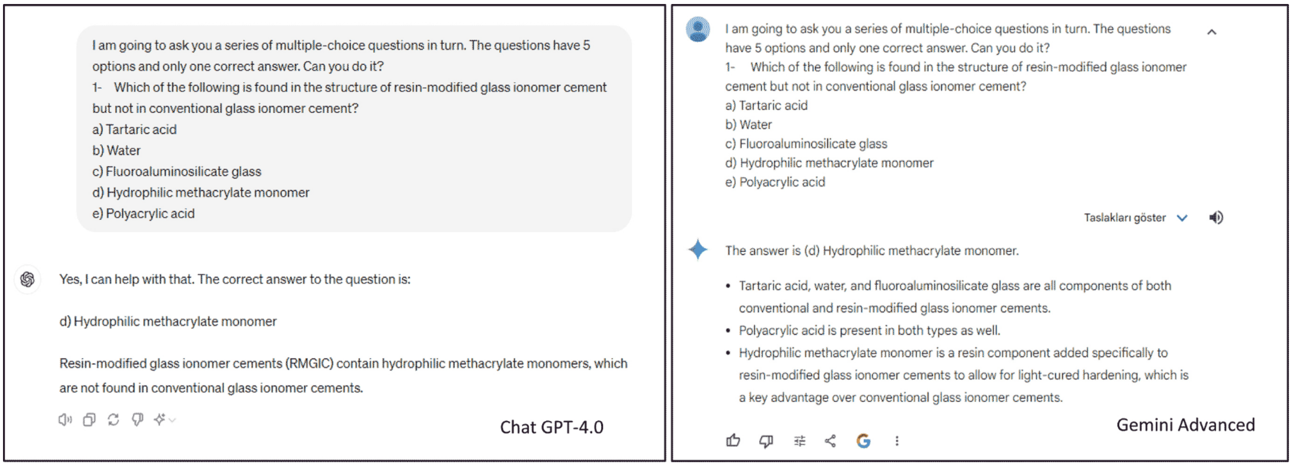
### Study design and questions

All DUS questions for 2020 and 2021 were collected from the database of the Student Selection and Placement Centre (ÖSYM) which is an institution established by the Republic of Türkiye to assess and place proficient applicants who seek admission to higher education programs by means of centralized examinations. Questions after 2021 are excluded as ChatGPT covers data up to January 2022 at the time of this study was conducted. DUS is an exam consisting of a total of 120 questions in which professional knowledge in basic medicine and clinical dentistry is measured. While the basic medicine section consists of 40 questions, the clinical dentistry section includes a total of 80 questions, 10 each from specialties such as endodontics, oral and maxillofacial radiology, oral and maxillofacial surgery, orthodontics, pedodontics, periodontology, prosthodontics, restorative dentistry. In the 2021 exam, one of the questions asked in the field of orthodontics and pedodontics was canceled after the exam, so the evaluation was made over 118 questions in 2021. Since both ChatGPT-4.0 and Gemini Advanced are capable of answering image-involving queries, one question (for endodontics) from 2020 and six questions (one for basic sciences, three on endodontics, one for oral and maxillofacial radiology, and one for restorative dentistry) from 2021 were not excluded and directed to these AI-powered chatbots.

Bloom's taxonomy is a hierarchical system used in education to categorize learning objectives and consists of six categories, namely remembering, understanding, applying, analyzing, evaluating, and creating. Both investigators (SS and BSC) independently categorized the questions according to Bloom's taxonomy. In case of a contradiction between the two authors, the contradictory questions were re-evaluated and consensus was reached. The first two, namely remembering and understanding, were categorized as lower-order and the other four were categorized as high-order [17].

### Administration of questions

The DUS questions, originally written in Turkish, were entered individually into ChatGPT and Gemini Advanced as prompts, and both were asked to respond from February 14 and 15, 2024, respectively. Data collection was conducted within 24 h for each separate AI-powered chatbot, to minimize heterogeneity among provided answers. The prompt "I am going to ask you a series of multiple-choice questions in turn. The questions have 5 options and only one correct answer. Can you do it?" and then the DUS questions were asked in order (Fig. 1). Since both ChatGPT-4.0 and Gemini Advanced have multi-language features, all communication with the chatbots was carried out in Turkish and questions were asked in Turkish. The first answer received was recorded



I am going to ask you a series of multiple-choice questions in turn. The questions have 5 options and only one correct answer. Can you do it?
1- Which of the following is found in the structure of resin-modified glass ionomer cement but not in conventional glass ionomer cement?
a) Tartaric acid
b) Water
c) Fluoroaluminosilicate glass
d) Hydrophilic methacrylate monomer
e) Polyacrylic acid

Yes, I can help with that. The correct answer to the question is:

d) Hydrophilic methacrylate monomer

Resin-modified glass ionomer cements (RMGIC) contain hydrophilic methacrylate monomers, which are not found in conventional glass ionomer cements.

I am going to ask you a series of multiple-choice questions in turn. The questions have 5 options and only one correct answer. Can you do it?
1- Which of the following is found in the structure of resin-modified glass ionomer cement but not in conventional glass ionomer cement?
a) Tartaric acid
b) Water
c) Fluoroaluminosilicate glass
d) Hydrophilic methacrylate monomer
e) Polyacrylic acid

The answer is (d) Hydrophilic methacrylate monomer.

- Tartaric acid, water, and fluoroaluminosilicate glass are all components of both conventional and resin-modified glass ionomer cements.
- Polyacrylic acid is present in both types as well.
- Hydrophilic methacrylate monomer is a resin component added specifically to resin-modified glass ionomer cements to allow for light-cured hardening, which is a key advantage over conventional glass ionomer cements.

**Fig. 1** Example prompt input for AI-powered chatbots: During the study, the questions were asked to the AI-powered chatbots in Turkish. For better understanding by the readers, the example shown is a translated version for the English audience

without using the "Regenerate" feature. The recorded answers were compared with the provided answer key and recorded as true or false (Fig. 2). Questions that AI-powered chatbots preferred not to answer were recorded as empty. The questions that Gemini Advanced left empty in all exams are questions containing an image of a human.

### Scoring of responses

In order to give a general idea and to better understand the performance in a real-life scenario, the exam scores of the AI-powered chatbots were calculated according to that year using an online scoring calculation tool. Thus, chatbots can be compared both against each other and against the best performing participant of that year. In the score calculation, the procedure of 4 false answers canceling one true answer, which is called net score, was taken into account in accordance with the DUS rules. Candidates who score at least 45 points on this centralized exam are deemed to have passed and are eligible to select their preferred department and institution.

### Statistical analysis

Statistical analyses were performed using statistical software (SPSS v22; IBM, Armonk, NY, USA). In the statistical analysis of the study data, in addition to descriptive statistical methods, Pearson Chi-Square test was used for comparison of qualitative variables. Significance was assessed at $p < 0.05$.

## Results

The findings of the present study revealed that ChatGPT-4.0 outperformed Gemini Advanced in both the basic sciences and clinical sciences sections ($p < 0.001$), achieving better overall scores. Although there were statistically significant differences in the clinical sciences section as a whole, there were no statistically significant differences in the subgroup analysis for each individual specialty within the clinical sciences (Fig. 3). While periodontology was the clinical specialty in which both AI-powered chatbots performed best, the lowest performances were observed in endodontics and orthodontics (Table 1; Fig. 4).

Higher-order questions comprised 96 (40.3%) of the 238 questions asked in the exams. ChatGPT-4.0 achieved statistically higher performance in both categories (higher-order and lower-order) when the performance of both AI-powered chatbots was evaluated based on the taxonomy levels ($p < 0.05$). For both AI-powered chatbots, the correct response rate decreased as the Bloom's taxonomy level of the questions increased (Table 2).

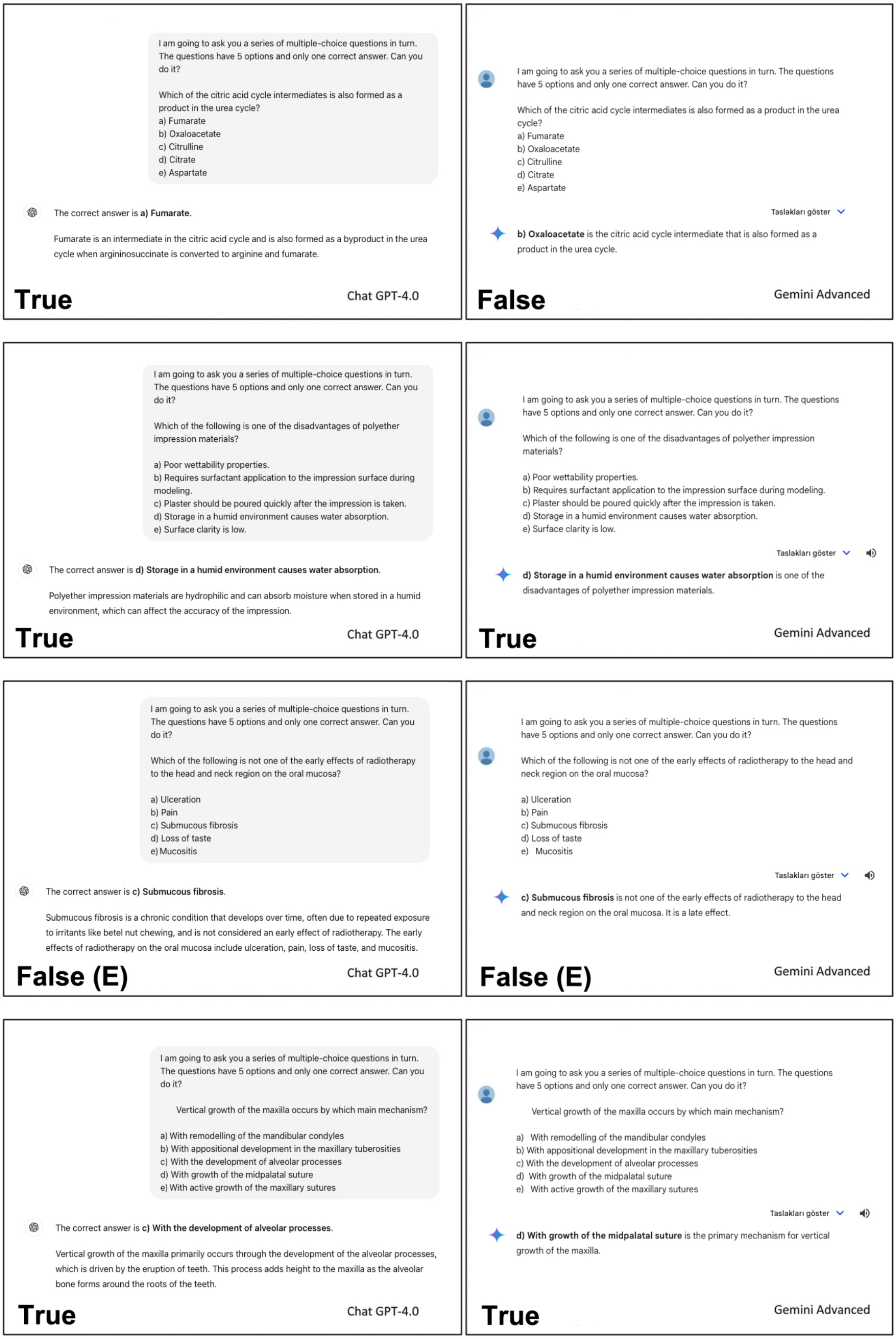
ChatGPT-4.0 scored 65.5 points with 83.3% correct response rate on the 2020 exam. In comparison, Gemini Advanced received 50.1 points with a 65% correct response rate. On the 2021 exam, ChatGPT-4.0 received 65.6 points with 80.5% correct response rate, whereas Gemini Advanced received 48.6 points with 60.2% correct response rate. The findings of this study revealed a statistically significant difference in performance between ChatGPT-4.0 and Gemini Advanced on both the 2020 and 2021 exams. ChatGPT-4.0 outperformed Gemini Advanced in both exams, with statistical significance ($p < 0.05$). When compared to that year's best performer, AI-powered chatbots performed worse in clinical sciences and overall score ($p < 0.001$), although ChatGPT performed the best in basic sciences for the 2020 exam, as shown in Table 3. ChatGPT-4.0 achieved perfect score in the 2020 exam's basic sciences sect. (40 out of 40), whereas Gemini Advanced had the lowest performance across all sections (Table 3).

Notably, a total of 7 images were asked in the exams administered in this study. Four of these seven questions were from endodontics. When the performance of AI-powered chatbots in these questions was examined, ChatGPT-4.0 had 4 incorrect, 3 correct replies, whereas Gemini Advanced had 3 empty, 4 incorrect answers. The questions that Gemini Advanced left unanswered in all exams contained images with human visuals.

## Discussion

The increasing popularity of language-based AI models, such as ChatGPT and Gemini, can be attributed to their ability to generate coherent conversations and maintain contextual continuity. They are trained on vast amounts of text data from various sources, including books, articles, webpages, and more, utilizing deep learning techniques such as neural networks. This study demonstrated the applicability of two popular AI-powered chatbots in the field of dentistry by evaluating their performance in the DUS and comparing it with the top-performing participants in this exam. ChatGPT-4.0 demonstrated significantly better performance across all sections and specialties compared to Gemini Advanced. The findings of the present study highlight the potential effectiveness and limitations of AI-powered chatbots in dentistry. One of the main differences between Gemini and ChatGPT is that Gemini can access and use online data in real time when creating answers. In comparison, ChatGPT does not currently have web crawling capabilities and instead depends on previous training data up to January 2022 [1, 18]. Therefore, this study aimed to assess their performance on the 2020 and 2021 exams.

LLMs, such as ChatGPT and Gemini were not primarily designed for answering inquiries about medicine and dentistry. As a result, it lacks the context and dental expertise necessary to completely comprehend the complex connections between various problems and treatments [19]. However, previous studies have reported



I am going to ask you a series of multiple-choice questions in turn. The questions have 5 options and only one correct answer. Can you do it?

Which of the citric acid cycle intermediates is also formed as a product in the urea cycle?
a) Fumarate
b) Oxaloacetate
c) Citrulline
d) Citrate
e) Aspartate

The correct answer is **a) Fumarate**.

Fumarate is an intermediate in the citric acid cycle and is also formed as a byproduct in the urea cycle when argininosuccinate is converted to arginine and fumarate.

**True** Chat GPT-4.0

I am going to ask you a series of multiple-choice questions in turn. The questions have 5 options and only one correct answer. Can you do it?

Which of the citric acid cycle intermediates is also formed as a product in the urea cycle?
a) Fumarate
b) Oxaloacetate
c) Citrulline
d) Citrate
e) Aspartate

**b) Oxaloacetate** is the citric acid cycle intermediate that is also formed as a product in the urea cycle.

**False** Gemini Advanced

I am going to ask you a series of multiple-choice questions in turn. The questions have 5 options and only one correct answer. Can you do it?

Which of the following is one of the disadvantages of polyether impression materials?

a) Poor wettability properties.
b) Requires surfactant application to the impression surface during modeling.
c) Plaster should be poured quickly after the impression is taken.
d) Storage in a humid environment causes water absorption.
e) Surface clarity is low.

The correct answer is **d) Storage in a humid environment causes water absorption.**

Polyether impression materials are hydrophilic and can absorb moisture when stored in a humid environment, which can affect the accuracy of the impression.

**True** Chat GPT-4.0

I am going to ask you a series of multiple-choice questions in turn. The questions have 5 options and only one correct answer. Can you do it?

Which of the following is one of the disadvantages of polyether impression materials?

a) Poor wettability properties.
b) Requires surfactant application to the impression surface during modeling.
c) Plaster should be poured quickly after the impression is taken.
d) Storage in a humid environment causes water absorption.
e) Surface clarity is low.

**d) Storage in a humid environment causes water absorption** is one of the disadvantages of polyether impression materials.

**True** Gemini Advanced

I am going to ask you a series of multiple-choice questions in turn. The questions have 5 options and only one correct answer. Can you do it?

Which of the following is not one of the early effects of radiotherapy to the head and neck region on the oral mucosa?

a) Ulceration
b) Pain
c) Submucous fibrosis
d) Loss of taste
e) Mucositis

The correct answer is **c) Submucous fibrosis**.

Submucous fibrosis is a chronic condition that develops over time, often due to repeated exposure to irritants like betel nut chewing, and is not considered an early effect of radiotherapy. The early effects of radiotherapy on the oral mucosa include ulceration, pain, loss of taste, and mucositis.

**False (E)** Chat GPT-4.0

I am going to ask you a series of multiple-choice questions in turn. The questions have 5 options and only one correct answer. Can you do it?

Which of the following is not one of the early effects of radiotherapy to the head and neck region on the oral mucosa?

a) Ulceration
b) Pain
c) Submucous fibrosis
d) Loss of taste
e) Mucositis

**c) Submucous fibrosis** is not one of the early effects of radiotherapy to the head and neck region on the oral mucosa. It is a late effect.

**False (E)** Gemini Advanced

I am going to ask you a series of multiple-choice questions in turn. The questions have 5 options and only one correct answer. Can you do it?

Vertical growth of the maxilla occurs by which main mechanism?

a) With remodelling of the mandibular condyles
b) With appositional development in the maxillary tuberosities
c) With the development of alveolar processes
d) With growth of the midpalatal suture
e) With active growth of the maxillary sutures

The correct answer is **c) With the development of alveolar processes**.

Vertical growth of the maxilla primarily occurs through the development of the alveolar processes, which is driven by the eruption of teeth. This process adds height to the maxilla as the alveolar bone forms around the roots of the teeth.

**True** Chat GPT-4.0

I am going to ask you a series of multiple-choice questions in turn. The questions have 5 options and only one correct answer. Can you do it?

Vertical growth of the maxilla occurs by which main mechanism?

a) With remodelling of the mandibular condyles
b) With appositional development in the maxillary tuberosities
c) With the development of alveolar processes
d) With growth of the midpalatal suture
e) With active growth of the maxillary sutures

**d) With growth of the midpalatal suture** is the primary mechanism for vertical growth of the maxilla.

**True** Gemini Advanced

**Fig. 2** Screenshots of sample questions: The screenshots show sample questions answered by ChatGPT-4.0 and Gemini Advanced, covering topics from basic sciences and various clinical specialties. The answers on the left are from ChatGPT-4.0, while those on the right are from Gemini Advanced. The correctness or incorrectness of the responses given by the two AI-powered chatbots to the same questions is indicated at the bottom left of each question box

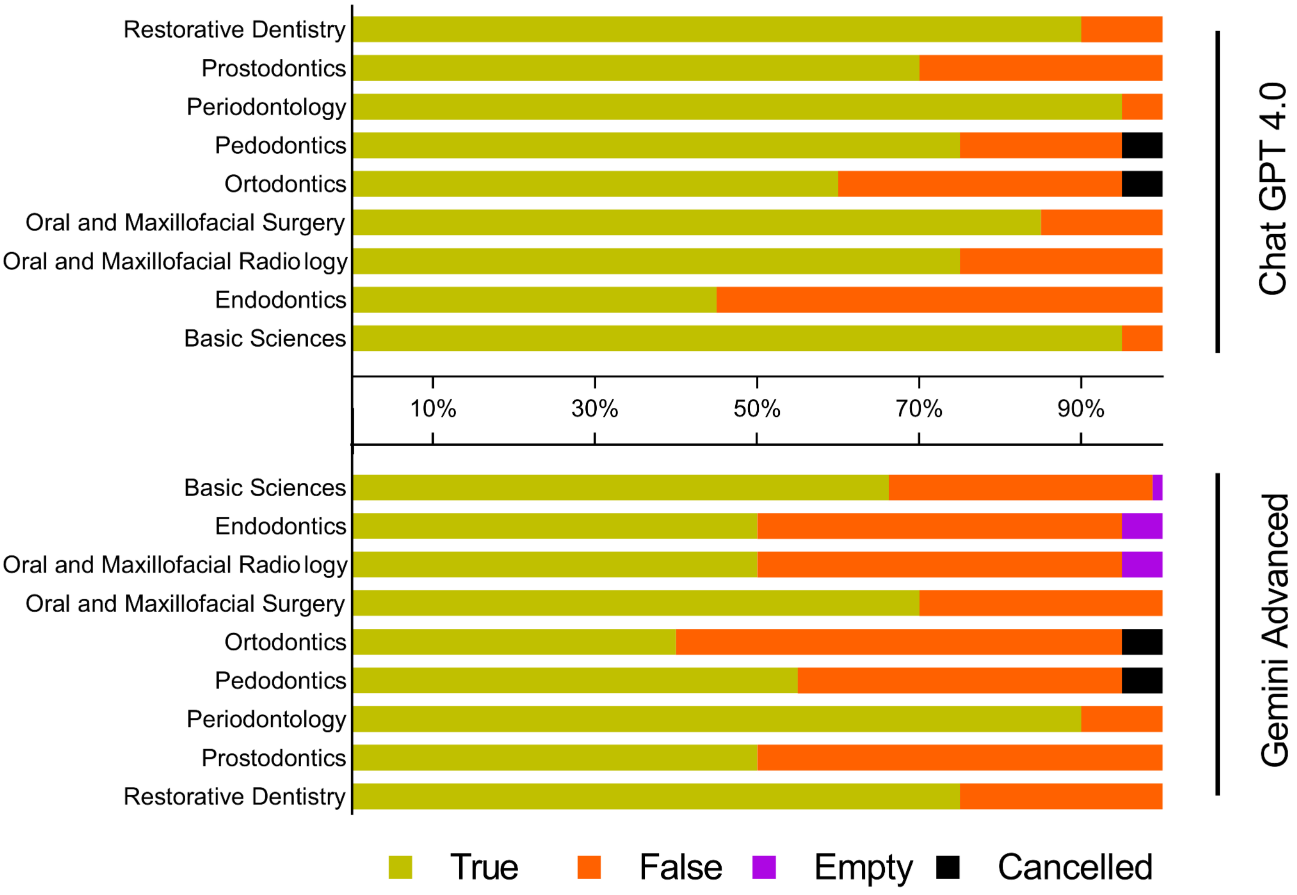


**Fig. 3** Stacked bar graphs displaying the correct and false response rates for each subspecialty. It shows combined 2020/2021 exam performance of chatbots. The correct response rates are represented by green bars, while the false response rates are represented by red bars. The subspecialties are ordered in alfa-numeric

**Table 1** Correct answer rates of ChatGPT-4.0 and Gemini Advanced by exam specialties

| | Corect answer rates | | |
|---|---|---|---|
| **Category** | **ChatGPT-4.0** | **Gemini Adv.** | ***p*-Value** |
| Basic medicine | 76/80 (95%) | 53/80 (66.3%) | < 0.001* |
| Endodontics | 9/20 (45%) | 10/20 (50%) | 0.752 |
| O. Maxillofacial Radio. | 15/20 (75%) | 10/20 (50%) | 0.102 |
| O. Maxillofacial Surg. | 17/20 (85%) | 14/20 (70%) | 0.256 |
| Orthodontics | 12/19 (63.2%) | 8/19 (42.1%) | 0.194 |
| Pedodontics | 15/19 (79%) | 11/19 (57.9%) | 0.163 |
| Periodontology | 19/20 (95%) | 18/20 (90%) | 0.548 |
| Prosthodontics | 14/20 (70%) | 10/20 (50%) | 0.197 |
| Restorative Dentistry | 18/20 (90%) | 15/20 (75%) | 0.212 |

The data shows the combined performance for 2020 and 2021exams. Statistically significant differences between AI models are indicated by asterisks (*) and calculated using Pearson's chi-square test

that these AI-powered chatbots perform successfully in both medical licensing exams and even questions that require expertise [7, 20]. Kung et al. [6] demonstrated that ChatGPT was able to pass all three United States Medical Licensing Examination (USMLE) exams, without the need for any specific training or user support. Furthermore, it was shown that ChatGPT and Bard, performed successfully in a variety of medical specialties, including otolaryngology, toxicology, cardiology, urology, and parasitology [1, 7, 21, 22]. According to a recent study assessing ChatGPT's performance on board-style questions meant to help dental students prepare for the Integrated National Board Dental Examination (INBDE), ChatGPT-3.5 did not perform well enough on the board-style knowledge assessment. On the other hand, ChatGPT-4.0 showed a competent ability to produce accurate dental content [23]. Ali et al., in their evaluation of ChatGPT's responses to dental queries, found that 90% of the AI system's responses to multiple-choice questions were accurate [9]. Other studies that examined the ChatGPT-4.0's performance in answering periodontology, endodontics, oral radiology, and oral implantology-related questions revealed a correct response rate of 73.6%, 57.3%, 80.7% and 84% respectively [24–27]. In accordance with previous studies, both the ChatGPT-4.0 and Gemini Advanced programs in this study passed the DUS with 80% (corresponds to approximately 65 points) and 60% (corresponds to approximately 50 points) correct

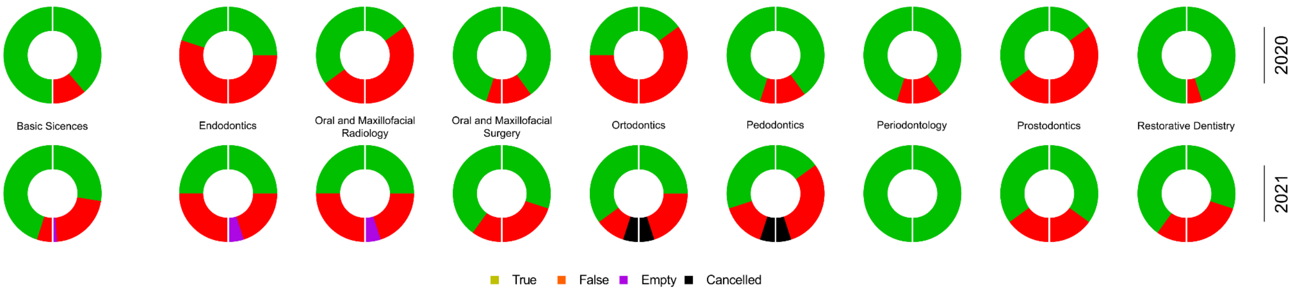


**Fig. 4** Donut charts illustrating the correct versus false rates for ChatGPT-4.0 and Gemini Advanced, stratified by subspecialties. The correct rates are represented by the green sections of the charts, while the false rates are represented by the red sections. The size of each donut chart is proportional to the total number of questions in each subspecialty. ChatGPT-4.0 is represented on left hand side and Gemini Advanced is represented on right hand side of the donut charts

**Table 2** Correct answer rates of ChatGPT-4.0 and Gemini Advanced according to Bloom's taxonomy order

| | Corect answer rates | | |
|---|---|---|---|
| **Taxonomy** | **ChatGPT-4.0** | **Gemini Adv.** | ***p*-Value** |
| High | 74/96 (77.1%) | 54/96 (56.3%) | 0.002* |
| Low | 121/142 (85.2%) | 98/142 (69%) | 0.001* |

The data shows the combined performance for 2020 and 2021exams. Statistically significant differences between AI models are indicated by asterisks (*) and calculated using Pearson's chi-square test

response rates, respectively. The first null hypothesis that both AI-powered chatbots are capable of passing the DUS was accepted.

The DUS, the exam utilized in this study, comprises of MCQs. Generally, MCQs should be designed to assess participants' critical thinking abilities. Using the cognitive domains of Bloom's taxonomy is one way to make sure that MCQs assess higher-order thinking as opposed to an examinee's capacity to only recall basic information. Blooms' taxonomy consists of 6 cognitive domains, including knowledge, complication, application, analysis, synthesis, and evaluation. Lower-order questions are those that focus on the first two cognitive domains, and higher-order questions are those that address the remaining four [15, 28]. The present research revealed that while the correct response rate for both AI-powered chatbots decreased as the questions' taxonomy level increased; and that ChatGPT-4.0 outperformed Gemini Advanced for higher-order questions. This finding is consistent with similar studies in the literature that examines the performance of AI-powered chatbots in the fields of neurology and orthopedics [17, 29]. Thus, it could be inferred that ChatGPT-4.0 might serve as a more reliable supplementary teaching tool in dentistry compared to Gemini Advanced.

Studies in the literature showed that there are variances in performance-based comparisons between AI-powered chatbots. In a study comparing the Nephrology Board test performance of ChatGPT-3.5, ChatGPT-4.0, and Bard programs, ChatGPT-4.0 outperforms the other two systems in the subcategories of question type, taxonomy level, and image [3]. Contradictory outcomes between AI-powered chatbots have been recorded in the field of dentistry. In a study where the performance only on prosthodontics and oral and maxillofacial radiology questions was assessed, ChatGPT-3.5 and Bard performed 52.8% in oral and maxillofacial radiology questions and 30–40% correctly in prosthodontics questions. There was no significant difference between the two AI-powered chatbots, while their dentistry-related success was regarded as poor [5]. Similarly, Azadi et al. [30] reported that different AI powered chatbots performed 26–38% on oral maxillofacial surgery questions. Among these chatbots, ChatGPT-4.0 showed the highest performance. In another study, that compared the performance of various AI-powered chatbots in queries regarding pediatric dentistry, ChatGPT-4.0 was shown to be the most successful

**Table 3** Correct answer rates of chat modal compared to best performers by exam year

| | Corect answer rates | | | | |
|---|---|---|---|---|---|
| **Exam Year** | **Category** | **ChatGPT-4.0** | **Gemini Adv.** | **Best Performer** | ***p*-Value** |
| 2020 | Basic medicine | 40/40 (100%) | 31/40 (77.5%) | 34/40 (85%) | 0.008* |
| | Clinical Dentistry | 60/80 (75%) | 47/80 (58.75%) | 76/80 (95%) | < 0.001* |
| | Total | 100/120 (83.3%) | 78/120 (65%) | 110/120 (91.7%) | < 0.001* |
| 2021 | Basic Medicine | 36/40 (90%) | 22/40 (55%) | 38/40 (95%) | < 0.001* |
| | Clinical Dentistry | 59/78 (75.6%) | 49/78 (62.8%) | 73/78 (93.6%) | < 0.001* |
| | Total | 95/118 (80.5%) | 71/118 (60.2%) | 111/118 (94.1%) | < 0.001* |

Statistically significant differences between AI models are indicated by asterisks (*) and calculated using Pearson's chi-square test. Best performer is the first placed attendant of that year exam

AI-powered chatbot. Google Bard provided responses to over ten queries in the same survey with the phrase "I'm a text-based AI and can't assist with…" and therefore, Bard was not included for additional assessment [31]. In this study, as in previous studies, there was a significant difference in their performances, despite the fact that both AI-powered chatbots passed the DUS. In terms of overall score for each of the two years, ChatGPT-4.0 outperformed Gemini Advanced. Therefore, the second hypothesis that there is no significant difference between the ChatGPT-4.0 and Gemini Advanced performances was rejected. To the best of the authors' knowledge there is no study examining Gemini Advanced performance in the literature. Most of the studies were carried out with the oldest version, Bard. This is the first study in the field of dentistry to use Gemini Advanced.

Previous studies in the field of dentistry revealed that AI-powered chatbots exhibit varying levels of performance across several clinical specialties. The results of these studies indicate that ChatGPT-4.0 answered 73.6% of periodontology questions correctly, whereas in endodontic questions had a success rate of 57.3% [24, 25]. In other studies, ChatGPT-3.5 and Bard demonstrated an overall success rate of 30–40% in the prosthodontics, 52.8% success rate in the oral and maxillofacial radiology and 57.5% success rate in dental traumatology questions [5, 32]. Performance variances were seen across clinical specialties in the current study. Compared to previous studies, correct response rates were higher in oral and maxillofacial radiology (75%), prosthodontics (70%), and periodontology (95%), but were lower in endodontics (45%). Periodontology was shown to have the best performance for both AI-powered chatbots, whereas the endodontics and orthodontics had the lowest performance. The difficulty of orthodontic-related questions and the technical aspect of the ideas demanding accurate assessment may be the cause of the poor performance. Furthermore, compared to other dental specialties, there's a possibility that the dataset used to train ChatGPT contains a less amount of information relating to orthodontics. The fact that the majority of image-containing questions in both years belonged to the endodontics contributes to the poor performance in this specialty. Gemini Advanced is the latest version of the AI-powered chatbot of Google LLC.

Previous versions of AI-powered chatbots did not support questions containing images. Nevertheless, it has been noted that both ChatGPT-4.0 and Gemini Advanced support queries with images after the latest upgraded versions. Therefore, questions containing images were included in this study. A total of seven inquiries containing images were posed to the AI-powered chatbots in the present study. ChatGPT-4.0 correctly responded to three of them, whereas Gemini Advanced failed to respond any of them. In previous research, where ChatGPT-3.5, ChatGPT-4.0, and Bard are compared, it was found that ChatGPT-4.0 also performed well in questions containing images on neurosurgery [18]. Based on these findings, although it can be argued that AI-powered chatbots have made progress in queries involving images, it is still a question mark whether they meet the necessary requirements. This study is also significant, since it is the first to assess the performance of these AI-powered chatbots on questions involving images in the field of dentistry. In order to accurately assess the claims of developers, future studies are needed to measure the performances of AI-powered chatbots with questions involving images.

It is important to compare the performance of AI-powered chatbots in medical disciplines, as well as with the participants taking tests. Huh et al. reported that ChatGPT performed worse than students in parasitology questions [7]. In another research, pediatric dentistry questions were posed of general dentists, dental students, and publicly accessible chatbots. The AI-powered chatbots performed worse than the general dentists and students [30]. Similar to the previous studies, the performance of both AI-powered chatbots in the basic sciences, clinical sciences, and overall score were lower than that of the year's best performer. These findings indicate that although AI-powered chatbots are increasingly used in dentistry and other fields, human intelligence remains superior. However, it should be kept in mind that AI-powered technologies are constantly developing, with new versions being released.

AI-powered chatbots have recently been employed in medicine and dentistry, particularly for academic writing and responding to student inquiries. However, there are still a limited number of studies on the accuracy and reliability of these AI-powered chatbots in the field of medicine and dentistry. Both successful and unsuccessful performances of AI-powered chatbots have been documented in medical studies [2, 33]. In this study, which tests AI-powered chatbots' performance in the field of dentistry, both AI-powered chatbots have achieved promising results, yet there are clinical specialties with poor performance. The responses provided by the latest versions used in this study were generally accurate and meaningful, indicating their potential for integration into dental education and clinical practice. Since they grew up in an era when technology and the internet were readily available, students in the new generation are used to having quick and simple access to information. Therefore they can quickly adapt to this training method [34]. Albeit, AI systems may still produce incorrect responses due to the "hallucination phenomenon" effect [2]. As a consequence of this, AI systems can only be utilized as an auxiliary material for education and clinical practice, rather than as a primary method. The main limitation of this study is that the AI-powered chatbots are constantly updated, therefore, the versions used in the present study may not be the latest version used at the time of publication. It should be kept in mind that these systems are not intended for dental and medical use, and that different results can be obtained with the development of AI-powered chatbots. Therefore, additional research on the application of AI systems in dentistry is essential. Besides this, present study is limited by the fact that it only includes exam questions from two specific years (2020 and 2021), which may restrict the generalizability of the results. Additionally, the limited number of image containing questions and the exclusive focus on MCQs may not fully reflect the AI's performance across different exam formats or in response to a broader set of visual data.

## Conclusion

AI-powered chatbots, namely ChatGPT-4.0 and Gemini Advanced, passed the DUS by exceeding the threshold score of 45. However, despite this achievement, they still lagged behind the best performers of that year, particularly in basic sciences, clinical sciences, and overall score. Additionally, they unsuccessfully performed in some clinical specialties such as endodontics and orthodontics. Academicians and students in the dental field should be aware of the advancement of AI-powered chatbots and consider its potential adaptation in dental education and clinical practice.

### Abbreviations

| | |
|---|---|
| AI | Artificial Intelligence |
| LLM | Large Language Model |
| DUS | Dental Specialization Exam |
| MCQ | Multiple-Choice Question |
| USMLE | United States Medical Licensing Examination |
| INBDE | Integrated National Board Dental Examination |

### Acknowledgements

Not applicable.

### Author contributions

Conception- SS, BSC; Design- SS; Data collection- SS, BSC; Analysis/Interpretation- SS; Literature Research- BSC; Writing Manuscript- BSC; Critical review- SS.

### Funding

No funding was provided for this study.

### Data availability

The data used and analyses in this study are available upon reasonable request to the corresponding author.

## Declarations

### Ethical approval

This was not a study of human subjects, but an analysis of the results of an educational examination routinely conducted for dental specialization. Therefore, this study was exempted from ethical approval due to its nature and the use of publicly accessible data.

### Informed consent

This study does not require informed consent as it does not involve human participation.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing interests.

Received: 4 July 2024 / Accepted: 21 November 2024
Published online: 10 February 2025

### References

1. Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K, Cotofana S, Alfertshofer M. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol. 2023;280:4271–8. https://doi.org/10.1007/s00405-023-08051-4.
2. Huynh LM, Bonebrake BT, Schultis K, Quach A, Deibert CM. New Artificial Intelligence ChatGPT performs poorly on the 2022 self-assessment study program for Urology. Urol Pract 10:409–15. https://doi.org/10.1097/UPJ.0000000000000406
3. Noda R, Izaki Y, Kitano F, Komatsu J, Ichikawa D, Shibagaki Y. (2024) Performance of ChatGPT and Bard in self-assessment questions for nephrology board renewal. Clin Exp Nephrol. 2023;28:465–9. https://doi.org/10.1007/s10157-023-02451-w
4. Carlà MM, Gambini G, Baldascino A, Boselli F, Giannuzzi F, Margollicci F, Rizzo S. Large language models as assistance for glaucoma surgical cases: a ChatGPT vs. Google Gemini comparison. Graefes Arch Clin Exp Ophthalmol. 2024;1–15. https://doi.org/10.1007/s00417-024-06470-5.
5. Tunç-Oguzman R, Yurdabakan ZZ. Performance of chat generative pretrained transformer and bard on the questions asked in the dental specialty entrance examination in Turkey regarding bloom's revised taxonomy. Curr Res Dent Sci. 2024;34:25–34. https://doi.org/10.5152/CRDS.2024.23261.
6. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198. https://doi.org/10.1371/journal.pdig.0000198.
7. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination? A descriptive study. J Educ Eval Health Prof. 2023;20:1. https://doi.org/10.3352/jeehp.2023.20.1.
8. Mihalache A, Popovic MM, Muni RH. Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. JAMA Ophthalmol. 2023;141:589–97. https://doi.org/10.1001/jamaophthalmol.2023.1144.
9. Ali K, Barhom N, Tamimi F, Duggal M. ChatGPT-A double-edged sword for healthcare education? Implications for assessments of dental students. Eur J Dent Educ. 2024;28:206–11. https://doi.org/10.1111/eje.12937.
10. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: an assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. Med Teach. 2024;46:366–72. https://doi.org/10.1080/0142159X.2023.2249588.
11. Masalkhi M, Ong J, Waisberg E, Lee AG. Google DeepMind's gemini AI versus ChatGPT: a comparative analysis in ophthalmology. Eye (Lond). 2024. https://doi.org/10.1038/s41433-024-02958-w.
12. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent. 2023;35:1098–102. https://doi.org/10.1111/jerd.13046.
13. Mohammad-Rahimi H, Motamedian SR, Pirayesh Z, Haiat A, Zahedrozegar S, Mahmoudinia E, Rohban MH, Krois J, Lee JH, Schwendicke F. Deep learning in periodontology and oral implantology: a scoping review. J Periodontal Res. 2022;57:942–51. https://doi.org/10.1111/jre.13037.
14. Culhaoglu AK, Kilicarslan MA, Deniz KZ. Evaluation of dental specialty entrance examination on approach and preferences for different levels of dental education and post graduation. J Dent Fac Atatürk Uni. 2021;31:420–6. https://doi.org/10.17567/ataunidfd.911839.
15. Zaidi NLB, Grob KL, Monrad SM, Kurtz JB, Tai A, Ahmed AZ, Gruppen LD, Santen SA. Pushing critical thinking skills with multiple-choice questions:

does Bloom's taxonomy work? Acad Med. 2018;93:856–9. https://doi.org/10.1097/ACM.0000000000002087.

16. Gonsalves C. On ChatGPT: what promise remains for multiple choice assessment? J Learn Develop High Educ. 2023;27. https://doi.org/10.47408/jldhe.vi27.1009.
17. Ali R, Tang OY, Connolly ID, Zadnik Sullivan PL, Shin JH, Fridley JS, Asaad WF, Cielo D, Oyelese AA, Doberstein CE, Gokaslan ZL, Telfeian AE. Performance of ChatGPT and GPT-4 on Neurosurgery Written Board examinations. Neurosurgery. 2023;93:1353–65. https://doi.org/10.1227/neu.0000000000002632.
18. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, Cielo D, Oyelese AA, Doberstein CE, Telfeian AE, Gokaslan ZL, Asaad WF. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards Preparation Question Bank. Neurosurgery. 2023;93:1090–8. https://doi.org/10.1227/neu.0000000000002551.
19. 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.
20. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312. https://doi.org/10.2196/45312.
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. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, Fournier S. ChatGPT takes on the European exam in Core Cardiology: an artificial intelligence success story? Eur Heart J Digit Health. 2023;4:279–81. https://doi.org/10.1093/ehjdh/ztad029.
23. Danesh A, Pazouki H, Danesh K, Danesh F, Danesh A. The performance of artificial intelligence language models in boardstyle dental knowledge assessment: a preliminary study on Chat-GPT. J Am Dent Assoc. 2023;154:970–4. https://doi.org/10.1016/j.adaj.2023.07.016.
24. Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: evaluating the consistency and accuracy of endodontic question answers. Int Endod J. 2024;57:108–13. https://doi.org/10.1111/iej.13985.
25. 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. https://doi.org/10.1002/JPER.23-0514.
26. 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(2):221–4. https://doi.org/10.11607/ijp.8852.
27. Bhayana R, Fawzy A, Deng Y, Bleakney RR, Krishna S. Retrieval-Augmented Generation for large Language models in Radiology: another leap forward in board examination performance. Radiol. 2024;313(1):e241489. https://doi.org/10.1148/radiol.241489.
28. Grainger R, Dai W, Osborne E, Kenwright D. Medical students create multiple-choice questions for learning in pathology education: a pilot study. BMC Med Educ 18:201. https://doi.org/10.1186/s12909-018-1312-1
29. Lum ZC. Can Artificial Intelligence pass the American Board of Orthopaedic Surgery Examination? Orthopaedic residents Versus ChatGPT. Clin Orthop Relat Res. 2023;481:1623–30. https://doi.org/10.1097/CORR.0000000000002704.
30. Azadi A, Gorjinejad F, Mohammad-Rahimi H, Tabrizi R, Alam M, Golkar M. Evaluation of AI-generated responses by different artificial intelligence chatbots to the clinical decision-making case-based questions in oral and maxillofacial surgery. Oral surg, oral Med, oral pathol. Oral Radiol. 2024;137(6):587–93. https://doi.org/10.1016/j.oooo.2024.02.018.
31. Rokhshad R, Zhang P, Mohammad-Rahimi H, Pitchika V, Entezari N, Schwendicke F. Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: a pilot study. J Dent. 2024;144:104938. https://doi.org/10.1016/j.jdent.2024.104938.
32. Ozden I, Gokyar M, Ozden ME, Sazak Ovecoglu H. Assessment of artificial intelligence applications in responding to dental trauma. Dent Traumatol. 2024;00:1–8. https://doi.org/10.1111/edt.12965.
33. Suchman K, Garg S, Trindade AJ. Chat Generative Pretrained Transformer fails the Multiple-Choice American College of Gastroenterology Self-Assessment Test. Am J Gastroenterol. 2023;118:2280–2. https://doi.org/10.14309/ajg.0000000000002320.
34. Szymkowiak A, Melović B, Dabić M, Jeganathan K, Kundi GS, Information technology, Gen Z. The role of teachers, the internet, and technology in the education of young people. Technol Soc. 2021;65:101565. https://doi.org/10.1016/j.techsoc.2021.101565.

## Publisher's note