ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA

DOI: 10.5336/dentalsci.2025-110603

Performance of ChatGPT 4, ChatGPT 3.5, Gemini 1.5, and Copilot in Solving Oral and Maxillofacial Surgery Questions Asked in the Turkish Dentistry Specialization Education Entrance Exam: Comparison Study

Türk Diş Hekimliği Uzmanlık Eğitimi Giriş Sınavı'nda Sorulan Ağız, Diş ve Çene Cerrahisi Sorularını Çözmede ChatGPT 4, ChatGPT 3.5, Gemini 1.5 ve Copilot'un Performansı: Karşılaştırma Çalışması

Ömer EKİCİ[a]

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

ABSTRACT Objective: The study aims to analyze and compare the performance of 4 leading large language models (LLMs) in answering questions related to oral and maxillofacial surgery, as posed in the Turkish Dentistry Specialization Education Entrance Exam. **Material and Methods:** A total of 123 oral and maxillofacial surgery questions, without figures or graphs, published between 2012-2021, were analyzed. The study evaluated the performance of ChatGPT 4, ChatGPT 3.5, Gemini 1.5, and Copilot. The correct answer rates of LLMs were compared according to the years in which the questions were asked and oral and maxillofacial surgery topics. **Results:** In the study, the highest correct response rate was obtained with ChatGPT-4 (91.06%), followed by Copilot (86.99%), ChatGPT 3.5 (82.11%), and Gemini 1.5 (79.67%). However, no statistically significant difference was observed regarding correct response rates among the 4 LLMs examined in the study (p=0.059). All LLMs correctly answered 66.66% of orofacial infection questions, 80% of orthognathic surgery questions, and 100% of orofacial pain questions. ChatGPT 4 and Copilot answered 100% of dental implantology questions correctly. **Conclusion:** The LLMs examined in the study exhibited acceptable correct response rates (79.67% to 91.06%), and their performances were similar to each other. The results of the study demonstrate the possible of LLMs to be used as educational support instruments in oral and maxillofacial surgery education.

Keywords: Large language model; oral and maxillofacial surgery; GPT 4; GPT 3.5; Copilot; Gemini 1.5

ÖZET Amaç: Çalışmanın amacı, Türkiye Diş Hekimliği Uzmanlık Eğitimi Giriş Sınavı'nda sorulan oral ve maksillofasiyal cerrahi ile ilgili soruları cevaplamada önde gelen 4 büyük dil modelinin (BDM) performansını analiz etmek ve karşılaştırmaktır. **Gereç ve Yöntemler:** Yayımlanmış, şekil ve grafik içermeyen toplam 123 oral ve maksillofasiyal cerrahi sorusu, 2012-2021 yılları arasında analiz edildi. Çalışmada; ChatGPT 4, ChatGPT 3.5, Gemini 1.5 ve Copilot'un performansı değerlendirildi. BDM'lerin doğru yanıt oranları soruların sorulduğu yıllara ve oral ve maksillofasiyal cerrahi konularına göre karşılaştırıldı. **Bulgular:** Çalışmada, en yüksek doğru yanıt oranı ChatGPT 4 (%91,06) ile elde edildi, bunu Copilot (%86,99), ChatGPT 3.5 (%82,11) ve Gemini 1.5 (%79,67) izledi. Ancak çalışmada incelenen 4 BDM arasında doğru yanıt oranları açısından istatistiksel olarak anlamlı bir fark gözlenmedi (p=0,059). Tüm BDM'ler orofasiyal enfeksiyon sorularının %66,66'sını, ortognatik cerrahi sorularının %80'ini ve orofasiyal ağrı sorularının %100'ünü doğru yanıtladı. ChatGPT 4 ve Copilot dental implantoloji sorularının %100'ünü doğru yanıtladı. **Sonuç:** Çalışmada incelenen BDM'ler kabul edilebilir doğru yanıt oranları gösterdi (%79,67'ye %91,06) ve performansları birbirine benzerdi. Çalışmanın sonuçları, BDM'lerin oral ve maksillofasiyal cerrahi eğitiminde eğitim destek araçları olarak kullanılma potansiyelini göstermektedir.

Anahtar Kelimeler: Büyük dil modeli; oral ve maksillofasiyal cerrahi; GPT 4; GPT 3.5; Copilot; Gemini 1.5

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: 21 Mar 2025 ***Received in revised form:*** 01 Jul 2025 ***Accepted:*** 07 Jul 2025 ***Available online:*** 03 Sep 2025

Artificial intelligence (AI) has advanced significantly in recent years, especially in the area of large language models (LLMs). These models generate human-like words by analyzing massive textual data and are progressively used in several areas, including question-answering and language translation.[1] LLMs, which are used in many areas of medicine and have many potential applications in dentistry, such as clinical decision-making, dental telemedicine, dental education, and patient education.[2]

Natural language processing (NLP), which allows computers to interpret and process human language, aims to improve the efficiency and usefulness of human-computer interaction. LLM-based AI implementations, which apply NLP approaches to create replies to text-based inputs, have received considerable interest following the introduction of OpenAI's ChatGPT in November 2022.[3] Among the most extensively studied LLMs is ChatGPT, which OpenAI released in November 2022 and has different versions, such as GPT 3.5 and GPT 4. Then, Microsoft Copilot (Formerly Bing Chat) was launched in February 2023 and Google Gemini (Formerly Bard) in March 2023.[4] The ability of LLMs to analyze and interpret large amounts of data has made them important in developing answers to complex questions.[5]

While AI and the increasing availability of LLMs are providing patients with a wealth of information in the fields of medicine and dentistry, the data presented must be trustworthy and accurate. LLMs' correct response rates have been tested in national examinations for dentists, pharmacists, and nurses, and the data suggest that LLMs can be used as a teaching aid.[5] It has been emphasized that LLMs can pass the United States Medical Licensing Examination and other national medical examinations in Japan and China.[5-7] ChatGPT 3.5 and ChatGPT 4 have been reported to perform successfully on the U.S. Integrated National Dental Board Examination, with 80 percent correct answers, especially on knowledge-based questions. However, according to one study, GPT 3.5 failed to pass the Iranian Endodontics Board Certification exam.[8] ChatGPT 3.5 and Gemini 1.5 were found to have similar performance in answering prosthetic dentistry questions asked in the Dentistry Specialization Education Entrance Exam in Türkiye.[9] There is no study in the literature that focuses on investigating the performance of LLMs in answering questions in the field of oral and maxillofacial surgery. This study aims to asses performance of four leading LLMs [ChatGPT 4 (OpenAI, San Francisco, California, USA), ChatGPT 3.5 (OpenAI, San Francisco, California, USA), Gemini 1.5 (Google LLC, Mountain View, California, USA) and Copilot (Microsoft, Redmond, Washington, USA)] in answering oral and maxillofacial surgery questions asked in the Turkish Dentistry Specialist Education Entrance Exam (TDSE).

MATERIAL AND METHODS

ETHICAL CONSIDERATIONS

This study was exempt from ethics committee approval because it did not use human volunteers and merely used publicly accessible internet data.

Large Language Models

Four LLMs were evaluated in this study: ChatGPT 4, ChatGPT 3.5, Gemini 1.5 and Copilot.

Turkish Dentistry Specialization Education Entrance Exam

The TDSE is a centralized test given by the Student Selection and Placement Center [Ölçme, Seçme ve Yerleştirme Merkezi (ÖSYM)] for dentists seeking specialist training. The TDSE, which was first administered in the spring of 2012, was administered twice a year between 2012-2014 and once a year between 2015-2022. It is applied twice a year again starting from 2023. The TDSE was administered 18 times between 2012-2025. There are 120 multiple-choice questions on the TDSE, 40 of which are in the basic sciences and 80 of which are in the clinical sciences. Each year, 10 questions related to oral and maxillofacial surgery are asked in the clinical sciences exam.[10]

INCLUSION AND EXCLUSION CRITERIA

The study included 130 oral, dental and maxillofacial surgery questions asked in 13 TDSE examinations administered between 2012-2021 and published on the ÖSYM website. Exams conducted after 2021

were not included in the study because they were not published. Questions that were canceled by ÖSYM and contained figures were excluded. Two questions were not included in the study because they were cancelled by ÖSYM. Five questions were excluded because they contained figures. As a result, the final analysis included 123 questions.

Prompt Engineering

The original TDSE questions were manually put into each LLM's chat window one at a time in Turkish. Before each question, the following instructions were entered in Turkish:

"You are asked to answer the oral and maxillofacial surgery questions asked in the TDSE exam. There are multiple-choice questions on the test, and you can only choose the best response. Which of the following best represents the most appropriate answer?"

This answer was considered "correct" if it corresponded with the ÖSYM's official response. This test was carried out on November 5, 2024, between 9 am-5 pm.

DATA ANALYSIS

The SPSS statistical software, version 27 (SPSS Inc., Chicago, IL, USA), was used for all data analyses. Standard descriptive statistics were used for statistical analysis. Chi-square analysis was used to compare correct response rates among LLMs. $p<0.05$ values were considered significant.

RESULTS

A total of 123 oral and maxillofacial surgery questions from the TDSE exam were evaluated. When the questions were examined according to the distribution of topics, the most commonly asked topic was "oral diseases (26.82%)" while the least frequently asked topic was "orofacial pain (1.62%)" (Figure 1).

The minimum and maximum accuracy rates of the four LLMs according to years were as follows: ChatGPT 4 (70-100%), ChatGPT 3.5 (55.55-100%), Gemini 1.5 (60-100%), and Copilot (60-100%). In total, ChatGPT 4 answered 5 of the 13 exams, Copilot answered 4, and ChatGPT 3.5 and Gemini 1.5 answered 2 with 100% accuracy. It was seen that the four LLMs answered 100% of the oral and maxillofacial surgery questions asked in 2015 correctly (Table 1).

The minimum and maximum accuracy rates of the four LLMs related to oral and maxillofacial surgery topics were as follows: ChatGPT 4 (66.66-

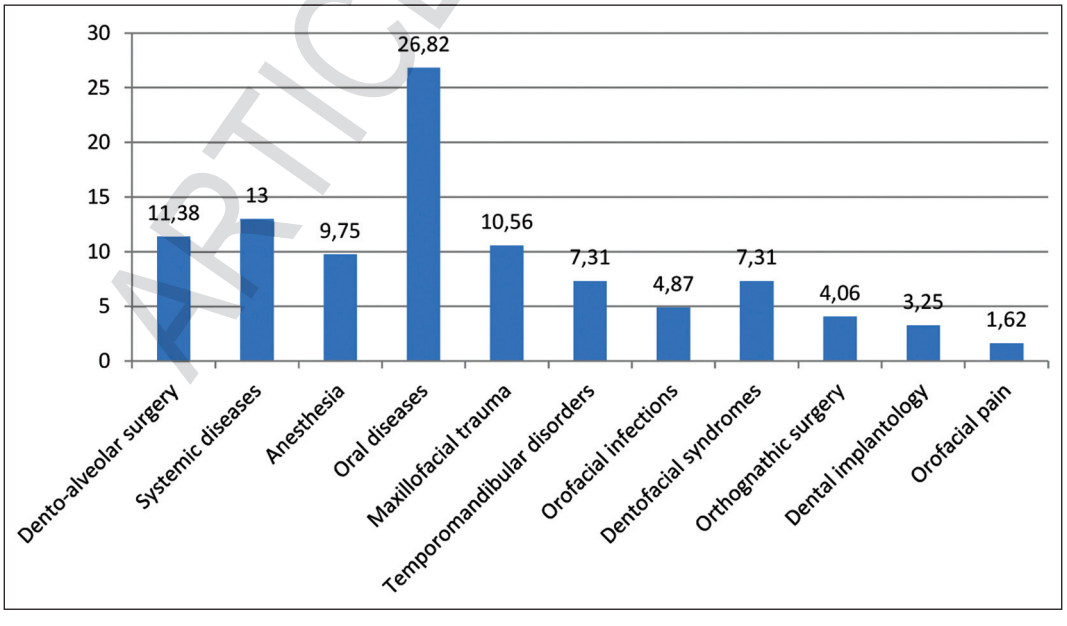

FIGURE 1: Distribution of oral and maxillofacial surgery questions asked in TDSE according to their subjects (%)
TDSE: Turkish Dentistry Specialization Education Entrance Exam

TABLE 1: Comparison of correct answers given by LLMs to oral and maxillofacial surgery questions asked in the TDSE exam by year

		Correct response rate			
		ChatGPT 4	ChatGPT 3.5	Gemini 1.5	Copilot
TDSE	n	n (%)	n (%)	n (%)	n (%)
2012-I	10	7 (70)	7 (70)	6 (60)	6 (60)
2012-II	10	10 (100)	9 (90)	8 (80)	9 (90)
2013-I	10	9 (90)	8 (80)	7 (70)	9 (90)
2013-II	9	9 (100)	9 (100)	8 (88.88)	8 (88.88)
2014-I	10	10 (100)	8 (80)	9 (90)	9 (90)
2014-II	10	9 (90)	8 (80)	7 (70)	7 (70)
2015	8	8 (100)	8 (100)	8 (100)	8 (100)
2016	10	9 (90)	8 (80)	9 (90)	10 (100)
2017	8	8 (100)	7 (87.50)	8 (100)	8 (100)
2018	9	7 (77.77)	5 (55.55)	7 (77.77)	6 (66.66)
2019	9	8 (88.88)	6 (66.66)	6 (66.66)	9 (100)
2020	10	9 (90)	9 (90)	8 (80)	9 (90)
2021	10	9 (90)	9 (90)	7 (70)	9 (90)
Total	123	112 (91.06)	101 (82.11)	98 (79.67)	107 (86.99)

LLMs: Large language models; TDSE: Turkish Dentistry Specialization Education Entrance Exam

TABLE 2: Comparison of correct answers given by LLMs according to oral and maxillofacial surgery topics

		Correct response rate			
		ChatGPT 4	ChatGPT 3.5	Gemini 1.5	Copilot
	n	n (%)	n (%)	n (%)	n (%)
Dento-alveolar surgery	14	13 (92.85)	12 (85.71)	11 (78.57)	13 (92.85)
Systemic diseases	16	14 (87.5)	13 (81.25)	14 (87.5)	11 (68.75)
Anesthesia	12	10 (83.33)	9 (75)	9 (75)	11 (91.66)
Oral diseases	33	32 (96.96)	31 (93.93)	30 (90.9)	32 (96.96)
Maxillofacial trauma	13	11 (84.61)	9 (69.23)	10 (76.92)	11 (84.61)
Temporomandibular disorders	9	9 (100)	7 (77.77)	5 (55.55)	7 (77.77)
Orofacial infections	6	4 (66.66)	4 (66.66)	4 (66.66)	4 (66.66)
Dentofacial syndromes	9	9 (100)	8 (88.88)	9 (100)	8 (88.88)
Orthognathic surgery	5	4 (80)	4 (80)	4 (80)	4 (80)
Dental implantology	4	4 (100)	2 (50)	0 (0)	4 (100)
Orofacial pain	2	2 (100)	2 (100)	2 (100)	2 (100)
Total	123	112 (91.06)	101 (82.11)	98 (79.67)	107 (86.99)

LLMs: Large language models

100%), ChatGPT 3.5 (50-100%), Gemini 1.5 (0-100%), and Copilot (66.66-100%). The 4 LLM models answered 66.66% of the orofacial infections questions correctly, 80% of the orthognathic surgery questions and 100% of the orofacial pain questions correctly. ChatGPT 4 and Copilot also answered 100% of the dental implantology questions correctly. In contrast, Gemini 1.5 could not answer any of the dental implantology questions, while ChatGPT 3.5 could only answer half of them (Table 2, Figure 2).

When the correct answer rates of the four LLM models were examined for all questions, the performance ranking was as follows: ChatGPT 4 (91.06%), Copilot (86.99%), ChatGPT 3.5 (82.11%) and Gemini 1.5 (79.67%) (Figure 3). However, as a result of the chi-square analysis, no statistically significant difference was observed between the correct answer rates of the four LLM models (p=0.059) (Table 3).

DISCUSSION

In this study, we assessed the correct response rates of ChatGPT 4, ChatGPT 3.5, Gemini 1.5, and Copilot in answering oral and maxillofacial surgery questions that did not contain figures and graphics asked in the TDSE exam between 2012-2021. The highest correct response rate in the study was obtained with ChatGPT 4 (91.06%), followed by Copilot (86.99%), ChatGPT 3.5 (82.11%) and Gemini 1.5 (79.67%). But there was no discernible statistically significant change between the performances of the 4 LLM models examined.

The performance of LLMs has been evaluated in many national dental examinations to date.

ChatGPT 4 performed significantly better than ChatGPT 3.5 on questions generated by INBDE Bootcamp, and ITD (76.9% vs. 61.3%, respectively).[11] Similarly, another study found that ChatGPT 4 exceeded ChatGPT 3.5 on the INBDE, ADAT, and DAT exams.[12] In the Korean Dental Hygienist Exam (KDHE) questions asked in the last 5 years, GPT 4 achieved an average success rate of 65.9% (63.6-70.3%), leaving Gemini (49.4-58.2%) and GPT 3.5 (39.2-45.5%) behind.[13] In the 2023 Japanese National Dental Hygienist Examination, where 73 questions were analyzed, the highest correct answer rates were reported to be seen in GPT 4 (75.3%), followed by Bing (68.5%) and GPT 3.5 (63.0%).[14] This finding is very similar to the results of this study, and here was no discernible statistically significant change between LLMs in this study. In the 2023 Japanese National Dentist Examination (JNDE) exam, where 185 questions were used, ChatGPT 4

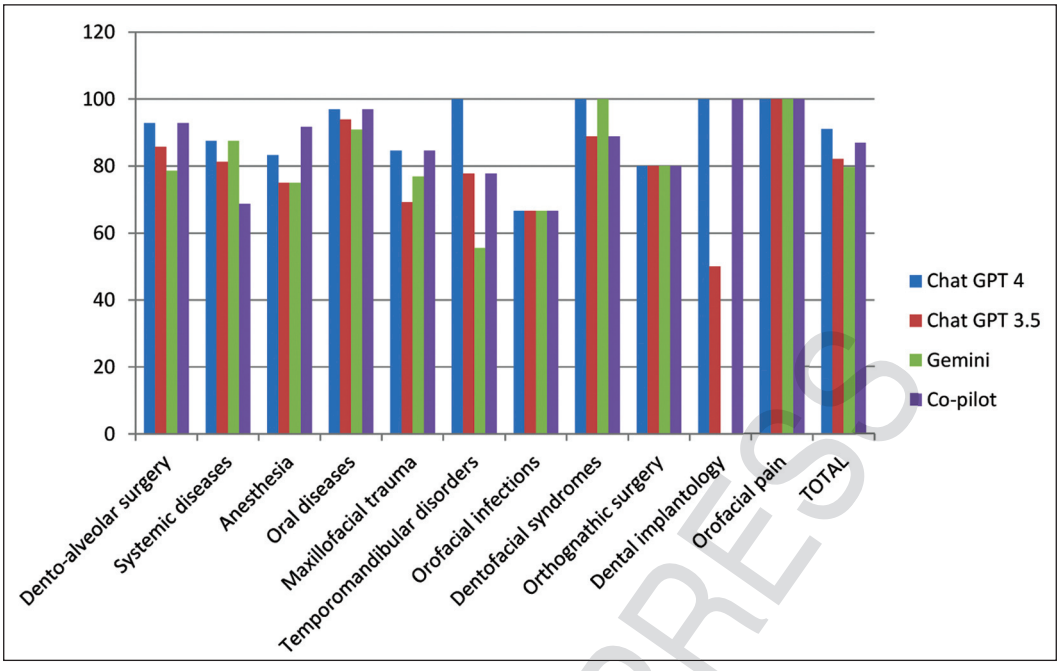

FIGURE 2: Comparison of correct answers of LLMs according to oral and maxillofacial surgery topics (%)
LLMs: Large language models

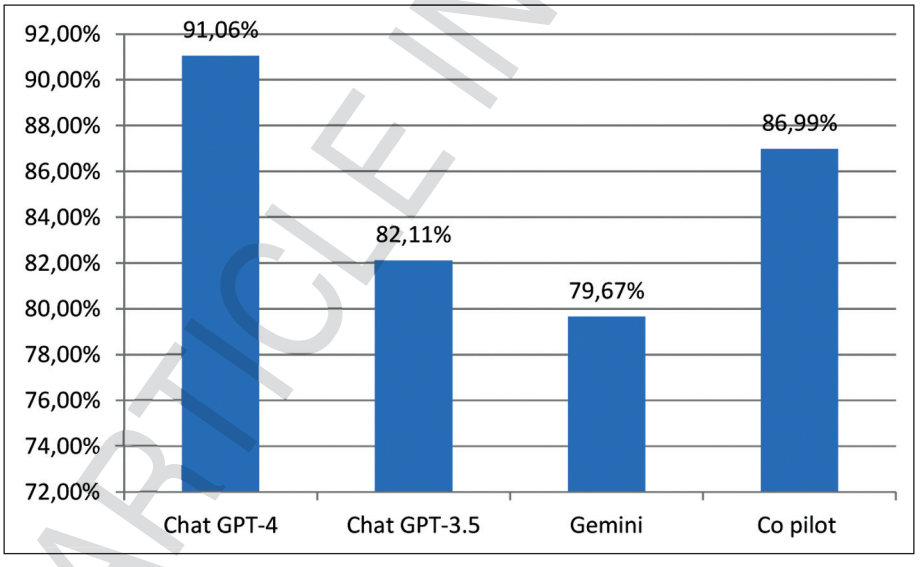

FIGURE 3: Comparison of correct response rates of LLMs to all questions (%)
LLMs: Large language models

TABLE 3: Comparison of LLMs' accuracy rates in all oral and maxillofacial surgery questions

		Correct response rate				
	n	ChatGPT 4	ChatGPT 3.5	Gemini 1.5	Copilot	p value
All questions	123	112 (91.06)	101 (82.11)	98 (79.67)	107 (86.99)	0.059

LLMs: Large language models

was reported to have the highest score (73.5%) for all questions, followed by Bard (66.5%) and ChatGPT 3.5 (51.9%). In professional dentistry questions, correct response rates were 51.6% for GPT 4, 45.3% for Bard and 35.9% for GPT 3.5, and there were no statistically significant differences between the models

for these questions.[4] Similar to this result in our study, the highest correct answer rate was seen in ChatGPT 4 (91.06%), while Copilot (formerly Bing) (86.99%) came second. Unlike our study, the accuracy rate of ChatGPT 3.5 (82.11%) was higher than Gemini (formerly Bard) (79.67%). In our study, it is seen that the accuracy rates of all LLMs we examined are higher than similar studies in the literature. We think that the most important reason for this may be the rapid change in AI tools and the increase in their performance day by day. Additionally, results may vary depending on the version of the LLMs. Even when the same LLMs are used, the success of LLMs may vary from country to country or from exam to exam, depending on the content and language of the dataset used and the time at which the research was conducted.

In this study, the rates of correct answers of the 4 LLMs we examined in oral and maxillofacial surgery exams varied between 79.67-91.06%. However, in the Polish Final Dentistry Examination (PFDE), GPT 4's success in oral surgery questions was found to be lower than our study with a rate of 64%. In this study, ChatGPT 4 showed better performance in areas such as endodontics and restorative dentistry (71.74%) and prosthodontics (80%), while it exhibited lower correctness in oral surgery (64%), pediatric dentistry (62.07%) and orthodontics (52.63%). In clinically more challenging areas such as pediatric dentistry and oral surgery, the decrease in AI accuracy was interpreted as AI models may have difficulty in questions requiring clinical reasoning. In addition, ChatGPT 4's performance in clinical case-based questions (36.36% accuracy) was found to be lower than its performance in other questions (72.87% accuracy).[15] In a German study evaluating the capabilities of LLMs in restorative dentistry and endodontics, 151 student evaluation questions were used for analysis. ChatGPT 4.0o showed the highest performance (72%), followed by ChatGPT 4.0 (62%), Google Gemini 1.0 (44%), and ChatGPT 3.5 (25%).[16] In a study conducted in Spain, it was reported that the percentage of correct answers for GPT 4 in endodontic questions was only 57.3% and that these models cannot currently replace the clinical decision-making processes of dentists.[17]

The ability to analyze visual results and radiographic pictures is important for maxillofacial surgery applications. Since the LLMs we examined do not have image analysis features, we did not include 5 questions containing figures and pictures. However, in a previous study, ChatGPT- 4V was found to have an overall correct response rate of 35% for image-based JNDE questions. While ChatGPT 4V achieved higher correct response rates in specialties such as dental anesthesia and endodontics, response rates were lower for orthodontics (25%) and oral surgery (38.2%). They stated that because there are so many images available for oral surgery and because answering the questions requires integrating a lot of data, the percentage of accurate answers may be lower. Similarly, it was reported that orthodontics questions also contained a large number of pictures and a polygon table. According to researchers, ChatGPT 4V's picture identification capabilities are not yet trustworthy or extensive enough to be employed as a training tool in the medical and dental disciplines.[18] Healthcare professionals who are thinking about using LLMs in clinical settings should also be mindful of "hallucinations", a condition in which LLMs pass off fictitious or misleading information as fact.

In this study, we were unable to compare the success of LLM models with candidates who took the exam, as there was no data on the response rates of oral and maxillofacial surgery questions asked on the TDSE exam. In a study comparing the performance of GPT 4.0o on the PFDE with the results of human candidates, GPT 4.0o achieved a success rate of 70.85%, but fell short of the highest human scores of 94.97%. Although the success of GPT 4.0o suggests its possible as a complementary educational instrument in dentistry, the researchers noted that its clinical reasoning abilities are limited and do not yet match the critical thinking and clinical judgment exhibited by human candidates.[15] Although GPT 4 achieved the highest average performance compared to Gemini and GPT 3.5 on the KDHE with an average success rate of 65.9%, it contrasted sharply with the consistent passing rates of human candidates above 74%. This result illustrates the difficulties in developing AI systems that can match human under-

standing and reasoning in specialized domains.[19] Even while AI has made encouraging strides, its shortcomings in clinical reasoning highlight the need for more research and development. According to these findings, LLMs are still used in professional exams as a supplemental tool rather than as a substitute for human competence.

In this study, the questions were entered into the LLM interface in "Turkish". In an exam that used both the original (Polish) and English languages, there was no discernible difference in performance across question sets written in various languages since LMAs are highly skilled at translating between languages.[20] In contrast, GPT 4 performed better than Gemini and GPT 3.5 on the KDHE in both English and Korean, but GPT 4 performed better in English. GPT 3.5 and Gemini failed in the Korean version and only reached passing scores in the English version. The higher performance of GPT 4 in English is likely due to the more extensive training data available for processing English.[19] The variability in performance across topics and languages highlights the need for continued improvements in these models.[4] Future research should focus on developing methodologies for evaluating LLMs in multilingual environments and improving their ability to handle specialized content in different languages. More research on LLMs' performance on national dental exams in other languages is required to evaluate their global applicability in clinical dentistry.

There are several limitations to this study. First, LLMs were not evaluated on imaging questions. As image analysis skills improve, a new evaluation including these questions will be required. Second, LLMs were tested only once in this study. Multiple attempts may provide a more accurate assessment of response consistency. Third, the rapid advancement of LLM technology means that responses to TDSE questions can vary, so testing was conducted over a single day to minimize this problem. Fourth, LLM models were evaluated for their correct response rates; the quality and adequacy of the information and comments providing the rationale behind their responses were not assessed. Future research is expected to undertake a full analysis that includes both the correctness of responses and the justification of the responses supplied by the LLMs. Although the sample size of 127 questions may be considered small, it has a similar question bank size covering a variety of topics when considering only one topic/specialty area. Despite these limitations, this study is the first to evaluate the performance of leading LLMs in answering oral and maxillofacial surgery questions asked in the TDSE.

CONCLUSION

In the study, the highest correct response rate in answering the oral and maxillofacial surgery questions asked in the TDSE exam was obtained with ChatGPT 4, followed by Copilot (86.99%), ChatGPT 3.5 (82.11%) and Gemini 1.5 (79.67%). However, the correct response rates of these four LLM models were found to be statistically similar to each other. Despite their limitations, the study's findings indicate that the LLMs under investigation have the possible to be utilized as a tool for students' educational support. Further studies on the performance of LLMs in other languages and question sets are needed to better understand their use globally in oral and maxillofacial surgery.

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

This study is entirely author's own work and no other author contribution.

REFERENCES

1. 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.
2. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent. 2023;35(7):1098-102. PMID: 37017291.
3. 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. PMID: 39373739; PMCID: PMC11458639.
4. 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.
5. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ. 2023;9:e48002. PMID: 37384388; PMCID: PMC10365615.
6. Kunitsu Y. The potential of GPT-4 as a support tool for pharmacists: analytical study using the Japanese National Examination for pharmacists. JMIR Med Educ. 2023;9(1). doi:10.2196/48452
7. 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. PMID: 37368470; PMCID: PMC10337249.
8. Farajollahi M, Modaberi A. Can ChatGPT pass the "Iranian Endodontics Specialist Board" exam? Iran Endod J. 2023;18(3):192. PMID: 37431530; PMCID: PMC10329760.
9. Bilgin Avşar D, Atila A, Diş Ü, et al. Diş Hekimliğinde Uzmanlık Sınavında sorulan protetik diş tedavisi sorularının ChatGPT-3.5 ve Gemini tarafından cevaplanma performanslarının karşılaştırmalı olarak incelenmesi: kesitsel araştırma [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 Klin J Dent Sci. 2024;30(4):668-73. doi:10.5336/DENTALSCI.2024-104610
10. Ekici Ö, Çalışkan İ. Retrospective analysis of oral and maxillofacial surgery questions asked in the dentistry specialization training entrance exam. Turkiye Klinikleri Journal of Dental Sciences. 2024;30(4):564-71. doi:10.5336/DENTALSCI.2024-104564
11. 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):970-4. PMID: 37676187.
12. Dashti M, Ghasemi S, Ghadimi N, Hefzi D, Karimian A, Zare N, et al. Performance of ChatGPT 3.5 and 4 on U.S. dental examinations: the INBDE, ADAT, and DAT. Imaging Sci Dent. 2024;54(3):271-5. PMID: 39371301; PMCID: PMC11450412.
13. Song ES, Lee SP. Comparative analysis of the response accuracies of large language models in the Korean National Dental Hygienist Examination across Korean and English questions. Int J Dent Hyg. 2025;23(2):267-76. PMID: 39415339; PMCID: PMC11982589.
14. Yamaguchi S, Morishita M, Fukuda H, Muraoka K, Nakamura T, Yoshioka I, 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):2262-7. PMID: 39347065; PMCID: PMC11437298.
15. Jaworski A, Jasiński D, Sławińska B, Błecha Z, Jaworski W, Kruplewicz M, et al. GPT-4o vs. human candidates: performance analysis in the polish final dentistry examination. Cureus. 2024;16(9):e68813. PMID: 39371744; PMCID: PMC11456324.
16. 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. PMID: 39373739; PMCID: PMC11458639.
17. Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: evaluating the consistency and accuracy of endodontic question answers. Int Endod J. 2024;57(1):108-13. PMID: 37814369.
18. 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.
19. Song ES, Lee SP. Comparative analysis of the response accuracies of large language models in the Korean National Dental Hygienist Examination across Korean and English questions. Int J Dent Hyg. 2025;23(2):267-76. PMID: 39415339; PMCID: PMC11982589.
20. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in Dermatology. Clin Exp Dermatol. 2024;49(7):686-91. PMID: 37540015.