Acta Medica Europa

ISSN Online: 2996-1211 doi: 10.5281/zenodo.17826326

Performance of Leading Large Language Models in Answering Questions in Turkish Dentistry Specialist Education Entrance Exams

Ömer Ekici, DDS, PhD[1], İsmail Çalışkan[2]

1. Assoc. Prof. Dr., Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Afyonkarahisar Health Sciences University, Afyonkarahisar, Turkey.
2. Research Assistant, Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Afyonkarahisar Health Sciences University, Afyonkarahisar, Turkey.

Article Info

Received: 18 August 2025

Revised: 4 December 2025

Accepted: 5 December 2025

Published: 5 December 2025

Keywords:

Artificial Intelligence, Large Language Model, Dentistry Education, Dentistry Specialty Education Entrance Exam.

Corresponding author:

Ömer Ekici, DDS, PhD

Assoc. Prof. Dr., Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Afyonkarahisar Health Sciences University, Afyonkarahisar, Turkey.

dromerekici@hotmail.com

ABSTRACT

The use of artificial intelligence (AI) based large language models (LLMs) in medical and dental education has been increasing in recent years. The purpose of this study is to comparatively examine the accuracy of the answers given by leading LLMs to the questions in the Turkish Dentistry Specialization Education Entrance Exam (TDSE). 1503 questions without figures and pictures from 13 exams published on the official website of the Student Selection and Placement Center (SSPC) were included in the study. The performance of 6 LLMs, namely GPT-4o GPT-4, GPT-3.5, Co-pilot, Gemini 2.0 and Gemini 1.5, was evaluated in answering the questions. Questions were directed to the LLMs simultaneously by a single operator. Chi-square analysis was used to compare the correct response rates between the LLMs. The order of performance of LLMs in terms of correct response rates to all questions was as follows: GPT-4o (92.2%), GPT-4 (88.7%), Co-pilot (87.3%), Gemini 2.0 (86.4%), GPT-3.5 (81.5%) and Gemini 1.5 (76.1%). LLMs performed better in basic sciences than in clinical sciences. Statistically significant differences were observed between correct response rates of LLMs according to years and branches. The LLMs evaluated in this study performed quite well according to the literature. The best performance was shown by GPT-4o, and the worst performance was shown by Gemini 1.5. The findings show that LLMs have the potential to be used as a supportive educational tool in basic and clinical dentistry education, despite certain limitations.

INTRODUCTION

Artificial intelligence (AI)-based chatbots and large language models (LLMs) are advanced systems that provide information by interacting with people through written, audio, and visual communication channels (1). Thanks to advances in natural language processing and deep learning techniques, these technologies have become more intelligent, consistent, and usable in various fields (2,3). Especially in the field of dentistry, AI-based chatbots are used for various purposes such as supporting diagnostic processes, accelerating clinical decision-making, creating digital data records, performing radiographic image analysis, and reducing error rates in treatment processes (4,5). One of the most common areas of use of AI-supported chatbots is medical and dental education (6-8).

Natural language processing (NLP), the ability of computers to understand and process human language, is focused on making human-computer interaction more efficient and useful. LLMs, which use NLP techniques to produce responses to text-based inputs, have received special attention since the introduction of OpenAI's ChatGPT (Generative Pre-trained Transformer) in November 2022 (9). ChatGPT-3.5, developed by OpenAI in November 2022 and having 175 billion parameters, attracted attention with its natural language processing capabilities, Copilot (formerly Bing Chat) was released by Microsoft in February 2023, followed by ChatGPT-4 in March 2023, and Gemini (formerly Bard) by Google in May 2023 (10-12). Thanks to their high capacity to analyze and interpret large data sets, LLMs play an important role in the process of producing

solutions for complex problems and continue to develop rapidly every day (6).

The use of new generation chatbots such as ChatGPT, Gemini and Co-pilot developed by OpenAI, Google and Microsoft has become widespread in medical and academic fields (2). The knowledge generation processes of these chatbots are based on NLP and provide predictive answers based on the information they learn from training data (13). The knowledge levels and correct response capacities of these systems are increasingly becoming a subject of research (14). Various studies have evaluated the effectiveness of ChatGPT in solving text-based questions in national exams for doctors, nurses and pharmacists (15,16). It has been shown that these systems have demonstrated successful performance in national medical exams administered in the USA, Japan and China and have the capacity to understand complex medical information, and important findings have been presented that they can be used as an educational support tool (17-19). However, there are various questions about the accuracy of the answers they give in areas that require specific knowledge, such as health sciences and especially dentistry (20,21). Some recent studies with small sample sizes and branch-based evaluations evaluating the performance of LLMs in answering questions asked in the Turkish Dentistry Specialization Education Entrance Exam (TDSE) have been included in the literature (22,23). However, no comprehensive study has been conducted to evaluate the performance of several leading LLMs on published questions asked in all TDSEs. The aim of this study is to evaluate the accuracy of responses given by different AI-powered chatbots (ChatGPT-3.5, ChatGPT-4, ChatGPT-4o, Co-Pilot, Gemini 1.5 and Gemini 2.0) in common use today to questions asked in TDSEs over a 10-year period and to compare the performances of these LLMs in the basic sciences and clinical sciences branches of dentistry.

METHODOLOGY

Ethics

Since our study was not conducted on humans or human samples but on a publicly accessible website, ethics committee approval was not required.

LLMs

Six LLMs were evaluated in this study: GPT-3.5, GPT-4 and GPT-4o (OpenAI, San Francisco, California, USA), Co-pilot (Microsoft, Redmond, Washington, USA), Gemini 1.5, and Gemini 2.0 (Google LLC, Mountain View, California, USA).

Turkish Dentistry Specialization Education Entrance Exam (TDSE)

In Türkiye, specialization training in dentistry began in 2011 with the legal regulation as 8 separate branches: oral and maxillofacial surgery, oral and maxillofacial radiology, pediatric dentistry, endodontics, orthodontics, periodontics, prosthetic dentistry and restorative dentistry. Oral pathology was added to these specialization branches in 2018.

TDSE is a centralized exam administered by the Student Selection and Placement Center (SSPC) for dentists who want to receive specialization education at university dentistry faculties. The TDSE exam, first implemented in the spring semester of 2012, was held twice a year in September and April between 2012 and 2014, once a year between 2015 and 2022, and twice a year from 2023 onwards. TDSE consists of 120 multiple-choice questions, 40 from basic sciences and 80 from clinical sciences (Table 1).

Table 1. Distribution of TDSE questions by branches.

CLINICAL SCIENCES	n	%	BASIC SCIENCES	n	%
Restorative dentistry	10	12.5	Anatomy	6	15
Prosthetic dentistry	10	12.5	Physiology	6	15
Oral and maxillofacial surgery	10	12.5	Histology and embryology	4	10
Oral and maxillofacial radiology	10	12.5	Medical biochemistry	6	15
Periodontology	10	12.5	Medical microbiology	6	15
Endodontics	10	12.5	Medical pathology	4	10
Pediatric dentistry	10	12.5	Medical pharmacology	4	10
Ortodontics	10	12.5	Medical biology and genetics	4	10
TOTAL	80	100	**TOTAL**	40	100

TDSE: Turkish Dentistry Specialization Education Entrance Exam; n: number of questions; %: percentage.

Inclusion and Exclusion Criteria

TDSE was conducted 17 times between 2012-2024. Thirteen TDSEs were conducted between 2012-2021 and published on the SSPC official website, and 1560 questions asked in these exams were included in this study (24). However, since SSPC has not published TDSE questions since 2022, exam questions from 2022 and later were not included in the study. In this study, 14 questions that were canceled by SSPC and 43 questions containing shapes were not included in the study. As a result, the final analysis included 1503 questions.

Question Directing and Evaluation Method to LLMs

In the study, 1503 multiple-choice questions were directed simultaneously to ChatGPT-3.5, ChatGPT-4, ChatGPT-4o, Co-pilot, Gemini 1.5, and Gemini 2.0 models by a single operator between February 10 and February 20, 2025. Before questions were directed to LLMs, new accounts were created for each LLM evaluated in this study. Cookies and internet history were deleted before queries were made. It was planned to re-ask questions in case of freezing or delay in answering, but no such problem was experienced in any LLM. LLMs were tested using default configurations without any parameter adjustments or additional prompts. Multiple choice questions were directed to LLMs using the same format, wording and order without any modifications. After grouping the questions according to years, the questions were directed to LLMs in chronological order. Original TDSE questions were presented in Turkish and manually entered into the chat interfaces of LLMs one by one to receive answers. Questions were entered only once to avoid biases from repeated exposure or learning effects.

LLMs' answers were carefully reviewed and were considered "correct" if they matched the official answers provided by the SSPC (24). The answers of LLMs were compared according to years and branches.

Data Analysis

All statistical analyses were performed using SPSS statistical program, version 27 (SPSS Inc., Chicago, IL, USA). Standard descriptive statistics were used for statistical analysis. Mean and standard deviation values were given in descriptive statistics of continuous data, and number and percentage values were given in nominal data. Pearson Chi-Square analysis was used to compare correct answer rates between LLMs. After the significant findings obtained from Pearson Chi-Square analysis, pairwise comparisons between LLMs evaluated based on their responses to all TDSE questions were performed using Post-Hoc Chi-Square analysis with Bonferroni correction. This Post-Hoc analysis was performed for all questions. For Pearson Chi-Square tests, $p<0.05$ was considered statistically significant, and for Post-Hoc comparisons, the significance level was set as $p<0.0041$ after Bonferroni correction.

RESULTS

A total of 1560 questions were analyzed, excluding 14 canceled and 57 questions containing 43 figures. The correct answer rates of the six LLMs used in the study were compared according to years and branches.

Performance of LLMs by Year

The minimum and maximum correct answer rates of LLMs according to years in basic sciences were as follows: GPT-4 (87.2%-100%), GPT-3.5 (81.6%-97.5%), GPT-4o (94.9%-100%), Co-pilot (95%-100%), Gemini 1.5 (76.3%-92.5%) and Gemini 2.0 (89.7%-97.5%). In clinical sciences, they were as follows: GPT-4 (69.2%-93.1%), GPT-3.5 (70%-84.8%), GPT-4o (81.4%-93.1%), Co-pilot (75.7%-90%), Gemini 1.5 (61%-82.9%) and Gemini 2.0 (71.4%-85.7%) (Table 2).

While GPT-4o answered 7 out of 13 exams in basic sciences with 100% correct answers, Co-pilot and GPT-4 answered 2 correctly. GPT-3.5, Gemini 1.5 and Gemini 2.0 did not answer any exam with 100% correct answers. In clinical sciences questions, no LLM could reach 100% correct answer rate in any exam (Table 2).

There were statistically significant differences between the correct answer rates of LLMs except for two of the 13 exams (2014/2 and 2015 exams). In some exams, only in basic sciences (2016, 2018 and 2020 exams), in some exams only in clinical sciences (2012/I, 2012/II, 2013/I, 2013/II, 2014/I and 2017), and in some exams both (2019 and 2021 exams) there were significant differences between the answer percentages of LLMs (Table 2).

Table 2. Comparison of correct answers of LLMs by year.

		Correct Response Rate n(%)						
YEARS	**TYPE of SCIENCE**	**Co-pilot n(%)**	**GPT-4 n(%)**	**GPT-3.5 n(%)**	**Gemini 1.5 n(%)**	**Gemini 2.0 n(%)**	**GPT-4o n(%)**	**p value**
2012/1	**Basic sciences**	38(95)	37(92.5)	36(90)	36(90)	38(95)	39(97.5)	0.708
	Clinical sciences	65(83.3)	54(69.2)	59(75.6)	54(69.2)	61(78.2)	68(87.2)	0.037*
2012/2	**Basic sciences**	39(97.5)	39(97.5)	39(97.5)	36(90)	39(97.5)	40(100)	0.203
	Clinical sciences	63(78.8)	61(76.3)	59(73.8)	52(65)	67(83.8)	73(91.3)	0.002**
2013/1	**Basic sciences**	38(95)	36(90)	37(92.5)	37(92.5)	38(95)	40(100)	0.514
	Clinical sciences	63(81.8)	61(79.2)	54(70.1)	47(61)	65(84.4)	69(89.6)	0.000**
2013/2	**Basic sciences**	38(97.4)	34(87.2)	33(84.6)	32(82.1)	35(89.7)	37(94.9)	0.205
	Clinical sciences	60(77.9)	71(92.2)	59(76.6)	48(62.3)	62(80.5)	68(88.3)	0.000**
2014/1	**Basic sciences**	38(95)	39(97.5)	38(95)	35(87.5)	39(97.5)	39(97.5)	0.277
	Clinical sciences	63(80.8)	71(91)	59(75.6)	56(71.8)	64(82.1)	70(89.7)	0.010*
2014/2	**Basic sciences**	38(95)	39(97.5)	37(92.5)	35(87.5)	38(95)	39(97.5)	0.397
	Clinical sciences	65(82.3)	67(84.8)	59(74.7)	57(72.2)	67(84.8)	69(87.3)	0.084
2015	**Basic sciences**	39(97.5)	38(95)	37(92.5)	37(92.5)	39(97.5)	39(97.5)	0.736
	Clinical sciences	64(84.2)	66(86.8)	62(81.6)	63(82.9)	65(85.5)	68(89.5)	0.789
2016	**Basic sciences**	39(97.5)	39(97.5)	37(92.5)	32(80)	37(92.5)	39(97.5)	0.016*
	Clinical sciences	72(90)	72(90)	65(81.3)	63(78.8)	68(85)	74(92.5)	0.072
2017	**Basic sciences**	39(97.5)	37(92.5)	36(90)	36(90)	38(95)	40(100)	0.300
	Clinical sciences	60(83.3)	67(93.1)	55(76.4)	54(75)	59(81.9)	67(93.1)	0.006**
2018	**Basic sciences**	37(97.4)	34(89.5)	31(81.6)	29(76.3)	36(94.7)	38(100)	0.003**
	Clinical sciences	53(75.7)	57(81.4)	49(70)	49(70)	50(71.4)	57(81.4)	0.364
2019	**Basic sciences**	39(100)	39(100)	37(94.9)	34(87.2)	37(94.9)	39(100)	0.019*
	Clinical sciences	57(81.4)	64(91.4)	57(81.4)	48(68.6)	60(85.7)	63(90)	0.005**
2020	**Basic sciences**	40(100)	40(100)	38(95)	32(80)	38(95)	40(100)	0.000**
	Clinical sciences	63(79.7)	72(91.1)	67(84.8)	60(75.9)	65(82.3)	70(88.6)	0.102
2021	**Basic sciences**	38(97.4)	38(97.4)	33(84.6)	31(79.5)	38(97.4)	39(100)	0.001**
	Clinical sciences	64(88.9)	61(84.7)	52(72.2)	51(70.8)	55(76.4)	62(86.1)	0.021*
TOTAL	**Basic sciences**	500(97.1)	489(95)	469(91.1)	442(85.8)	490(95.1)	508(98.6)	0.000**
	Clinical sciences	812(82.2)	844(85.4)	756(76.5)	702(71.1)	808(81.8)	878(88.9)	0.000**

Pearson Chi Square. *p-value is significant at p<0.05 level; **p-value is significant at p<0.01; GPT: Generative Pre-trained Transformer; LLM: large language model; n: number of questions; %: percentage.

Performance of LLMs by Branch

The minimum and maximum correct response rates of LLMs in basic science branches were as follows: GPT-4o (96.1%-100%), Co-pilot (90.4%-100%), GPT-4 (92.3%-98.1%), Gemini 2.0 (93.2%-98.1%), GPT-3.5 (87.2%-96.2%), and Gemini 1.5 (73.1%-94.2%) (Table 3). GPT-4o provided 100% correct responses in 4 out of 8 branches, while Co-pilot provided 100% correct responses in 3. None of the LLMs provided 100% correct responses in the Anatomy and Medical Pharmacology questions. The lowest correct response rates in LLMs in basic sciences were as follows, respectively: Co-pilot (Medical Pathology; 90.4%), GPT-3.5 (Medical Biochemistry; 87.2%) and GPT-4 (Medical Biochemistry; 92.3%) Gemini 1.5 (Medical Pathology; 73.1%), Gemini 2.0 (Anatomy; 93.2%) and GPT-4o (Medical Physiology; 96.1%) (Figure 1).

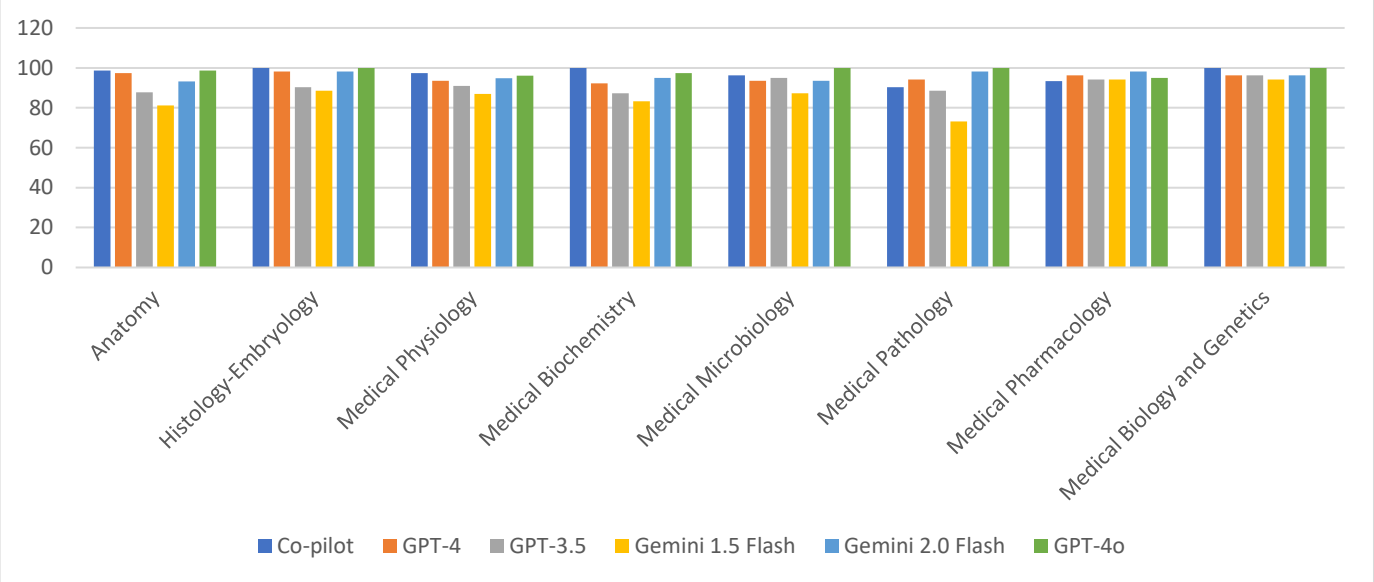

Figure 1. Comparison of correct answers given by LLMs according to basic science subjects (%).

The highest correct response rate in basic sciences was GPT-4o (98.6%), followed by Co-pilot (97.1%). The lowest correct response rate was seen in Gemini 1.5 (85.8%) (Figure 3). There was a significant difference between the correct response rates of LLMs in all basic sciences branches and in all basic sciences questions except Medical Physiology, Medical Pharmacology, Medical Biology and Genetics (Table 3).

Table 3. Comparison of correct answers of LLMs in basic sciences by subject.

	Correct Response Rate n(%)						
SUBJECTS	**Co-pilot n(%)**	**GPT-4 n(%)**	**GPT-3.5 n(%)**	**Gemini 1.5 n(%)**	**Gemini 2.0 n(%)**	**GPT-4o n(%)**	**p value**
Anatomy	73(98.6%)	72(97.3%)	65(87.8%)	60(81.1%)	69(93.2%)	73(98.6%)	0.000**
Histology-Embryology	52(100%)	51(98.1%)	47(90.4%)	46(88.5%)	51(98.1%)	52(100%)	0.005**
Medical Physiology	75(97.4%)	72(93.5%)	70(90.9%)	67(87%)	73(94.8%)	74(96.1%)	0.114
Medical Biochemistry	78(100%)	72(92.3%)	68(87.2%)	65(83.3%)	74(94.9%)	76(97.4%)	0.000**
Medical Microbiology	75(96.2%)	73(93.6%)	74(94.9%)	68(87.2%)	73(93.6%)	78(100%)	0.027*
Medical Pathology	47(90.4%)	49(94.2%)	46(88.5%)	38(73.1%)	51(98.1%)	52(100%)	0.000**
Medical Pharmacology	48(92.3%)	50(96.2%)	49(94.2%)	49(94.2%)	49(94.2%)	51(98.1%)	0.834
Medical Biology and Genetics	52(100%)	50(96.2%)	50(96.2%)	49(94.2%)	50(96.2%)	52(100%)	0.398
TOTAL	500(97.1%)	489(95%)	469(91.1%)	442(85.8%)	490(95.1%)	508(98.6%)	0.000**

Pearson Chi Square. *p-value is significant at p<0.05 level; **p-value is significant at p<0.01; GPT: Generative Pre-trained Transformer; LLM: large language model; n: number of questions; %: percentage.

The lowest correct response rates of LLMs in clinical sciences branches were as follows: GPT-4o (Prosthetic Dentistry; 81%), GPT-4 (Endodontics; 75.4%), Gemini 2.0 (Orthodontics; 72.9%), Co-pilot (Orthodontics; 71.2%), GPT-3.5 (Endodontics; 65.6%) and, Gemini 1.5 (Orthodontics; 55.9%). The highest correct response rates of LLMs in clinical sciences branches were as follows: GPT-4o (Restorative dentistry; 95.3%), GPT-4 (Periodontology; 91.3%), Gemini 2.0 (Oral and maxillofacial surgery; 92.7%), Co-pilot (Restorative dentistry; 90.6%), GPT-3.5 (Oral and maxillofacial surgery; 85.4%, Gemini 1.5 (Oral and maxillofacial surgery; 80.5%). No LLM in clinical sciences branches reached 100% correct response rate (Table 4)(Figure 2).

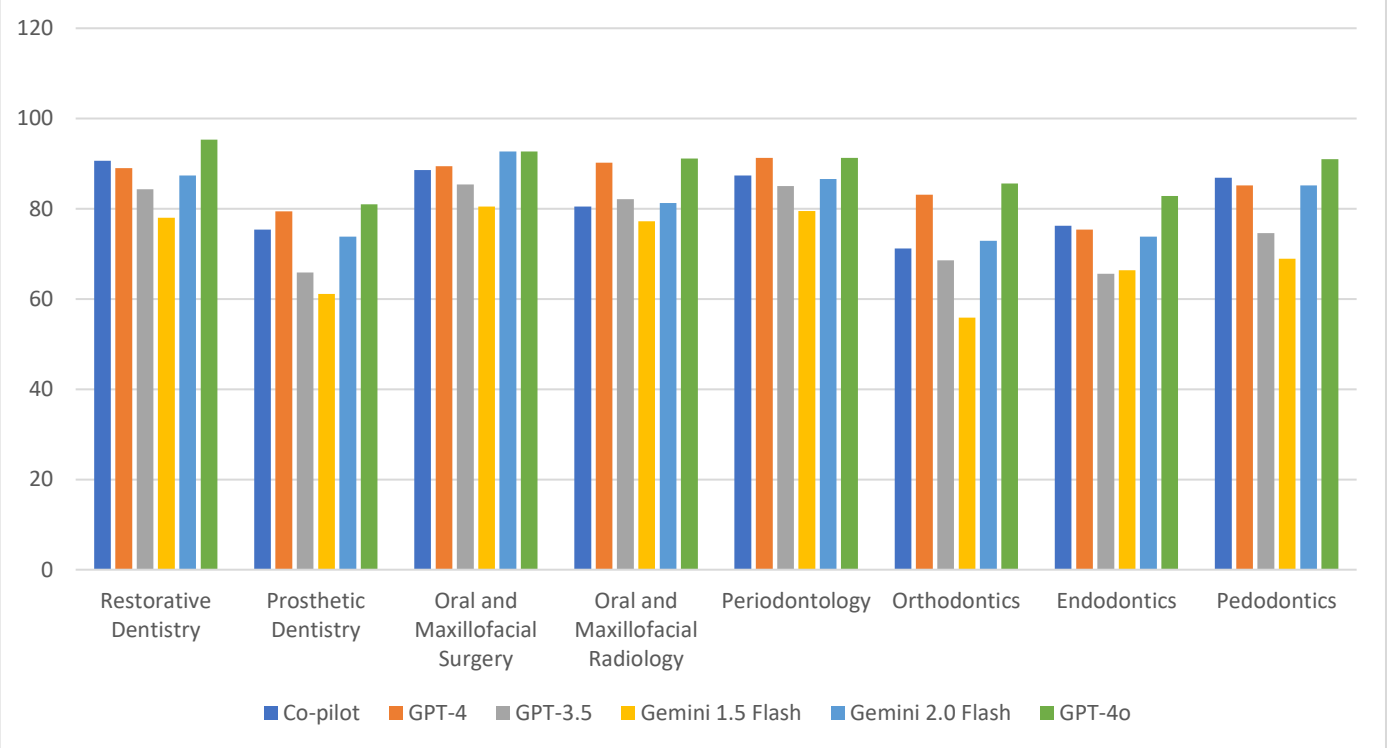

Figure 2. Comparison of correct answers given by LLMs according to clinical science subjects (%).

The highest correct response rate in clinical sciences branches was GPT-4o (88.9%), followed by GPT-4 (85.4%) and the lowest correct response rate was Gemini 1.5 (71.1%) (Figure 3). There were significant differences between the correct answer rates of LLMs in clinical sciences except periodontics ($p<0.05$) (Table 4).

Table 4. Comparison of correct answers of LLMs in clinical sciences by subject.

	Correct Response Rate n(%)						
SUBJECTS	**Co-pilot n(%)**	**GPT-4 n(%)**	**GPT-3.5 n(%)**	**Gemini 1.5 n(%)**	**Gemini 2.0 n(%)**	**GPT-4o n(%)**	**p value**
Restorative Dentistry	115(90.6%)	113(89%)	107(84.3%)	99(78%)	111(87.4%)	121(95.3%)	0.001**
Prosthetic Dentistry	95(75.4%)	100(79.4%)	83(65.9%)	77(61.1%)	93(73.8%)	102(81%)	0.002**
Oral and Maxillofacial Surgery	109(88.6%)	110(89.4%)	105(85.4%)	99(80.5%)	114(92.7%)	114(92.7%)	0.024*
Oral and Maxillofacial Radiology	99(80.5%)	111(90.2%)	101(82.1%)	95(77.2%)	100(81.3%)	112(91.1%)	0.014*
Periodontology	111(87.4%)	116(91.3%)	108(85%)	101(79.5%)	110(86.6%)	116(91.3%)	0.054
Orthodontics	84(71.2%)	98(83.1%)	81(68.6%)	66(55.9%)	86(72.9%)	101(85.6%)	0.000**
Endodontics	93(76.2%)	92(75.4%)	80(65.6%)	81(66.4%)	90(73.8%)	101(82.8%)	0.022*
Pedodontics	106(86.9%)	104(85.2%)	91(74.6%)	84(68.9%)	104(85.2%)	111(91%)	0.000**
TOTAL	812(82.2%)	844(85.4%)	756(76.5%)	702(71.1%)	808(81.8%)	878(88.9%)	0.000**

Pearson Chi Square. *p-value is significant at $p<0.05$ level; **p-value is significant at $p<0.01$; GPT: Generative Pre-trained Transformer; LLM: large language model; n: number of questions; %: percentage.

When the correct answer rates of LLMs to all questions were examined, the performance ranking was as follows: GPT-4o (92.2%), GPT-4 (88.7%), Co-pilot (87.3%), Gemini 2.0 (86.4%), GPT-3.5 (81.5%) and Gemini 1.5 (76.1%) (Table 5). However, as a result of the analysis, a statistically significant difference was observed between the correct answer rates of six LLMs ($p<0.05$) (Table 5). GPT-4o (98.6%) gave the highest correct answer rate in basic sciences, followed by Co-pilot (97.1%). Similarly, in clinical sciences, GPT-4o (88.9%) gave the highest correct response rate, followed by GPT-4 (85.4%). Gemini 1.5 gave the lowest correct response rate in both basic (85.8%) and clinical sciences (71.1%) (Figure 3). The results of pairwise comparisons in terms of LLMs' responses to all questions were given in detail (Table 6).

Table 5. Comparison of the correct answer rates of six LLMs to all TDSE questions.

	Correct Response Rate n(%)						
	Co-pilot n(%)	**GPT-4 n(%)**	**GPT-3.5 n(%)**	**Gemini 1.5 n(%)**	**Gemini 2.0 n(%)**	**GPT-4o n(%)**	**p value**
Basic sciences	500(97.1%)	489(95%)	469(91.1%)	442(85.8%)	490(95.1%)	508(98.6%)	0.000**
Clinical sciences	812(82.2%)	844(85.4%)	756(76.5%)	702(71.1%)	808(81.8%)	878(88.9%)	0.000**
TOTAL	1312 (87.3%)	1333 (88.7%)	1225 (81.5%)	1144 (76.1%)	1298 (86.4%)	1386 (92.2%)	0.000**

Pearson Chi Square. *p-value is significant at $p<0.05$ level; **p-value is significant at $p<0.01$; GPT: Generative Pre-trained Transformer; LLM: large language model; TDSE: Turkish Dentistry Specialization Education Entrance Exam; n: number of questions; %: percentage.

Table 6. Pairwise comparisons of LLMs' for all questions.

	Co-pilot	**GPT-4**	**GPT-3.5**	**Gemini 1.5**	**Gemini 2.0**	**GPT-4o**
Co-pilot	-	0.416	<0.001**	<0.001**	0.242	<0.001**
GPT-4	0.416	-	<0.001**	<0.001**	0.130	0.004**
GPT-3.5	<0.001**	<0.001**	-	<0.001**	<0.001**	<0.001**
Gemini 1.5	<0.001**	<0.001**	<0.001**	-	<0.001**	<0.001**
Gemini 2.0	0.242	0.130	<0.001**	<0.001**	-	<0.001**
GPT-4o	<0.001**	0.004**	<0.001**	<0.001**	<0.001**	-

Pearson Chi Square. *p-value is significant at $p<0.05$ level; **p-value is significant at $p<0.01$; GPT: Generative Pre-trained Transformer; LLM: large language model ($p<0.0041$, bonferroni correction).

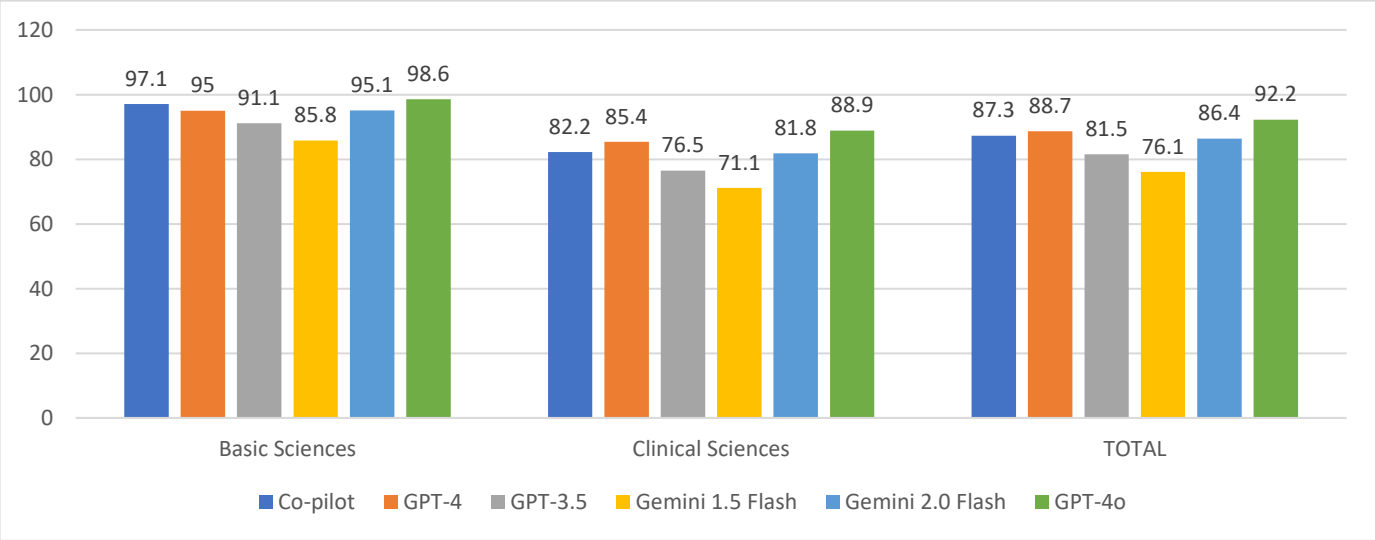

Figure 3. Comparison of percentages of correct answers given by LLMs.

DISCUSSION

In this study, the correct response rates of GPT-3.5, GPT-4, GPT-4o, Gemini 1.5, Gemini 2.0, and Co-pilot were evaluated in answering all questions (n=1503) that did not contain figures and graphs asked in TDSE between 2012-2021. GPT-4o (92.2%) gave the highest correct response rate, followed by GPT-4 (88.7%) and Co-pilot (87.3%). The lowest correct response rate was determined in Gemini 1.5 (76.1%). However, when the correct response rates of the LLMs examined in this study were compared, a statistically significant difference was observed ($p<0.05$).

The performance of leading LLMs in answering questions asked in national dental examinations has been the subject of a number of recent studies. In the 2023 Japanese National Dentist Examination (JNDE), GPT-4 achieved the highest correct answer rate (73.5%) for all questions, followed by Bard (66.5%) and GPT-3.5 (51.9%). The correct response rates of GPT-4 and Bard were observed to be higher than those of GPT-3.5. For dentistry questions, the correct answer rates were 51.6% for GPT-4, 45.3% for Bard, and 35.9% for GPT-3.5, and no statistically significant difference was found between the LLMs (14). In the 2023 Japanese National Dental Hygienist Examination, where 73 questions were evaluated, the highest correct answer rates were seen in GPT-4 (75.3%), followed by Bing (68.5%) and GPT-3.5 (63%). However, this study found no statistically significant difference between the LLMs (25). In this study, the highest success rate was found in GPT-4o (92.2%). GPT-4o was followed by GPT-4 (88.7%), Co-pilot (formerly Bing) (87.3%), Gemini 2.0 (formerly Bard) (86.4%), GPT-3.5 (81.5%), and Gemini 1.5 (formerly Bard) (76.1%). In our study, similar to these studies, after GPT-4, Co-pilot or Gemini 2.0 appear to be the most successful LLMs. In our study, unlike this study, it is noteworthy that there are statistically significant differences between LLMs.

The performance of the GPT has been investigated in many national examinations. In a study conducted in the USA examining the performance of GPT-3.5 and GPT-4 on questions asked in the field of dentistry, 253 questions asked in the Integrated National Board Dental Examination (INBDE), Dental Admission Test (DAT), and Advanced Dental Admission Test (ADAT) exams were analyzed excluding visual questions. While GPT-4 and GPT-3.5 (80%) showed the same success rate in knowledge-based questions asked in INBDE, GPT-4 (69%) showed a higher success rate than GPT-3.5 (66%) in case questions. GPT-4 (83%) showed a higher success rate than GPT-3.5 (66%) in knowledge-based questions asked in ADAT, while GPT-4 and GPT-3.5 (76%) showed the same success rate in case questions. On the other hand, GPT-4 (94%) showed a higher success rate than GPT-3.5 (83%) in knowledge-based questions asked in DAT. As a result of the analysis, it was seen that the correct answer performance of GPT-4 was statistically significant compared to GPT-3.5 (26). Similarly, in our study, GPT-4 showed a higher success rate than GPT-3.5 in both basic sciences (95% vs. 91%), clinical sciences (85.4% vs. 76.5%), and all questions (88.7% vs. 81.5%). In our study, unlike this study, the performance of the higher model GPT-4o was examined and GPT-4o showed significantly better performance than the other GPT models (GPT-3.5 and GPT-4) with a correct answer rate of 98.6% in basic sciences, 88.9% in clinical sciences and 92.2% in all questions.

There have been some recent studies that evaluated the performance of LLMs in answering questions published in the TDSE. Sismanoglu and Capan compared GPT-4 and Gemini Advanced in the TDSE exams published in 2020 and 2021. GPT-4 performed better than Gemini Advance in both the 2020 exam (83.3% vs. 65%) and the 2021 exam (80.5% vs. 60.2%). In addition, GPT-4 showed a higher success rate in both basic sciences and clinical sciences than Gemini Advance in both years. In this study, the success of Gemini 1.5 and Gemini 2 was examined instead of Gemini Advance in 13 TDSEs between 2012 and 2021. The results obtained were similar to the findings of this study and showed that GPT-4 outperformed Gemini models in both clinical and basic sciences in both the 2020 exam and the 2021 exam (23). In the study by Şişmanoğlu and Çapan, both LLMs showed the highest correct response rate in periodontics questions in clinical sciences (GPT-4 95%, Gemini Advanced 90%), while GPT-4 showed the lowest correct response rate in endodontics (45%) and Gemini Advanced in orthodontics (42.1%) (23). Similarly, in our study, GPT-4 showed the best performance in periodontics in clinical sciences (91.3%), while Gemini 1.5 and Gemini 2.0 showed the best performance in oral, dental, and maxillofacial surgery (80.5% and 92.7%, respectively). GPT-4 showed the worst performance in endodontics (75.4%), while Gemini 1.5 and Gemini 2.0 showed the worst performance in orthodontics (55.9% and 72.9%, respectively).

Some recent studies have been conducted to evaluate the questions asked in TDSE on the basis of dental clinical branches. Avsar et al reported that there was no statistically significant difference in performance between Gemini and GPT-3.5 in answering prosthetic dental treatment questions asked in the TDSE (27). In our study, when LLMs were compared in terms of their responses to the field of prosthetic dentistry, their performance rankings were as follows: GPT-4o (81%), GPT-4 (79.4%), Co-pilot (75.4%), Gemini 2.0 (73.8%), GPT-3.5 (65.9%) and Gemini 1.5 (61.1%). Unlike the study by Avsar et al., statistically significant differences were found between LLMs when compared in terms of their responses to the field of prosthetic dentistry. In a study evaluating the performance of LLMs in answering oral, dental and maxillofacial radiology questions in TDSE, the highest correct response rate of 86.1% was obtained with GPT-4o, followed by Bard (61.8%), GPT-3.5 (43.9%) and Co-pilot (41.5%), respectively (22). In our study, similar to this study, GPT-4o showed the highest correct response rate with 91.1%, followed by GPT-4 (90.2%), GPT-3.5 (82.1%), Gemini 2.0 (81.3%), Co-pilot (80.5%), and Gemini 1.5 (77.2). In our study, unlike this study, it was observed that GPT-3.5, Co-pilot, and Gemini models exhibited higher correct response rates in oral and maxillofacial radiology questions. In Tassoker's study, Bard (newly named Gemini) answered the questions with more words, in detail, while GPT-3.5 answered with the fewest words (22). In addition, in this study, GPT-3.5 gave the fastest

responses, while GPT-4o gave the slowest responses. It is thought that the slower response of GPT-4o may be due to the difference in the way it processes messages. Although the number of words in the response and response time were not evaluated in our study, it was observed that GPT-4o showed the best performance in basic and clinical sciences and all questions.

Some studies evaluating the performance of LLMs on dental clinical specialty questions have shown lower success rates. In a study examining the Polish Final Dentistry Examination conducted in Poland, it was reported that GPT-4 had a 64% success rate in oral and maxillofacial surgery questions (28). In this study, when LLMs were compared in terms of correct response rates in oral and maxillofacial surgery questions, they showed accuracy rates ranging from 80.5% to 92.7%. GPT-4o and Gemini 2.0 gave the highest response rate with 92.7%, followed by GPT-4 (89.4%) and Co-pilot (88.6%). The lowest response percentage was seen in GPT-3.5 (85.4%) and Gemini 1.5 (80.5%), but they were found to have higher performance than the study conducted in Poland. In the study by Künzle and Paris, in which endodontics and restorative dentistry student evaluation questions were analyzed, 151 questions were analyzed and GPT-4o obtained the highest correct response rate with 72%, followed by GPT-4 (62%), Gemini 1.0 (44%) and GPT-3.5 (25%), respectively (9). In our study, GPT-4o reached the highest correct response rate in the fields of restorative dentistry (95.3%) and endodontics (82.8%). In restorative dentistry and endodontics questions, GPT-4 showed 89% and 75.4% correct response rates, GPT-3.5 84.3% and 65.6%, Gemini 2 87.7% and 73.8%, and Gemini 1.5 78% and 66.4% correct response rates, respectively. Similarly, in Suarez's study, it was emphasized that the correct answer rate of GPT-4 in the field of endodontics was only 57.3% and that LLMs cannot replace the clinical decision-making processes of dentists in the current situation (29). It is seen that the correct answer percentages of the LLMs examined in our study are higher than those of studies conducted in similar branches in other countries. This situation reveals that the performance of LLMs may vary from country to country or from exam to exam, depending on the content, language, and time of the data sets examined.

This study has some limitations. First of all, since not all LLMs examined have image analysis capabilities, questions containing figures were not included. As image analysis capabilities improve, such questions will need to be re-evaluated. Secondly, questions were asked only once to LLMs. Asking questions more than once may allow for a more reliable analysis of the consistency of the answers. Thirdly, although questions were directed to LLMs simultaneously, the tests were conducted on different days since a total of 1503 questions were asked to 6 models one by one. Factors such as the frequency of updating the models and contextual memory status may have affected the performance. Fourthly, since there is not enough data on the success of the candidates taking the exams in this study, the performances of LLMs were only compared among themselves and no comparison could be made with the candidates taking the exam. Although LLMs are trained to translate across languages and can work independently of language, asking questions in Turkish may have affected the results. Finally, the correct answer rates of LLMs were evaluated in this study, but the logic behind their answers and the scientific adequacy and consistency of the information given were not analyzed. Other studies can be conducted to analyze both the accuracy and scientific basis of the answers provided by LLMs. Despite these limitations, this study is the first to evaluate the correct response rates of leading LLMs for all questions asked within the scope of TDSE. Unlike previous studies, this study used a larger data set and comprehensively examined the performance of leading large LLMs by considering all basic and clinical science branches together.

CONCLUSION

The success of the LLMs evaluated in this study varied between 76.1% and 92.2%, and the examined LLMs performed better than similar studies in the literature. The highest correct response rate was observed in ChatGPT-4o, followed by GPT-4, Co-pilot, Gemini 2.0, GPT-3.5, and the lowest correct response rate was seen in Gemini 1.5. Significant differences were observed between the performances of the LLMs according to the exams conducted over the years and the branches examined. All of the examined LLMs showed higher performance in basic sciences than in clinical sciences. The lower performance of LLMs in clinical sciences indicates that AI may not yet be sufficient to replace human expertise in clinical skills such as combining anamnesis with clinical examination, interpretation, and critical thinking. However, the findings show that LLMs have the potential to be used as a supportive educational tool in basic and clinical dentistry education, despite certain limitations. Further studies are needed to demonstrate the performance of LLMs in dental education and clinical practice.

DECLARATIONS

Ethics Approval And Consent To Participate

Not applicable. Open-source public data was used in this study.

Consent For Publication

Not applicable since there was no direct human contact.

Conflict of Interests

The authors declare that they have no competing interests.

Funding

No financial support was obtained for the completion of this study.

Author Contributions

Study conceptualisation: ÖE. Study protocol and design: ÖE. Data collection: ÖE, İÇ, Data analysis and interpretation: ÖE, İÇ, Writing original manuscript: İÇ Reviewing and editing of manuscript: ÖE, Reading and approval of the manuscript: all authors.

Acknowledgements

None

ORCIDs:

ÖE: 0000-0002-7902-9601

İÇ: 0009-0004-4879-2403

REFERENCES

1. Feigenbaum EA. Some challenges and grand challenges for computational intelligence. Journal of the ACM. 2003;50(1):32-40. doi:10.1145/602382.602400
2. Clusmann J, Wagner SJ, Kornblum HI, et al. The future landscape of large language models in medicine. Commun Med. 2023;3(1):141. doi:10.1038/s43856-023-00370-1
3. Russell SJ, Norvig P. Artificial intelligence: a modern approach. 4th ed. Hoboken, NJ: Pearson; 2021.
4. Sharma M, Savage C, Nair M, et al. Artificial intelligence applications in health care practice: scoping review. J Med Internet Res. 2022;24(10):e40238. doi:10.2196/40238
5. Ding H, Wu J, Zhao W, et al. Artificial intelligence in dentistry: a review. Front Dent Med. 2023;4:1085251. doi:10.3389/fdmed.2023.1085251
6. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8
7. Nori H, King N, McKinney SM, et al. Capabilities of GPT-4 on medical challenge problems. arXiv [Preprint]. Published March 2023. doi:10.48550/arXiv.2303.13375
8. 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. doi:10.2196/52346
9. Künzle P, Paris S. Performance of large language artificial intelligence models on solving restorative dentistry and endodontics student assessments. Clin Oral Investig. 2024;28(11):575. doi:10.1007/s00784-024-05968-w
10. OpenAI. Introducing ChatGPT. Available from: https://openai.com/blog/chatgpt. Accessed May 2025.
11. Microsoft 365 Co-pilot. Available from: https://learn.microsoft.com/tr-tr/copilot/microsoft-365/microsoft-365-copilot-overview Accessed May 2025.
12. Google AI. Gemini: The next-generation AI model. Available from: https://gemini.google/overview/?hl=tr Accessed May 2025.
13. Livberber T, Ayvaz S. The impact of artificial intelligence in academia: views of Turkish academics on ChatGPT. Heliyon. 2023;9(9):e19688. doi:10.1016/j.heliyon.2023.e19688
14. Ohta K, Ohta S. The performance of GPT-3.5, GPT-4, and Bard on the Japanese National Dentist Examination: a comparison study. Cureus. 2023;15(12):e50369. doi:10.7759/cureus.50369
15. Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the National Nurse Examinations in Japan: evaluation study. JMIR Nurs. 2023;6:e47305. doi:10.2196/47305
16. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. J Chin Med Assoc. 2023;86(7):653-658. doi:10.1097/JCMA.0000000000000942
17. 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:e45312. doi:10.2196/45312
18. Yanagita Y, Yokokawa D, Uchida SJ, et al. Accuracy of ChatGPT on medical questions in the National Medical Licensing Examination in Japan: evaluation study. JMIR Form Res. 2023;7:e48023. doi:10.2196/48023
19. Wang X, Zhou H, Zhang Y, et al. ChatGPT performs on the Chinese National Medical Licensing Examination. J Med Syst. 2023;47(1):86. doi:10.1007/s10916-023-01961-0
20. Choi RY, Coyner AS, Kalpathy-Cramer J, et al. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol. 2020;9(2):14. doi:10.1167/tvst.9.2.14
21. Haze T, Kawano R, Takase H, et al. Influence on the accuracy in ChatGPT: differences in the amount of information per medical field. Int J Med Inform. 2023;180:105283. doi:10.1016/j.ijmedinf.2023.105283
22. Tassoker M. ChatGPT-4 Omni's superiority in answering multiple-choice oral radiology questions. BMC Oral Health. 2025;25(1):173. doi:10.1186/s12903-025-05554-w
23. 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(1):214. doi:10.1186/s12909-024-06389-9
24. ÖSYM. Diş Hekimliğinde Uzmanlık Eğitimi Giriş Sınavında Çıkmış Sorular. Available from: https://www.osym.gov.tr/TR,15070/dus-cikmis-sorular.html Accessed May 2025.
25. Yamaguchi S, Nakamura K, Tanaka M, et al. Evaluating the efficacy of leading large language models in the Japanese national dental hygienist examination: a comparative analysis of ChatGPT, Bard, and Bing Chat. J Dent Sci. 2024;19(4):2262-2267. doi:10.1016/j.jds.2024.02.019
26. Dashti M, Lee J, Kim S, et al. Performance of ChatGPT-3.5 and GPT-4 on US dental examinations: the INBDE, ADAT, and DAT. Imaging Sci Dent. 2024;54(3):271. doi:10.5624/isd.20240037
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(4):668-673. doi:10.5336/dentalsci.2024-104610
28. Jaworski A, Kowalska M, Nowak P, et al. GPT-4o vs human candidates: performance analysis in the Polish Final Dentistry Examination. Cureus. 2024;16(9):e68813. doi:10.7759/cureus.68813
29. Suárez A, Díaz-Flores García V, Algar J, et al. Unveiling the ChatGPT phenomenon: evaluating the consistency and accuracy of endodontic question answers. Int Endod J. 2024;57(1):108-113. doi:10.1111/iej.13985