ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA

DOI: 10.5336/dentalsci.2025-111376

Comparison of Performance of Leading Large Language Models in Answering Medical Pathology Questions in Dentistry Specialization Education Entrance Exams: A Cross-Sectional Research

Diş Hekimliği Uzmanlık Eğitimi Giriş Sınavlarındaki Tıbbi Patoloji Sorularının Önde Gelen Büyük Dil Modelleri Tarafından Cevaplanma Performanslarının Karşılaştırılması: Kesitsel Araştırma

Ömer EKİCİ[a], İsmail ÇALIŞKAN[a]

[a]Afyonkarahisar Health Sciences University Faculty of Dentistry, Department of Oral and Maxillofacial Surgery, Afyonkarahisar, Türkiye

ABSTRACT Objective: Artificial intelligence (AI) based large language models (LLMs) have recently become an effective and efficient tool in education and learning. The purpose of this study is to comparatively evaluate the accuracy of the answers given by different AI-supported chatbots to the Medical Pat hology questions asked in the Dentistry Specialization Education Entrance Exams (DSE). **Material and Methods:** A total of 52 pathology questions from 13 exams published on the official website of Student Selection and Placement Center (Öğrenci Seçme ve Yerleştirme Merkezi) were included in the study. Questions were directed simultaneously to LLMs by a single operator. Chi-square analysis was used to compare correct response rates among LLMs in all questions. **Results:** The order of correct answer rates of LLMs to all questions was as follows: ChatGPT-4o (100%), Chat GPT-4 (96.15%), Gemini 2.0 (90.38%) and Claude 3 Sonnet (90.38%), Gemini 1.5 (86.53%), Co-pilot (76.92%). In general, correct answer percentages of LLMs in basic pathology questions were higher than in clinical pathology questions. While no statistically significant difference was observed between correct answers of LLMs to basic pathology questions (p=0.542), a significant difference was observed between correct answers of LLMs in clinical pathology and all questions (p<0.05). **Conclusion:** In this study, the highest accuracy rate was found in GPT-4o and the lowest rate was found in Co-Pilot. The findings show that LLMs have the potential to be used as a supportive tool for students and academics in pathology education.

Keywords: Artificial intelligence;
dentistry specialist training entrance exam;
large language model; medical pathology

ÖZET Amaç: Yapay zekâ (YZ) tabanlı büyük dil modelleri [large language models (LLMs)] son zamanlarda eğitim ve öğrenmede etkili ve verimli bir araç hâline gelmektedir. Bu çalışmanın amacı, farklı YZ destekli sohbet robotlarının Diş Hekimliği Uzmanlık Eğitimi Giriş Sınavlarında (DUS) sorulan Tıbbi Patoloji sorularına verdikleri yanıtların doğruluğunu karşılaştırmalı olarak değerlendirmektir. **Gereç ve Yöntemler:** Öğrenci Seçme ve Yerleştirme Merkezi'nin resmi web sitesinde yayınlanan 13 sınavdan, toplamda 52 patoloji sorusu çalışmaya dâhil edildi. Sorular, LLM'lere tek bir operatör tarafından eş zamanlı olarak yönlendirildi. Tüm sorularda LLM'ler arasındaki doğru yanıt oranlarını karşılaştırmak için ki-kare analizi kullanıldı. **Bulgular:** LLM'lerin tüm sorulara verdikleri doğru cevap oranı sıralaması şu şekilde idi: ChatGPT-4o (%100), ChatGPT-4 (%96.15), Gemini 2.0 (%90.38) ve Claude 3 Sonnet (%90.38), Gemini 1.5 (%86.53), Co-pilot (%76.92). Genel olarak LLM'lerin temel patoloji sorularındaki doğru yanıt yüzdeleri klinik patoloji sorularına göre daha yüksek idi. LLM'lerin temel patoloji sorularına verdikleri doğru yanıtlar arasında istatistiksel açıdan fark gözlenmez iken (p=0.542), klinik patoloji ve tüm sorularda LLM'lerin doğru yanıtları arasında anlamlı farklılık görüldü (p<0.05). **Sonuç:** Bu çalışmada, en yüksek doğruluk oranı GPT-4o'te, en düşük oran Co-Pilot'ta tespit edilmiştir. Bulgular, LLM'lerin patoloji eğitiminde öğrenciler ve akademisyenler için destekleyici bir araç olarak kullanılma potansiyeline sahip olduğunu göstermektedir.

Anahtar Kelimeler: Yapay zekâ;
diş hekimliğimde uzmanlık eğitimi giriş sınavı;
büyük dil modeli; tıbbi patoloji

Correspondence: Ömer EKİCİ
Afyonkarahisar Health Sciences University Faculty of Dentistry, Department of Oral and Maxillofacial Surgery, Afyonkarahisar, Türkiye
E-mail: dromerekici@hotmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Dental Sciences.

Received: 18 Apr 2025 ***Accepted:*** 12 Jun 2025 ***Available online:*** 02 Jul 2025

The rapid development in the field of artificial intelligence (AI) in recent years, especially with the evolution of large language models (LLMs), has led to significant changes in many sectors.[1] These advances have had an impact on a wide range of areas, especially healthcare and language processing tasks.[2-5] The use of AI-supported LLMs (LLMAs) allows for a large number of tasks to be performed more practically and quickly in many languages.[6] This aspect has made LLMAs an effective and efficient tool in education and learning processes.[7] LLMAs have gained increasing attention with the release of the Generative Pre-trained Transformer (ChatGPT) by OpenAI in November 2022. With its updated versions such as ChatGPT-4 and ChatGPT-4o, it has become more advanced over time and has been the subject of many studies.[8] In this process, different LLMs such as Gemini (formerly Bard) by Google, Claude by Anthropic, and Microsoft 365 Co-pilot (formerly Bing) have been developed and the competition in the AI field has increased.[9-11] Recently, the use of LLMs as educational support tools for students in basic and clinical dentistry education has been increasing. Various studies have evaluated the effectiveness of LLMs in solving text-based questions in national exams organized for doctors, nurses, and pharmacists.[12,13] It has been demonstrated that LLMs have successful performance and have the capacity to understand complex medical information, especially in national medical exams administered in the USA, Japan, and China.[14-16] There is no study in the literature investigating the effectiveness of LLMs in answering Medical Pathology questions asked in the Dentistry Specialization Education Entrance Exam (DSE) held in Türkiye. The aim of this study is to evaluate the accuracy of the responses of different leading AI-powered chatbots to Medical Pathology questions asked in DSE. In this context, the accuracy rates of the ChatGPT-4, ChatGPT-4o (OpenAI, San Francisco, California, USA), Co-Pilot (Microsoft, Redmond, Washington), Gemini 1.5, Gemini 2.0 (Google LLC, Mountain View, California, USA) and Claude 3 Sonnet (San Francisco, California, USA) models will be compared to the Medical Pathology questions asked in DSE between 2012-2021 and the potential of these LLMs to be used as educational support tools in dentistry education will be revealed.

MATERIAL AND METHODS

ETHICS

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

LLMS

Six LLMs were evaluated in this study: Co-pilot, Claude 3 Sonnet, Gemini 1.5 and Gemini 2.0, GPT-4 and GPT-4o.

Dentistry Specialist Education Entrance Examination

DSE is administered by the Student Selection and Placement Center (SSPC) once a year between 2015-2022 and twice a year (spring and fall semesters) from 2012-2014 and 2023. DSE was conducted 17 times between 2012-2024.[17] Exams consist of a total of 120 multiple choice questions, 40 from basic sciences and 80 from clinical sciences. Each exam asks 4 questions in the field of Medical Pathology.

Inclusion and Exclusion Criteria

13 DSEs were conducted between 2012-2021 and published on the SSPC official website and 52 questions asked in these exams were included in this study. However, since SSPC has not published DSE questions since 2022; exam questions from 2022, 2023 and 2024 were not included in the study. Since there were no Medical Pathology questions that were canceled by SSPC and included figures, all Medical Pathology questions were analyzed.

Subject headings and subheadings were examined under 10 headings for Pathology by examining the sources shown as reference books by SSPC. The distribution of questions asked in Medical Pathology branches by year and the number of questions by each subject subheading were analyzed.[18-21] The correct answers given by LLMs to Medical Pathology questions were examined by year and subject. In addition, LLMs were compared in terms of correct answer rates.

Question Directing and Evaluation Method to LLMs

In the study, 52 multiple choice questions were directed to Co-pilot, ChatGPT-4, ChatGPT-4o, Gemini 1.5, Gemini 2.0 and Claude 3 Sonnet models simultaneously by a single operator on the same day. The questions were presented in Turkish and were entered manually one by one into the chat interfaces of LLMs to receive answers. Before each question was asked, the following instructions were given in Turkish:

> *"I want you to respond to the Pathology questions asked in the Dentistry Specialization Entrance Exam as a participant. I will soon send you the Pathology questions, multiple choice and some with premises, one by one, which include Pathology topics. I want you to give the most probable answers to these questions."*

This answer was considered "correct" if it matched the official answers provided by the SSPC.

DATA ANALYSIS

All statistical analyses were performed with the SPSS statistical program, version 27 (SPSS Inc., Chicago, IL, USA). Standard descriptive statistics were used for statistical analysis. In the descriptive statistics of continuous data, numbers and percentages were given in nominal data. Chi-square analysis was used to compare the correct response rates between LLMs in all questions. p<0.05 was considered statistically significant.

RESULTS

Medical pathology questions asked in DSE were divided into 2 as basic pathology and clinical pathology. When clinical pathology questions were categorized according to their topics, it was seen that the most questions came from the topic of infections (25.92%), while the least questions came from the topics of premalignant lesions, pigmented lesions and fibro osseous lesions with an equal rate of 3.7% (Figure 1). The study also examined the distribution of medical pathology questions by year (Table 1).

The minimum and maximum accuracy rates of the 6 LLMs by year were as follows: Co-pilot (25-100%), ChatGPT-4 (75-100%), ChatGPT-4o (100-100%), Gemini 1.5 (50-100%), Gemini 2.0 (75-100%) and Claude 3 Sonnet (75-100%). GPT-4o answered 100% of Pathology questions correctly in all 13 exams, while GPT-4 answered 100% of Medical Pathology questions correctly in 11 exams. Gem-

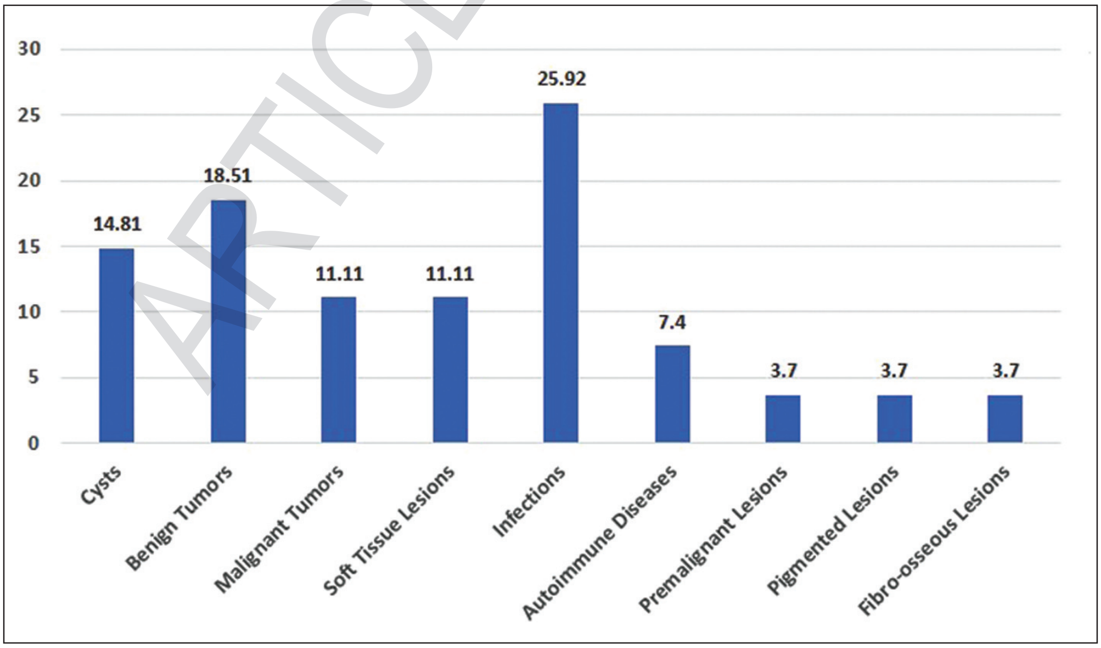

FIGURE 1: Percentage distribution of clinical pathology questions by subject (%)

TABLE 1: Distribution of questions on Medical Pathology in the DSE by year between 2012-2021

Topics	2012/1	2012/2	2013/1	2013/2	2014/1	2014/2	2015	2016	2017	2018	2019	2020	2021	Total
Basic pathology	1	2	1	2	2	2	1	2	2	2	3	3	2	25
Clinical pathology	3	2	3	2	2	2	3	2	2	2	1	1	2	27
Cysts	-	1	-	-	1	1	-	-	-	1	-	-	-	4
Benign tumors	1	-	1	1	-	-	-	1	-	-	-	-	1	5
Malignant tumors	-	-	-	-	-	1	-	-	1	-	-	1	-	3
Soft tissue lesions	1	-	-	-	-	-	1	-	1	-	-	-	-	3
Infections	-	-	2	1	1	-	1	1	-	-	-	-	1	7
Autoimmune diseases	1	-	-	-	-	-	-	-	-	-	1	-	-	2
Premalignant lesions	-	1	-	-	-	-	-	-	-	-	-	-	-	1
Pigmented lesions	-	-	-	-	-	-	-	-	-	1	-	-	-	1
Fibro-osseous lesions	-	-	-	-	-	-	1	-	-	-	-	-	-	1
Total	4	4	4	4	4	4	4	4	4	4	4	4	4	52

DSE: Dentistry Specialization Education Entrance Exam

ini 1.5, Gemini 2.0, and Claude 3 Sonnet answered 8 out of 13 exams 100% correctly, while Co-pilot answered 6 out of 100% correctly (Table 2). The minimum and maximum accuracy rates of LLMs according to Medical Pathology topics were as follows: Co-pilot (0-100%), ChatGPT-4 (75-100%), ChatGPT-4o(100-100%), Gemini 1.5 (25-100%), Gemini 2.0 (25-100%), and Claude 3 Sonnet (50-100%). While GPT-4o answered all 10 topics 100% correctly, GPT-4 answered 8, Gemini 1.5 and Gemini 2.0 answered 7, Claude 3 Sonnet answered 6, and Co-pilot answered 4 (Table 3, Figure 2). When the correct response rates given by LLMs to basic pathology questions were examined, it was seen that other LLMs, except for Co-pilot (96%) and Claude 3 Sonnet (96%), reached 100% correct response rate.

TABLE 2: Comparison of correct answers given by LLMs to Medical Pathology questions asked in the DSE exam by year

		Correct response rate (%)					
		Co-Pilot	GPT-4	GPT-4o	Gemini 1.5	Gemini 2.0	Claude 3 Sonnet
DSE Exams	n	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
2012/1	4	2 (50)	3 (75)	4 (100)	4 (100)	3 (75)	3 (75)
2012/2	4	3 (75)	4 (100)	4 (100)	3 (75)	3 (75)	4 (100)
2013/1	4	3 (75)	4 (100)	4 (100)	4 (100)	4 (100)	3 (75)
2013/2	4	4 (100)	4 (100)	4 (100)	4 (100)	4 (100)	3 (75)
2014/1	4	2 (50)	4 (100)	4 (100)	2 (50)	3 (75)	3 (75)
2014/2	4	4 (100)	4 (100)	4 (100)	3 (75)	4 (100)	4 (100)
2015	4	4 (100)	4 (100)	4 (100)	4 (100)	4 (100)	4 (100)
2016	4	2 (50)	4 (100)	4 (100)	2 (50)	3 (75)	4 (100)
2017	4	4 (100)	4 (100)	4 (100)	4 (100)	4 (100)	4 (100)
2018	4	1 (25)	3 (75)	4 (100)	4 (100)	3 (75)	3 (75)
2019	4	4 (100)	4 (100)	4 (100)	4 (100)	4 (100)	4 (100)
2020	4	4 (100)	4 (100)	4 (100)	4 (100)	4 (100)	4 (100)
2021	4	3 (75)	4 (100)	4 (100)	3 (75)	4 (100)	4 (100)
Total	52	40 (76.92)	50 (96.15)	52 (100)	45 (86.53)	47 (90.38)	47 (90.38)

DSE: Dentistry Specialization Education Entrance Exam; GPT: Generative Pre-trained Transformer

TABLE 3: Comparison of correct answers given by LLMs according to Medical Pathology topics

		Correct response rate (%)					
		Co-Pilot	GPT-4	GPT-4o	Gemini 1.5	Gemini 2	Claude 3 Sonnet
Topics	n	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
Basic pathology	25	24 (96)	25 (100)	25 (100)	25 (100)	25 (100)	24 (96)
Clinical pathology	27	16 (59.3)	25 (92.6)	27 (100)	20 (74.1)	22 (81.5)	23 (85.2)
Cysts	7	4 (57.14)	7 (100)	7 (100)	4 (57.14)	6 (85.71)	6 (85.71)
Benign tumors	4	1 (25)	3 (75)	4 (100)	1 (25)	1 (25)	2 (50)
Malignant tumors	5	2 (40)	4 (80)	5 (100)	4 (80)	4 (80)	4 (80)
Soft tissue lesions	3	3 (100)	3 (100)	3 (100)	3 (100)	3 (100)	3 (100)
Infections	3	3 (100)	3 (100)	3 (100)	3 (100)	3 (100)	3 (100)
Autoimmune diseases	2	1 (50)	2 (100)	2 (100)	2 (100)	2 (100)	2 (100)
Premalignant lesions	1	1 (100)	1 (100)	1 (100)	1 (100)	1 (100)	1 (100)
Pigmented lesions	1	0 (0)	1 (100)	1 (100)	1 (100)	1 (100)	1 (100)
Fibro osseous lesions	1	1 (100)	1 (100)	1 (100)	1 (100)	1 (100)	1 (100)
Total	52	40 (76.92)	50 (96.15)	52 (100)	45 (86.53)	47 (90.38)	47 (90.38)

GPT: Generative Pre-trained Transformer

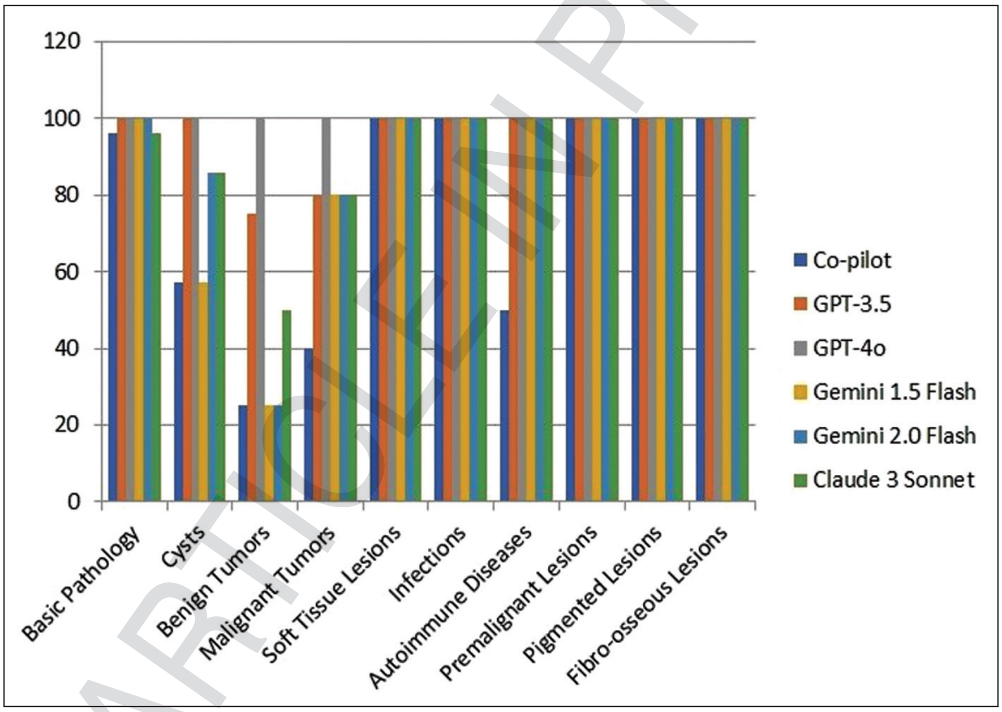

FIGURE 2: Comparison of correct answers given by LLMs by pathology subject (%)
GPT: Generative Pre-trained Transformer

While LLMs showed 96% and above correct response rate to basic pathology questions, the correct response percentage was lower in clinical pathology questions. When the correct response rates of LLMs to clinical pathology questions were examined, the performance was as follows: ChatGPT-4o (100%), ChatGPT-4 (92.59%), Gemini 2.0 (81.48%) and Claude 3 Sonnet (81.48), Co-pilot (77.77%), Gemini 1.5 (74.07%). All LLMs gave 100% correct responses to the questions on malignant tumors, soft tissue lesions, premalignant lesions and fibro osseous lesions. The lowest correct response rates were detected in the topics of odontogenic/non-odontogenic cysts (50%) and benign tumors (76.66%), respectively (Table 3, Figure 2). When the correct response rates of LLMs to all questions were examined, the performance was as follows: ChatGPT-4o (100%), ChatGPT-4 (96.15%), Gemini 2.0 (90.38%) and Claude 3 Sonnet (90.38%), Gemini 1.5 (86.53%), Co-pilot (76.92%). No statistically significant difference was observed between the correct response rates of LLMs to basic pathology questions (p=0.542). However, a statistically significant difference was observed in terms of correct response rates of LLMs in clinical pathology questions and all questions (p=0.002). ChatGPT-4o and ChatGPT-4 exhibited the best performance in clinical pathology questions, while Claude 3 Sonnet and Gemini 2.0 followed. Gemini 1.5 and Co-pilot exhibited significantly lower performance compared to other LLMs. When the performance of LLMs was compared in all questions, Co-pilot exhibited the worst performance, similar to clinical questions (Table 4, Figure 3).

DISCUSSION

In this study, the correct response rates of Claude 3 Sonnet, GPT-4, GPT-4o, Gemini 1.5, Gemini 2.0 and Co-Pilot models were evaluated for a total of 52 questions in 13 DSEs conducted between 2012-2021. According to the analysis results, the highest accuracy rate of 100% was obtained by GPT-4o in all questions, followed by GPT-4 (96.15%), Gemini 2.0 (90.38%) and Claude 3 Sonnet (90.38%), respectively. The lowest correct response rate was observed in the Co-pilot model with 76.92%. Statistically significant differences were observed in terms of correct response rates among the LLMs included in the study ($p<0.05$).

TABLE 4: Comparison of accuracy rates of six LLMs in basic pathology, clinical pathology and all questions

		Correct response rate (%)						
		Co-Pilot	GPT-4	GPT-4o	Gemini 1.5	Gemini	Claude 3 Sonnet	
	n	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	p value
Basic pathology	25	24 (96)	25 (100)	25 (100)	25 (100)	25 (100)	24 (96)	0.542
Clinical pathology	27	16 (59.3)^{d}	25 (92.6)^{a}	27 (100)^{a}	20 (74.1)^{c}	22 (81.5)^{b}	23 (85.2)^{b}	0.002*
Total	52	40 (76.92)^{c}	50 (96.15)^{a}	52 (100)^{a}	45 (86.53)^{b}	47 (90.38)^{b}	47 (90.38)^{b}	0.002*

*p<0.05. Different superscripts in each row indicate statistically significant difference between groups. GPT: Generative Pre-trained Transformer

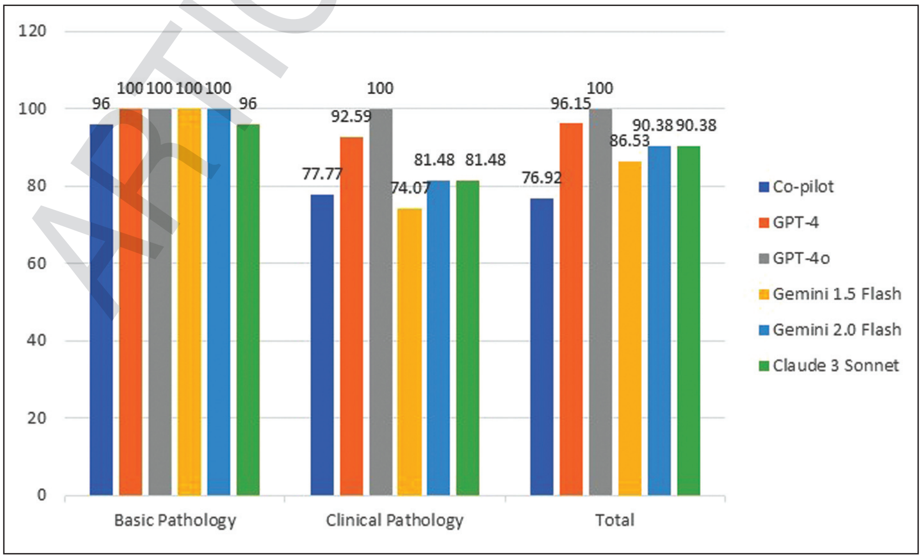

FIGURE 3: Comparison of correct answers of LLMs in Basic Pathology, Clinical Pathology and All questions (%)
GPT: Generative Pre-trained Transformer

In recent years, some studies have been conducted to evaluate the performance of LLM models in answering questions asked in the dentistry specialization exam in Türkiye. Sismanoglu and Çapan compared ChatGPT 4 with Gemini Advanced in the 2020-2021 DSE exams, and ChatGPT-4 outperformed Gemini Advanced in both exams. In this study, ChatGPT-4 achieved a correct response rate of 83.3% in the 2020 exam and 80.5% in the 2021 exam, while Gemini Advanced achieved a correct response rate of 65-60.2% in the exams in the same years, respectively.[22] In a recent DSE exam conducted by Tassoker in Türkiye, the correct response rates of the four major LLMs in answering questions in the field of radiology were compared, and the highest correct response rate of 86.1% was achieved by GPT-4o, followed by Bard (61.8%), GPT-4 (43.9%), and Co-Pilot (41.5%). It has been emphasized that GPT 4o has the potential to serve as a reliable source of information for both healthcare professionals and the general public.[23] Similar results were obtained in this study, with GPT-4o showing the best performance, followed by GPT-4, Claude 3 Sonnet, Gemini 2.0 and Gemini 1.5, respectively. Similarly, Co-pilot showed the lowest correct response rate.

In this study, the correct response rates of LLMs were examined under the titles of basic pathology and clinical pathology. While most of the LLMs answered questions close to 100% correctly in the basic pathology field, only GPT-4o answered 100% correctly in the clinical pathology field. Benirschke et al. evaluated the performance of GPT-4 on 61 pathology questions, and GPT-4 showed an accuracy rate of 98%, similar to the GPT-4o we used in our study. It was also revealed that 82% of the answers given by GPT-4 included all the information. Researchers have stated that GPT-4 can be used to help a pathologist or healthcare provider answer questions.[24] Similarly, Daungsupawong et al. in a study evaluating the performance and usefulness of GPT-4 in the field of pathology, emphasized that GPT-4 showed similar performance in anatomical and clinical pathology fields and that GPT-4 could be useful in the field of pathology.[25]

In a study comparing ChatGPT and Bard (now called Gemini) in pathology questions conducted in Thailand, GPT-4 achieved a 100% correct response rate in questions containing clinical information, while Bard achieved a correct response rate of 87.2%. The researchers stated that GPT-4 was superior to Bard in terms of performance, and that Bard tended to give reasonable but incorrect answers. Therefore, they emphasized that AI should be integrated in medical education in a careful and controlled manner.[26] Our study results, similar to the findings of this study, revealed that both GPT-4o and GPT-4 models provided clear superiority over Gemini 1.5 and Gemini 2.0 models in clinical pathology and all questions. While GPT 4o answered basic pathology, clinical pathology and all questions correctly, GPT-4 answered 100% of basic pathology questions and exhibited a correct response rate of over 90% in clinical pathology and all questions (92.6-96.15%, respectively). On the other hand, Gemini models answered 100% of basic pathology questions correctly, while Gemini 1.5 answered over 70% in clinical pathology and all questions (74.1-86.53%, respectively), and Gemini 2.0 answered over 80% in clinical pathology and all questions (81.5-90.38%, respectively).

In a study conducted by Huang et al. in Taiwan to evaluate the performance of ChatGPT in medical licensing exams, GPT-4 was asked a total of 75 pathology questions, and the correct answer rates for each year were 2022/1 (100%), 2022/2 (80%), and 2023 (84%). In the study, GPT-4 achieved a total correct answer rate of 86.7% in pathology questions. The researchers suggested that ChatGPT could be a useful tool for learning and exam preparation for medical students and specialists. They also stated that ChatGPT has the potential to improve the accessibility of medical education and support continuing education for medical specialists.[27] In this study, GPT-4o and GPT-4 models were asked a total of 52 pathology questions, and GPT-4o correctly answered 100% of the pathology questions in all exams conducted between 2012-2021. ChatGPT-4, on the other hand, showed a 75% correct response rate in 2 exams (2012/1 and 2018), while it showed a 100% correct response rate in all other 11 exams, and achieved an average success rate of 96.15% in all exams.

The evaluation of radiographic and visual data is of great importance in dentistry practices. However, there were no questions containing images in our study. It has been reported in the literature that ChatGPT-4V achieved only a 35% accuracy rate in image-based Japanese National Dentist Examination questions.[28] Interpreting complex data contained in images (such as radiographs and pathology images) requires integrated clinical knowledge and experience. This explains why existing LLMs perform less well in such questions. In the study, it was emphasized that ChatGPT-4V does not yet provide sufficient reliability in image recognition and interpretation, and therefore should be used with caution as an educational or decision support tool in the fields of medicine and dentistry.

Sinha et al. in their study, ChatGPT showed a relational level accuracy rate of over 80% (average 86%) in 100 pathology questions requiring high-level reasoning, and each question was solved in an average of 45.31±7.14 seconds. There was no difference between the scores of the answers given to questions asked from various organ systems in the pathology course. The researchers stated that academics or students can get help from these LLM programs to solve reasoning-type questions.[29] In our study, while no statistically significant difference was observed between the correct answers given by LLMs to basic pathology questions, a significant difference was observed between the correct answers of LLMs to clinical pathology and all questions ($p<0.05$). It is also noteworthy that the percentage of correct answers given by LLMs to clinical pathology questions was generally lower than to basic pathology questions. This shows that AI-based tools need to develop in terms of clinical reasoning, critical thinking, and interpreting multiple pieces of information together. Therefore, it should be kept in mind that these technologies still do not reach the level of in-depth knowledge and interpretation skills that human experts have. The study highlights this difference as a necessity for ongoing research and improvement efforts in the development of AI technologies.

Although most studies in the literature demonstrate the potential of LLMs as educational support tools, LLMs still have some limitations. It is clear that LLMs lag behind human experts, especially in matters such as clinical decision making, analyzing multidimensional scenarios, and ethical reasoning. In addition, the fact that these models sometimes tend to present incorrect or imaginary information as if it were true reveals the need for careful use.[14] LLMs are not considered appropriate to replace diagnosis, treatment planning, or clinical judgment in professional dental practice. In conclusion, these findings indicate that LLMs can make significant contributions, especially in exam-based medical education, and emphasize that these technologies should be used as a complementary tool to traditional education and as a tool to support human expertise.

This study has some limitations and the findings should be interpreted within this framework. Questions containing figures and images were not asked to LLMs. This means that the performance of these models against questions based on image analysis cannot be measured. The exclusion of visual content from the evaluation has shown itself with limited accuracy rates even in models with visual processing capabilities such as ChatGPT-4V.[2] Another limitation is that each question was asked to LLMs only once and the responses were analyzed based on this single evaluation. However, it is known that LLMs can give different answers to the same question at different times.[14,23] Therefore, asking the same questions more than once can provide more reliable results in terms of the consistency of the responses and the stability of the model. In this study, a total of 52 questions were asked to 6 different models and 306 question-answer transactions were performed. The number of samples can be expanded for more consistent results. Finally, only the accuracy rates of the responses given by LLMs were evaluated; the content, scientific adequacy and logical consistency of the explanations were excluded from the scope of the analysis. However, assessing whether AI models produce reliable information in a clinical context requires examining not only the answer accuracy but also the scientific validity of the explanatory rationales.[14] Despite all these limitations, this study is one of the first to comparatively evaluate the correct response rates of leading LLMs using pathology questions in the DSE. It is of great importance that future studies focus on evaluating the performance of LLMs more comprehensively, especially in terms of multilingual content, visual analysis capacity, and clinical reasoning levels.

CONCLUSION

The LLMs evaluated in this study performed quite successfully on the pathology questions asked in the DSE exam in Türkiye. ChatGPT-4o showed excellent performance by answering all questions 100% correctly. GPT-4o was followed by GPT-4 (96.15%), Gemini 2.0 (90.38%), Claude 3 Sonnet (90.38%), and Gemini 1.5 (86.53%). The lowest accuracy rate was observed in the Co-pilot model with 76.92%. The study findings revealed that there were statistically significant performance differences among LLMs. This suggests that models may exhibit different levels of success depending on the level of knowledge, language processing ability, and content of training data. Despite all the limitations, this study demonstrates that LLMs have the potential to be used as a complementary digital learning tool, especially in dentistry pathology education. Further research is needed to increase the reliability of LLMs, evaluate their performance in different languages and specialties, and make the use of models more systematic in clinical contexts.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

***Idea/Concept:** Ömer Ekici; **Design:** Ömer Ekici; **Control/Supervision:** Ömer Ekici; **Data Collection and/or Processing:** Ömer Ekici, İsmail Çalışkan; **Analysis and/or Interpretation:** Ömer Ekici, İsmail Çalışkan; **Literature Review:** Ömer Ekici, İsmail Çalışkan; **Writing the Article:** Ömer Ekici, İsmail Çalışkan; **Critical Review:** Ömer Ekici; **References and Fundings:** Ömer Ekici, İsmail Çalışkan; **Materials:** Ömer Ekici, İsmail Çalışkan.*

REFERENCES

1. Wu T, He S, Liu J, Sun S, Liu K, Han QL. A brief overview of ChatGPT: the history, status quo and potential future development. IEEE/CAA Journal of Automatica Sinica. 2023;10(5):1122-36. doi:10.1109/JAS.2023.123618
2. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. 2023. https://arxiv.org/pdf/2303.13375
3. Wang S, Zhao Z, Ouyang X, Liu T, Wang Q, Shen D. ChatCAD: interactive computer-aided diagnosis on medical image using large language models. 2023;3(133). https://doi.org/10.1038/s44172-024-00271-8
4. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med (Lond). 2023;3(1):141. PMID: 37816837; PMCID: PMC10564921.
5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-40. PMID: 37460753.
6. Yüzbaşıoğlu E. Attitudes and perceptions of dental students towards artificial intelligence. J Dent Educ. 2021;85(1):60-8PMID: 32851649.
7. Livberber T and AS. The impact of Artificial Intelligence in academia: views of Turkish academics on ChatGPT. Heliyon. 2023;9(9):e19688. https://www.sciencedirect.com/science/article/pii/S2405844023068962
8. OpenAI [Internet]. Introducing ChatGPT. OpenAI © 2015–2025 [Cited: February 2025]. Available from: https://openai.com/blog/chatgpt
9. Anthropic AI. The Claude 3 model family: Opus, Sonnet, Haiku. Claude-3 Model Card. 2024;1. https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf
10. Google AI [Internet]. Gemini: The next-generation AI model. ©2025 Google [Cited: February 2025]. Available from: https://gemini.google.com/faq?hl=tr
11. Microsoft 365 [Internet]. Microsoft 365 Copilot'a genel bakış. [Cited: February 2025]. Available from: https://learn.microsoft.com/tr-tr/copilot/microsoft-365/microsoft-365-copilot-overview
12. 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
13. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. J Chin Med Assoc. 2023;86(7):653-8. PMID: 37227901.
14. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, 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. Erratum in: JMIR Med Educ. 2024;10:e57594. PMID: 36753318; PMCID: PMC9947764.
15. 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. PMID: 38213361; PMCID: PMC10782219.
16. Wu J, Wu X, Qiu Z, Li M, Lin S, Zhang Y, et al. Large language models leverage external knowledge to extend clinical insight beyond language boundaries. J Am Med Inform Assoc. 2024;31(9):2054-64. PMID: 38684792; PMCID: PMC11339525.

17. ÖSYM [İnternet]. DUS Çıkmış Sorular. © T.C. ÖLÇME, SEÇME VE YERLEŞTİRME MERKEZİ BAŞKANLIĞI [Erişim tarihi:]. Erişim linki: https://www.osym.gov.tr/TR,15070/dus-cikmis-sorular.html.

18. ÖSYM. Diş Hekimliğinde Uzmanlık Eğitimi Giriş Sınavının Hazırlık Aşamasında Kullanılan Kitaplar. https://www.osym.gov.tr/Eklenti/2568,dus-kaynaklar-07062013pdf.pdf?0

19. Folly P. Cawson's Essentials of Oral Pathology and Oral Medicine. 10th ed. BDJ Student; 2024.

20. Neville B, Damm DD, Allen CM, Chi AC. Oral and Maxillofacial Pathology. 5th ed. E-Book. Elsevier Health Sciences; 2023.

21. Regezi JA, Sciubba J, Jordan RC. Oral Pathology: Clinical Pathologic Correlations. Elsevier Health Sciences; 2016.

22. 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. PMID: 39930399; PMCID: PMC11809121.

23. Tassoker M. ChatGPT-4 Omni's superiority in answering multiple-choice oral radiology questions. BMC Oral Health. 2025;25(173). doi:10.1186/s12903-025-05554-w

24. Benirschke RC, Wodskow J, Prasai K, Freeman A, Lee JM, Groth J. Assessment of a large language model's utility in helping pathology professionals answer general knowledge pathology questions. Am J Clin Pathol. 2024;161(1):42-8. PMID: 37658808.

25. Daungsupawong H, Wiwanitkit V. Large language model's utility in helping pathology professionals. Am J Clin Pathol. 2024;161(2):210. PMID: 37843400.

26. Apornvirat S, Namboonlue C, Laohawetwanit T. Comparative analysis of ChatGPT and Bard in answering pathology examination questions requiring image interpretation. Am J Clin Pathol. 2024;162(3):252-60. PMID: 38619043.

27. Huang CH, Hsiao HJ, Yeh PC, Wu KC, Kao CH. Performance of ChatGPT on Stage 1 of the Taiwanese medical licensing exam. Digit Health. 2024;10:20552076241233144. PMID: 38371244; PMCID: PMC10874144.

28. Morishita M, Fukuda H, Muraoka K, Nakamura T, Hayashi M, Yoshioka I, et al. Evaluating GPT-4V's performance in the Japanese national dental examination: a challenge explored. J Dent Sci. 2024;19(3):1595-600. PMID: 39035269; PMCID: PMC11259620.

29. Sinha RK, Deb Roy A, Kumar N, Mondal H. Applicability of ChatGPT in assisting to solve higher order problems in pathology. Cureus. 2023;15(2):e35237. PMID: 36968864; PMCID: PMC10033699.