ORIGINAL ARTICLE
DOI: 10.5152/EssentDent.2025.25029

ISTANBUL UNIVERSITY CERRAHPASA

ChatGPT-4 Omni's Accuracy in Multiple-Choice Dentistry Questions: A Multidisciplinary and Bilingual Assessment

Makbule Buse Dündar Sarı, Berkant Sezer

Department of Pediatric Dentistry, Çanakkale Onsekiz Mart University School of Dentistry, Çanakkale, Türkiye

Cite this article as: Dündar Sarı MB, Sezer B. ChatGPT-4 Omni's Accuracy in Multiple-Choice Dentistry Questions: A Multidisciplinary and Bilingual Assessment. *Essent Dent.* 2025, 4, 0029, doi: 10.5152/EssentDent.2025.25029.

Abstract

Background: This study evaluated the performance of ChatGPT-4 Omni (ChatGPT-4o) in answering multiple-choice questions from the Dentistry Specialty Examination (DUS), a nationwide exam conducted in Türkiye, assessing knowledge in basic medical and clinical dentistry sciences. Additionally, it examined performance variations based on question language (Turkish vs. English).

Methods: The dataset included 1504 unique questions from publicly available DUS exams (2012-2021) categorized into Basic Medical Sciences (n=514) and Clinical Dentistry Sciences (n=990). Each question was presented to ChatGPT-4o in both Turkish and English, generating 3008 responses. Accuracy was determined using the official answer key. McNemar's test compared accuracy between languages, while chi-square and Bonferroni post-hoc tests assessed differences across disciplines.

Results: ChatGPT-4o showed significantly higher accuracy for English questions (87.8%) than Turkish questions (84.0%) ($P < .001$). In Basic Medical Sciences, accuracy was significantly higher for English questions in Anatomy ($P=.004$) and Physiology ($P=.039$), while Biochemistry achieved 100% accuracy in both languages. In Clinical Dentistry Sciences, English responses were significantly more accurate in Periodontology ($P=.013$), Endodontics ($P=.003$), and Pediatric Dentistry ($P=.005$), whereas Turkish responses performed better in Maxillofacial Radiology ($P=.013$). The highest error rates were in Prosthetic Dentistry (20.1%) for English and Endodontics (18.3%) for Turkish.

Conclusion: ChatGPT-4o demonstrated high accuracy in DUS exam questions, with English responses generally outperforming Turkish ones. Performance varied across disciplines, indicating potential language-based limitations. These findings highlight large language models (LLMs)' potential for dental education while underscoring the need for improvements in language processing and discipline-specific knowledge.

Keywords: Artificial intelligence, ChatGPT, dental education, large language models

What is already known about this topic?

- *The use of artificial intelligence, and more specifically, LLM-based chatbots, is becoming increasingly widespread across various professions.*
- *In the scientific literature, the performance of ChatGPT and other AI-based chatbots in terms of accuracy, readability, and completeness when answering questions from various medical and dental examinations has been investigated.*
- *Current scientific evidence indicates that AI-based chatbots demonstrate varying levels of accuracy in different medical and dental examinations, while also being continuously improved.*

What does this study add to this topic?

- *The accuracy of responses provided by Omni (4o), the latest version of ChatGPT, to questions from different specialties of the Dentistry Specialization Exam (DUS), which plays a crucial role in postgraduate dental education in Türkiye, has been comprehensively evaluated for the first time.*
- *Additionally, the correct response rates of ChatGPT-4o in different languages have been thoroughly compared.*
- *The results indicate that ChatGPT-4o achieved varying levels of accuracy across different branches but demonstrated a generally high accuracy, performing better in answering questions in English.*

INTRODUCTION

Artificial intelligence (AI)-powered chatbots have significantly evolved with advancements in natural language processing (NLP) technologies. This transformation began with Alan Turing's conceptualization of the Turing Test in 1950 and continued with the development of ELIZA by Joseph Weizenbaum in the 1960s, which laid the groundwork for the first interactive chatbot systems. Since the 1990s, rule-based chatbots

Corresponding author: Berkant Sezer E-mail: dt.berkantsezer@gmail.com or berkant.sezer@comu.edu.tr

Received: March 4, 2025
Revision Requested: March 18, 2025
Last Revision Received: April 9, 2025
Accepted: April 25, 2025
Publication Date: July 11, 2025

such as ALICE and Mitsuku have been gradually replaced by large language models (LLMs) that utilize deep learning techniques.[1] Currently, AI-driven models, including ChatGPT, Google Gemini, and Microsoft Copilot, are trained on extensive datasets, allowing them to engage in natural, meaningful, and human-like interactions.[2]

The coronavirus disease 2019 (COVID-19) pandemic led to a rapid transition to remote education, accelerating the development and adoption of digital learning technologies. As traditional in-person education was disrupted, online learning platforms, virtual classrooms, and AI-powered tutoring systems became essential tools for maintaining educational continuity.[3,4] The demand for adaptive and interactive learning experiences increased, fostering advancements in AI-based educational tools.[5] In medical and dental education, remote learning posed unique challenges due to the necessity of hands-on clinical training.[6] However, AI-driven systems, including LLMs, played a crucial role in supplementing theoretical instruction, providing instant access to medical knowledge, and facilitating self-directed learning. As a result, the integration of AI into education has continued to expand beyond the pandemic, shaping the future of digital learning environments.[7]

Among existing LLMs, ChatGPT, developed by OpenAI, stands out for its high accuracy in academic examinations compared to other LLMs. A previous study reported that ChatGPT-4 outperformed Google Bard in the Diagnostic Radiology In-Training Exam, achieving an accuracy rate of 87.11%.[8] Similarly, when assessing the performance of different LLMs in answering questions from the Japan Dental Anesthesiology Society Certification Examination, ChatGPT-4 demonstrated superior accuracy compared to Gemini 1.0. These findings highlight ChatGPT's potential as an educational tool in specialized medical and dental examinations. However, ChatGPT's accuracy is significantly influenced by the language in which questions are posed.[9] As reliance on AI-based knowledge sources increases, evaluating the model's performance across different languages becomes essential. Prior research suggests that ChatGPT provides more precise responses to healthcare-related questions in English compared to other languages.[10]

Large language models are trained on massive multilingual datasets, yet their proficiency in different languages depends on the amount, quality, and diversity of available training data.[11] During pretraining, these models process vast text corpora from publicly accessible sources, including academic literature, online encyclopedias, and structured databases, with English accounting for a disproportionately large share of these data sources.[12] Consequently, LLMs exhibit higher accuracy and coherence in English responses, while performance in less-represented languages often suffers from grammatical inconsistencies, factual inaccuracies, or contextual misunderstandings.[13-15]

In Türkiye, dentists seeking specialization in a particular field must complete a 5-year undergraduate program and take the Dentistry Specialty Examination (DUS) to gain admission into specialty training programs. This multiple-choice exam, administered by the Student Selection and Placement Center (ÖSYM), evaluates candidates' knowledge in basic medical sciences and clinical dentistry. The exam consists of 120 questions, with 40 questions covering basic medical sciences and the remaining 80 questions assessing clinical dentistry science. Due to its broad scope and competitive nature, effective study strategies are crucial for success. With the growing reliance on digital study resources, exploring the potential of AI-assisted tools in DUS exam preparation may contribute to the development of modern educational methodologies.[16]

Previous studies have indicated that the performance of LLMs may vary depending on the language in which questions are presented, as models trained on a greater volume of data in certain languages tend to perform more accurately.[17] The null hypothesis of this study proposes that ChatGPT's accuracy in answering multiple-choice questions related to basic medical sciences and clinical sciences in dentistry does not significantly differ based on whether the questions are presented in Turkish or English.

MATERIALS AND METHODS

Study Design and Data Collection

The aim of this study was to evaluate the performance of ChatGPT-4o (OpenAI, San Francisco, USA, June 2024 version) in answering multiple-choice questions related to basic medical sciences and clinical dentistry sciences in the field of dentistry. Additionally, the model's performance was assessed based on responses to both English and Turkish versions of the questions. The dataset used in this study consisted of questions from the DUS exam, administered nationwide by the ÖSYM in Türkiye since 2012. The study included questions from DUS exams conducted between 2012 and 2021, as these questions were publicly available in the open-access question bank provided by ÖSYM (https://www.osym.gov.tr/TR,15070/dus-cikmis-sorular.html). Exams conducted after 2021 were excluded since their questions were not publicly accessible. Ethics committee approval and informed consent were not needed because this study used only publicly available online materials and not included any human or animal materials.

The DUS exam consists of 120 multiple-choice questions, 40 of which pertain to basic medical sciences and 80 to clinical dentistry sciences. The distribution of questions by discipline is as follows:

- *Basic Medical Sciences (40 questions):* 6 Anatomy, 4 Histology-Embryology, 6 Physiology, 6 Biochemistry, 6 Microbiology, 4 Pathology, 4 Pharmacology, and 4 Medical Biology and Genetics.

- *Clinical Dentistry Sciences (80 questions):* 10 Restorative Dentistry, 10 Prosthetic Dentistry, 10 Maxillofacial Surgery, 10 Maxillofacial Radiology, 10 Periodontology, 10 Orthodontics, 10 Endodontics, and 10 Pediatric Dentistry.

The exclusion criteria for questions were as follows: (1) questions officially annulled by ÖSYM due to structural errors in either the question stem or answer choices, and (2) questions containing visual elements. Large language models exhibit strong performance on text-based multiple-choice questions; however, their accuracy declines significantly when interpreting figures or radiographic images, which require spatial reasoning and domain-specific visual expertise. Apart from these exclusions, all available DUS exam questions from the specified period were included in the study. The Turkish questions were entered into the chatbot without any modifications, while the English translations were reviewed and verified by a native academician English speaker before being inputted into the chatbot.

Query Procedure and Evaluation

The AI-based chatbot used in this study was ChatGPT-4o. Due to the rapid advancement of AI-powered chatbots and frequent updates, the most recent version of ChatGPT available at the time of the study was selected. ChatGPT, developed by OpenAI, was chosen primarily for 2 reasons: it is one of the most widely used AI chatbots globally, and previous studies have reported high accuracy rates.

ChatGPT-4 Omni is a general language model designed to understand user queries on various topics. Given that responses may vary over time due to technological advancements, scientific discoveries, or other factors, a weekly variance assessment was conducted to track potential changes. The latest version of the application (GPT-4o) was accessed on December 10, 2024, at 10:00 AM from Çanakkale, Türkiye (https://openai.com/gpt-4). Each question was entered individually in Turkish to ensure precise understanding and minimize contextual contamination. Every interaction was treated as an independent session to prevent memory effects from influencing subsequent prompts. The full text of the questions, including punctuation and syntax, was preserved during input. No additional prompt optimization or pre-testing was performed, allowing conditions to closely reflect real-world usage.

All selected questions were asked in their original language (Turkish) and again in English after translation. The same process was repeated 2 weeks later, on December 31, 2024, at the same time and location, and responses were documented again. The first set of responses was utilized for the primary analysis, whereas the second set was analyzed to assess reliability. Consistency between the 2 query sessions was measured using Cohen's Kappa (κ), yielding a value of $\kappa = 0.94$. The chatbot's answers were classified as "correct" or "incorrect" based on the official answer key provided by the DUS exam question bank. Accordingly, the primary metric used to evaluate the chatbot's performance was the accuracy rate of its responses to both Turkish and English questions.

Statistical Analysis

Data entries were recorded in a Microsoft Excel spreadsheet (Microsoft Corporation, Redmond, WA, USA). All statistical analyses were conducted using SPSS version 26 (IBM Corp., Armonk, NY, USA). The accuracy of the chatbot's responses to Turkish and English questions across different branches and exam years was compared using the McNemar test. The chi-square test was employed to assess statistical differences in response distributions between Turkish and English questions across branches, and the post-hoc Bonferroni test was used to determine which groups exhibited significant differences. A significance level of .05 was applied for all analyses.

RESULTS

A total of 514 questions from the Basic Medical Sciences were presented to the chatbot. These included 73 Anatomy, 52 Histology-Embryology, 77 Physiology, 78 Biochemistry, 78 Microbiology, 52 Pathology, 52 Pharmacology, and 52 Medical Biology and Genetics questions. Since each question was asked in both Turkish and English, a total of 1018 responses were collected for Basic Medical Sciences. In the Clinical Dentistry Sciences, 990 questions were asked, comprising 127 Restorative Dentistry, 126 Prosthetic Dentistry, 123 Maxillofacial Surgery, 123 Maxillofacial Radiology, 127 Periodontology, 118 Orthodontics, 121 Endodontics, and 124 Pediatric Dentistry questions. Similarly, as all questions were asked in both languages, a total of 1980 responses were recorded for Clinical Dentistry Sciences. Overall, 1504 questions were presented to the chatbot, yielding a total of 3008 responses across both disciplines.

Table 1 presents the accuracy rates of ChatGPT-4o's responses to Turkish and English questions in Basic Medical Sciences and its branches. The chatbot demonstrated a significantly higher accuracy in answering English Anatomy (97.3%) and Physiology (96.1%) questions compared to their Turkish counterparts (84.9% and 87.0%, respectively), with statistical significance ($P = .004$ and $P = .039$, respectively). While no statistically significant difference was observed in accuracy rates between Turkish and English questions for other Basic Medical Sciences branches, the overall accuracy for English questions (96.7%) was significantly higher than for Turkish questions (92.6%) ($P = .001$). Notably, the chatbot answered all Biochemistry questions correctly in both Turkish and English.

Table 2 presents the accuracy rates of ChatGPT-4o's responses to Turkish and English Clinical Dentistry Sciences questions across various branches. The chatbot demonstrated a significantly higher accuracy in answering Turkish Maxillofacial Radiology questions (91.9%) compared to their English counterparts (83.7%) ($P = .013$). Conversely, the accuracy rates for

Table 1. Correct and Incorrect Answers Provided by ChatGPT-4 Omni for Turkish and English Dentistry Specialty Examination questions in basic medical sciences and branches

	Total Answers N (%)	English Answers N (%)	Turkish Answers N (%)	P*
Anatomy				
Correct	133 (91.1)	71 (97.3)	62 (84.9)	**.004**
Incorrect	13 (8.9)	2 (2.7)	11 (15.1)	
Histology-Embryology				
Correct	96 (92.3)	51 (98.1)	45 (86.5)	.070
Incorrect	8 (7.7)	1 (1.9)	7 (13.5)	
Physiology				
Correct	141 (91.6)	74 (96.1)	67 (87)	**.039**
Incorrect	13 (8.4)	3 (3.9)	10 (13)	
Biochemistry				
Correct	156 (100)	78 (100)	78 (100)	1.000
Incorrect	0	0	0	
Microbiology				
Correct	152 (97.4)	77 (98.7)	75 (96.2)	.500
Incorrect	4 (2.6)	1 (1.3)	3 (3.8)	
Patology				
Correct	97 (93.3)	47 (90.4)	50 (96.2)	.250
Incorrect	7 (6.7)	5 (9.6)	2 (3.8)	
Pharmacology				
Correct	97 (93.3)	49 (94.2)	48 (92.3)	1.000
Incorrect	7 (6.7)	3 (5.8)	4 (7.7)	
Medical Biology and Genetics				
Correct	101 (97.1)	50 (96.1)	51 (98.1)	1.000
Incorrect	3 (2.9)	2 (3.9)	1 (1.9)	
Basic Medical Sciences				
Correct	973 (94.6)	497 (96.7)	476 (92.6)	**.001**
Incorrect	55 (5.4)	17 (3.3)	38 (7.4)	

Bold values mean statistically significance.
N, number, *McNemar Test.

Table 2. Correct and Incorrect Answers Provided by ChatGPT-4 Omni for Turkish and English Dentistry Specialty Examination Questions In Clinical Dentistry Sciences and Branches

	Total Answers N (%)	English Answers N (%)	Turkish Answers N (%)	P*
Restorative Dentistry				
Correct	231 (90.9)	114 (89.8)	117 (92.1)	.549
Incorrect	23 (8.9)	13 (10.2)	10 (7.9)	
Prosthetic Dentistry				
Correct	178 (70.7)	89 (70.7)	89 (70.7)	1.000
Incorrect	74 (29.3)	37 (29.3)	37 (29.3)	
Maxillofacial Surgery				
Correct	220 (89.4)	109 (88.6)	111 (90.2)	.754
Incorrect	26 (10.6)	14 (11.4)	12 (9.8)	
Maxillofacial Radiology				
Correct	216 (87.8)	103 (83.7)	113 (91.9)	**.013**
Incorrect	30 (12.2)	20 (16.3)	10 (9.1)	
Periodontology				
Correct	215 (84.6)	113 (89)	102 (80.3)	**.013**
Incorrect	39 (15.4)	14 (11)	25 (19.7)	
Orthodontics				
Correct	182 (77.1)	95 (80.5)	87 (73.7)	.134
Incorrect	54 (22.9)	23 (19.5)	31 (26.3)	
Endodontics				
Correct	172 (69.4)	94 (77.7)	78 (64.5)	**.003**
Incorrect	76 (30.6)	23 (22.3)	43 (35.5)	
Pediatric Dentistry				
Correct	196 (79)	106 (85.5)	90 (72.6)	**.005**
Incorrect	52 (21)	18 (14.5)	34 (27.4)	
Clinical Dentistry Sciences				
Correct	1610 (81.3)	823 (83.1)	787 (79.5)	**.004**
Incorrect	370 (18.7)	167 (16.9)	203 (20.5)	

Bold values mean statistically significance.
N, number, *McNemar test.

English Periodontology (89.0%), Endodontics (77.7%), and Pediatric Dentistry (85.5%) questions were significantly higher than those for their Turkish equivalents (80.3%, 64.5%, and 72.6%, respectively) ($P = .013$, $P = .003$, and $P = .005$, respectively). While no statistically significant differences were observed in accuracy rates between Turkish and English questions for other Clinical Dentistry Sciences branches, the overall accuracy for English Clinical Dentistry Sciences questions (83.1%) was significantly higher than for Turkish questions (79.5%) ($P = .004$). The lowest accuracy rate was observed for Turkish Endodontics questions (64.5%), while the highest accuracy rate was found for Turkish Restorative Dentistry questions (92.1%).

When all questions were considered, the chatbot's accuracy for English questions (87.8%) was significantly higher than for Turkish questions (84.0%) ($P < .001$) (Table 3).

The accuracy rates of the chatbot's responses to Turkish and English DUS exam questions from different years are presented in Table 4. No statistically significant differences were observed in accuracy rates between Turkish and English questions for any exam year except 2021, where the chatbot's accuracy for English questions (91.1%) was

Table 3. Correct and Incorrect Answers Provided by ChatGPT-4o for Turkish and English all Dentistry Specialty Examination Questions

	Total Answers N (%)	English Answers N (%)	Turkish Answers N (%)	P*
All Questions				
Correct	2583 (85.9)	1320 (87.8)	1263 (84)	**< .001**
Incorrect	425 (14.1)	184 (12.2)	241 (16)	

Bold values mean statistically significance.
N, number, *McNemar Test.

Table 4. Correct and Incorrect Answers Provided by ChatGPT-4 Omni for Turkish and English Dentistry Specialty Examination Questions in Different Years

	Total Answers N (%)	English Answers N (%)	Turkish Answers N (%)	*P**
2012				
Correct	414 (87)	211 (88.7)	203 (85.3)	.170
Incorrect	62 (13)	27 (11.3)	35 (14.7)	
2013				
Correct	393 (84.7)	201 (86.6)	192 (82.8)	.108
Incorrect	71 (15.3)	31 (13.4)	40 (17.2)	
2014				
Correct	412 (86.9)	210 (88.6)	202 (85.2)	.230
Incorrect	62 (13.1)	27 (11.4)	35 (14.8)	
2015				
Correct	214 (92.2)	109 (94)	105 (90.5)	.219
Incorrect	18 (7.8)	7 (6)	11 (9.5)	
2016				
Correct	203 (84.6)	101 (84.1)	102 (85)	1.000
Incorrect	37 (15.4)	19 (15.9)	18 (15)	
2017				
Correct	193 (86.2)	97 (86.6)	96 (85.7)	1.000
Incorrect	31 (13.8)	15 (13.4)	16 (14.3)	
2018				
Correct	168 (77.8)	86 (79.6)	82 (75.9)	.388
Incorrect	48 (22.2)	22 (20.4)	26 (24.1)	
2019				
Correct	183 (83.2)	95 (86.4)	88 (80)	.143
Incorrect	37 (16.8)	15 (13.6)	22 (20)	
2020				
Correct	209 (87.8)	108 (90.8)	101 (84.9)	.143
Incorrect	29 (12.2)	11 (9.1)	18 (15.1)	
2021				
Correct	194 (86.6)	102 (91.1)	92 (82.1)	**.031**
Incorrect	30 (13.4)	10 (8.9)	20 (17.9)	

Bold values mean statistically significance.
N, number, *McNemar test.

significantly higher than for Turkish questions (82.1%) ($P = .031$).

The distribution of accuracy rates for responses to questions across different branches and languages is presented in Table 5. Among the incorrect responses to English questions, the highest proportion was observed in Prosthetic Dentistry (20.1%), while the lowest was in Biochemistry (0%). Similarly, for incorrect responses to Turkish questions, the highest proportion was in Endodontics (18.3%), whereas Biochemistry remained the only discipline with no incorrect responses. A statistically significant difference was observed in the distribution of accuracy rates among different branches for both English and Turkish questions ($P < .001$). Table 6 presents the accuracy distributions for the 2 main disciplines in both languages. Similarly, a statistically significant difference was found in the distribution of accuracy of responses ($P < .001$).

DISCUSSION

The evaluation of responses provided by LLMs-based chatbots in terms of accuracy, readability, and completeness has gained increasing attention in the scientific literature. Studies involving OpenAI's ChatGPT, which is regularly updated, represent a prominent group in this area. The accuracy of chatbot responses to various exam questions, particularly in the field of health sciences, continues to be investigated across different countries and languages. In this study, the authors conducted a comparative analysis of the accuracy of ChatGPT-4o, the most recent version, in answering all Turkish and English questions from various disciplines included in the DUS exam, an officially administered exam for postgraduate dental education in Türkiye. The findings of this study indicate that ChatGPT-4o demonstrated a higher overall accuracy rate in responding to questions in English compared to Turkish, and its accuracy varied across different disciplines. Thus, the null hypothesis is rejected.

Previous research has demonstrated that ChatGPT is capable of providing highly accurate responses to medical-related academic questions. In their study evaluating ChatGPT's performance in the United States Medical Licensing Examination (USMLE), Kung et al.[18] concluded that the model responded to medical questions with a high level of accuracy. Similarly, Tian et al.[19] examined the applications of ChatGPT and similar LLMs in biomedical and healthcare fields, highlighting that these models have made significant advancements in medical text generation compared to earlier approaches.

A recent study conducted in Türkiye found that ChatGPT outperformed Google Bard and Microsoft Copilot in answering oral radiology questions from the DUS exam.[20] Similarly, in an evaluation comparing the performance of GPT-3.5, GPT-4, GPT-4o, and Google Bard on the USMLE, the Professional and Linguistic Assessments Board (PLAB), the Hong Kong Medical Licensing Examination (HKMLE), and the National Medical Licensing Examination (NMLE), all questions were derived from official exam question banks. The results revealed that GPT-4o outperformed the other 3 LLMs in these medical licensing exams.[21] Conversely, Avşar et al.[22]examined DUS exam prosthetic dentistry questions from 2012 to 2021 by presenting them simultaneously to ChatGPT-3.5 and Google's Gemini. Their findings indicated that both chatbots exhibited similar knowledge levels, with their accuracy rates in answering prosthetic dentistry-related questions being relatively limited.[22] Studies like these provide insights into how different AI-based chatbots perform in medical licensing exams, facilitating a better understanding of their potential roles in medical education.

A meta-analysis comprising 23 studies evaluating the performance of different ChatGPT versions in national licensing exams for medicine, pharmacy, dentistry, and nursing reported varying accuracy levels, ranging from 36% to

Table 5. Distribution of correct and Incorrect Answers Provided by ChatGPT-4 Omni to Turkish and English Dentistry Specialty Examination Questions Among Different Branches

	English Answers			Turkish Answers		
	Incorrect N (%)	Correct N (%)	*P**	Incorrect N (%)	Correct N (%)	*P**
Anatomy	2 (1.1)	71 (5.4)	**< .001**	11 (4.6)	62 (4.9)	**< .001**
Histology-Embriology	1 (0.5)	51 (3.9)		7 (2.9)	45 (3.6)	
Physiology	3 (1.6)	74 (5.6)		10 (4.1)	67 (5.3)	
Biochemistry	0 (0)	78 (5.9)		0 (0)	78 (6.2)	
Microbiology	1 (0.5)	77 (5.8)		3 (1.2)	75 (5.9)	
Patology	5 (2.7)	47 (3.6)		2 (0.8)	50 (4)	
Pharmacology	3 (1.6)	49 (3.7)		4 (1.7)	48 (3.8)	
Medical Biology and Genetics	2 (1.1)	50 (3.8)		1 (0.4)	51 (4)	
Restorative Dentistry	13 (7.1)	114 (8.6)		10 (4.1)	117 (9.3)	
Prosthetic Dentistry	37 (20.1)	89 (6.7)		37 (15.4)	89 (7)	
Maxillofacial Surgery	14 (7.6)	109 (8.3)		12 (5)	111 (8.8)	
Maxillofacial Radiology	20 (10.9)	103 (7.8)		10 (4.1)	113 (8.9)	
Periodontology	14 (7.6)	113 (8.6)		25 (10.4)	102 (8.1)	
Orthodontics	23 (12.5)	95 (7.2)		31 (12.9)	87 (6.9)	
Endodontics	28 (15.2)	94 (7.1)		44 (18.3)	78 (6.2)	
Pediatric Dentistry	18 (9.8)	106 (8)		34 (14.1)	90 (7.1)	
Total	184 (100)	1320 (100)		241 (100)	1263 (100)	

Bold values mean statistically significance.
N, number, *chi-square test with post-hoc Bonferroni test.

77% for ChatGPT-3.5 and from 64.4% to 100% for GPT-4. Subgroup analyses revealed that GPT-4 provided significantly higher accuracy in correct responses compared to its predecessor, ChatGPT-3.5.[23] Similarly, Meyer et al.[24] assessed the medical competence of GPT-3.5 and GPT-4 using 937 multiple-choice questions from 3 written German medical licensing exams administered in October 2021, April 2022, and October 2022. GPT-4 scored an average of 85%, ranking within the 92.8th percentile for the October 2021 exam, the 99.5th percentile for the April 2022 exam, and the 92.6th percentile for the October 2022 exam. In contrast, GPT-3.5 passed only 1 of the 3 exams.[24] Danesh et al.,[25] in their study comparing ChatGPT-3.5 and ChatGPT-4 in dental knowledge assessment, found that the newer versions of ChatGPT exhibited enhanced capabilities in generating dental content.[25] This improvement was a key factor in choosing ChatGPT-4o for the present study, where it achieved an overall accuracy rate of 85.9%, demonstrating satisfactory performance.

Despite these promising results, ChatGPT's performance in the DUS exam and its discipline-specific accuracy levels remain underexplored. The current study identified significant differences in ChatGPT-4o's accuracy across basic medical sciences and clinical dentistry sciences. The model exhibited higher accuracy in fundamental sciences, which could be attributed to the universal nature of these subjects and the extensive training LLMs receive in these domains. For instance, ChatGPT-4o answered all biochemistry questions correctly in both Turkish and English, likely due to biochemistry being rooted in absolute scientific principles rather than subjective clinical scenarios. In contrast, the model's lower accuracy in clinical dentistry sciences suggests potential challenges in clinical decision-making processes. The findings indicate relatively lower performance in prosthetic dentistry, endodontics, and pediatric dentistry, which may stem from the interpretative nature of clinical questions and their specificity to real-world dental practice. This discrepancy suggests that while AI models excel at theoretical

Table 6. Distribution of Correct and Incorrect Answers Provided by ChatGPT-4 Omni to Turkish and English Dentistry Specialty Examination Questions Among Main Disciplines

	English Answers			Turkish Answers		
	Incorrect N (%)	Correct N (%)	*P**	Incorrect N (%)	Correct N (%)	*P**
Basic Medical Sciences	17 (9.2)	497 (37.7)	**< .001**	38 (15.8)	476 (37.7)	**< .001**
Clinical Dentistry Sciences	167 (90.8)	823 (62.3)		203 (84.2)	787 (62.3)	
Total	184 (100)	1320 (100)		241 (100)	1263 (100)	

Bold values mean statistically significance.
N, number, *chi-square test with post-hoc Bonferroni test.

knowledge, they may require further exposure to case-based learning to enhance their ability to process and respond to clinical scenarios.

Previous studies suggest that ChatGPT exhibits superior accuracy in answering medical terminology-based questions in English compared to other languages. Joshi et al.[26] conducted a study comparing ChatGPT's responses to vaccine hesitancy-related questions in English, Spanish, and French, concluding that its English responses were more detailed and consistent. Similarly, Schulz et al.[27] demonstrated that ChatGPT exhibited greater accuracy in medical terminology when responding in English, while its accuracy declined in other languages. This discrepancy is likely due to the model's training data being predominantly composed of English-language sources. If language variation significantly affects the accuracy of chatbot responses, this could suggest that LLMs are trained more extensively in certain languages, influencing their performance across different linguistic contexts. The findings of the present study confirm that ChatGPT-4o's accuracy in English questions was significantly higher than in Turkish questions. This trend was particularly evident in anatomy and physiology, where English questions yielded higher accuracy rates than their Turkish counterparts. These results support the hypothesis that ChatGPT may perform better in English due to greater exposure to English-language training data.

Additionally, differences in Turkish medical terminology compared to English could influence ChatGPT's response accuracy. Aytekin and Karabina[28] found that ChatGPT-4 was more effective than ChatGPT-3.5 in disambiguating homonyms in Turkish, suggesting that language-specific variations can impact AI model performance in multilingual contexts. Interestingly, ChatGPT-4o exhibited higher accuracy in Turkish than English in oral and maxillofacial radiology questions, which may be due to terminological distinctions unique to certain dental specialties in Turkish.

Medical and dental specialty exams frequently incorporate radiographic images, clinical photographs, histological slides, and anatomical illustrations, which were excluded from this study as ChatGPT-4o only processes text-based inputs. The inability to evaluate image-based questions represents a limitation, as it restricts the assessment of the model's clinical decision-making capabilities. Future research could explore OpenAI's DALL·E-based image analysis tools or multimodal models such as GPT-4V to assess AI-driven interpretation of visual medical data.

This study analyzed DUS exam questions from 2012 to 2021. Expanding the dataset to include more recent exam questions could provide a clearer picture of ChatGPT-4o's alignment with contemporary examination formats. Furthermore, as AI models undergo continuous updates, future iterations of ChatGPT may yield different results. While this study focused exclusively on multiple-choice questions, further research could evaluate ChatGPT's ability to handle open-ended clinical decision-making scenarios, offering deeper insights into its potential applications in real-world dental practice.

The most significant strength of the study is the evaluation of a very large dataset. The inclusion of all exam questions from the years 2012-2021 has expanded the sample size and increased the generalizability of the study results. Additionally, the inclusion of questions from various specialties and the evaluation of the latest ChatGPT model are other notable strengths. On the other hand, the most important limitation of the study is that the evaluation was conducted in only 2 different languages. Furthermore, assessing the accuracy of just 1 chatbot is another limitation.

The findings suggest that ChatGPT's high accuracy rates make it a potentially useful tool for students preparing for the DUS exam. However, it is important to recognize that ChatGPT functions primarily as an information retrieval tool and does not independently facilitate the learning process. More research is needed to determine how AI-powered systems can be optimally integrated into DUS exam preparation. For instance, feedback mechanisms could be incorporated to explain why ChatGPT provides incorrect responses, allowing students to understand and learn from their mistakes. Additionally, AI-driven education platforms could develop personalized study plans to help students focus on their weaker areas.

As medical knowledge continues to evolve, the ability of AI models to rapidly incorporate new developments remains a critical issue. Given the ongoing advancements in treatment protocols, pharmacology, and clinical approaches, future versions of ChatGPT may require more frequent updates incorporating the latest medical literature. To maintain relevance, AI models should have direct access to up-to-date medical databases, ensuring that they provide the most current and evidence-based information.

CONCLUSION

This study demonstrated that ChatGPT-4o provides highly accurate responses in dentistry but performs better in English than in Turkish. This emphasizes the need for AI-supported learning platforms to incorporate more Turkish training data for improved accuracy. The model excelled in basic medical sciences like biochemistry and physiology, where knowledge is absolute, but showed lower accuracy in clinical dentistry sciences such as prosthetic dentistry, endodontics, and pediatric dentistry, likely due to the complexity of clinical decision-making. While AI models are strong in theoretical knowledge, further refinements are needed to enhance their application in clinical reasoning and problem-solving within dental specialties.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

Ethics Committee Approval: N/A.

Informed Consent: N/A.

Peer-review: Externally peer reviewed.

Author Contributions: Concept – M.B.D.S., B.S.; Design – M.B.D.S., B.S.; Supervision – B.S.; Resources – M.B.D.S.; Materials – M.B.D.S., B.S.; Data Collection and/or Processing – M.B.D.S.; Analysis and/or Interpretation – B.S.; Literature Search – M.B.D.S., B.S.; Writing Manuscript – M.B.D.S., B.S.; Critical Review – B.S.

Declaration of Interests: The authors declare that they have no competing interest.

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

REFERENCES

1. Yigci D, Eryilmaz M, Yetisen AK, Tasoglu S, Ozcan A. Large language model-based chatbots in higher education. *Adv Intell Syst*. 2024;7(3):2400429.
2. Rossettini G, Rodeghiero L, Corradi F, et al. Comparative accuracy of ChatGPT-4, Microsoft Copilot and Google Gemini in the Italian entrance test for healthcare sciences degrees: a cross-sectional study. *BMC Med Educ*. 2024;24(1):694. [CrossRef]
3. Akbeyaz Şivet E, Altıntaş S, Atmaca N, Kargul B. COVID-19 Pandemisinin diş Hekimliği Öğrencilerinin anksiyete bozukluğu ve kariyer Seçimlerine Etkisinin Değerlendirilmesi. *ADO Klin Bilimler Derg*. 2024;13(3):503-516. [CrossRef]
4. Şen Yavuz B, Güneyligil Kazaz T, Akbeyaz Şivet E, Kargul B. Prediction of the spread of the COVID-19 pandemic with Google searches: an infodemiological approach. *ADO Klin Bilimler Derg*. 2024;13(2):358-367. [CrossRef]
5. Ibrahim U, Argungu JI, Mungadi IM, Yeldu AS. E-learning and remote education technologies: lessons from the pandemic. *Int J Educ Life Sci*. 2023;1(3):159-174.
6. Garcia PPNS, de Souza Ferreira F, Pazos JM. Stress among dental students transitioning from remote learning to clinical training during coronavirus disease 2019 pandemic: a qualitative study. *J Dent Educ*. 2022;86(11):1498-1504. [CrossRef]
7. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med Educ*. 2023;9(1):e48291. [CrossRef]
8. Patil NS, Huang RS, van der Pol CB, Larocque N. Comparative performance of ChatGPT and Bard in a text-based radiology knowledge assessment. *Can Assoc Radiol J*. 2024;75(2):344-350. [CrossRef]
9. Fujimoto M, Kuroda H, Katayama T, et al. Evaluating large language models in dental anesthesiology: a comparative analysis of ChatGPT, Claude 3 Opus, and Gemini 1.0 on the Japanese Dental Society of Anesthesiology Board Certification Exam. *Cureus*. 2024;16(9):e70302.
10. Fang C, Wu Y, Fu W, et al. How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language. *PLoS Digit Health*. 2023;2(12):e0000397. [CrossRef]
11. Khanna P, Dhillon G, Buddhavarapu V, Verma R, Kashyap R, Grewal H. Artificial intelligence in multilingual interpretation and radiology assessment for clinical language evaluation. *J Pers Med*. 2024;14(9):923. [CrossRef]
12. Wang D, Zhang S. Large language models in medical and healthcare fields: applications, advances, and challenges. *Artif Intell Rev*. 2024;57(11):299. [CrossRef]
13. Sallam M, Al-Mahzoum K, Alshuaib O, et al. Language discrepancies in the performance of generative artificial intelligence models: an examination of infectious disease queries in English and Arabic. *BMC Infect Dis*. 2024;24(1):799. [CrossRef]
14. Chang E, Sung S. Use of SNOMED CT in large language models: scoping review. *JMIR Med Inform*. 2024;12(1):e62924. [CrossRef]
15. Qiu P, Wu C, Zhang X, et al. Towards building multilingual language model for medicine. *Nat Commun*. 2024;15(1):8384. [CrossRef]
16. Yılmaz C, Erdem RZ, Uygun LA. Artificial intelligence knowledge, attitudes and application perspectives of undergraduate and specialty students of faculty of dentistry in Turkey: an online survey research. *BMC Med Educ*. 2024;24(1):1149. [CrossRef]
17. Watari T, Takagi S, Sakaguchi K, et al. Performance comparison of ChatGPT-4 and Japanese medical residents in the general medicine in-training examination: comparison Study. *JMIR Med Educ*. 2023;9:e52202. [CrossRef]
18. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. [CrossRef]
19. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9(1):e45312. [CrossRef]
20. Tassoker M. ChatGPT-4 Omni's superiority in answering multiple-choice oral radiology questions. *BMC Oral Health*. 2025;25(1):173. [CrossRef]
21. Chen Y, Huang X, Yang F, et al. Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: a comparison study. *BMC Med Educ*. 2024;24(1):1372. [CrossRef]
22. 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. *Turkiye Klinikleri J Dental Sci*. 2024;30(4):668-673. [CrossRef]
23. 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(1):1013. [CrossRef]
24. 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. [CrossRef]
25. 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(11):970-974. [CrossRef]

26. Joshi S, Ha E, Rivera Y, Singh VK. ChatGPT and vaccine hesitancy: a comparison of English, Spanish, and French Responses Using a Validated Scale. *AMIA Jt Summits Transl Sci Proc*. 2024;2024:266-275.
27. Abdulnazar A, Roller R, Schulz S, Kreuzthaler M. Large language models for clinical text cleansing enhance medical concept normalization. *IEEE Access*. 2024;12:147981-147990. [CrossRef]
28. Aytekin Ç, Karabina TB. ChatGPT'nin farklı büyük dil modelleri performanslarının türkçedeki eş adlı kelimeler üzerinden incelenmesi. *Istanb Aydın Univ Sosyal Bilimler Derg*. 2024;16(3):365-390.