Selcuk Dental Journal / Udeg 3. Uluslararası Diş Hekimliği Eğitimi Kongresi Özel Sayı (2025): 6-10

Research / Araştırma Makalesi

Comparative Evaluation of Four Large Language Models in Turkish Dentistry Specialization Exam

Türk Diş Hekimliği Uzmanlık Sınavında Dört Büyük Dil Modelinin Karşılaştırmalı Değerlendirilmesi

Ömer EKİCİ[α]

[α]Afyonkarahisar Health Sciences University, Faculty of Dentistry, Department of Oral and Maxillofacial Surgery, Afyonkarahisar, Türkiye
[α]Afyonkarahisar Sağlık Bilimleri Üniversitesi, Diş Hekimliği Fakültesi, Ağız, Diş ve Çene Cerrahisi AD, Afyonkarahisar, Türkiye

ABSTRACT

Background: The aim of the study is to evaluate the performance of four leading Large Language Models (LLMs) in the 2021 Dentistry Specialization Exam (DSE).

Methods: A total of 112 questions were used, including 39 questions in basic sciences and 73 questions in clinical sciences, which did not include the figures and graphs asked in the 2021 DSE. The study evaluated the performance of four LLMs: Claude-3.5 Haiku, GPT-3.5, Co-pilot, and Gemini-1.5.

Results: In basic sciences, Claude-3.5 Haiku and GPT-3.5 answered all questions correctly by 100%, while Gemini-1.5 answered by 94.9% and Co-pilot by 92.3%. In clinical sciences, Claude-3.5 Haiku showed an overall correct answer rate of 89%, Co-pilot 80.9%, GPT-3.5 79.7% and Gemini-1.5 65.7%. For all questions, Claude-3.5 Haiku showed a correct answer rate of 92.85%, GPT-3.5 86.6%, Co-pilot 84.8% and Gemini-1.5 75.9%. While the performance of LLMs in basic sciences was similar (p=0.134), there was a statistically significant difference between the performances of LLMs in clinical sciences and all questions (p=0.007 and p=0.005, respectively).

Conclusion: In all questions and clinical sciences, Claude-3.5 Haiku performed best, Gemini-1.5 performed worst, and GPT-3.5 and Co-pilot performed similarly. The 4 LLM models examined showed a higher success rate in basic sciences than in clinical sciences. The results showed that AI-based LLMs can perform well in knowledge-based questions such as basic sciences but perform poorly in questions that require knowledge as well as clinical reasoning, discussion, and interpretation, such as clinical sciences.

Keywords: Artificial intelligence, Dentistry, Dentistry specialization training, Large language model

ÖZ

Amaç: Çalışmanın amacı 2021 Diş Hekimliği Uzmanlık Eğitimi giriş sınavında (DUS) önde gelen dört Büyük Dil Modeli (LLM)' nin performansını değerlendirmektir.

Gereç ve Yöntemler: 2021 DUS sınavında sorulan şekil ve grafik içermeyen temel bilimlerde 39 soru ve klinik bilimlerde 73 soru olmak üzere 112 soru kullanıldı. Çalışmada Claude 3.5 Haiku, GPT-3.5, Copilot ve Gemini 1.5 olmak üzere dört LLM'nin performansı değerlendirildi.

Bulgular: Temel bilimlerde Claude-3.5 Haiku ve GPT-3.5 tüm soruları %100 doğru cevaplarken, Gemini 1.5 %94,9 ve Copilot %92,3 oranında cevapladı. Klinik bilimlerde toplamda Claude 3.5 Haiku %89, Copilot %80,9, GPT-3.5 %79,7 ve Gemini 1.5 %65,7 doğru cevap oranı sergiledi. Tüm sorularda ise Claude 3.5 Haiku %92,85, GPT-3.5 %86,6, Copilot %84,8 ve Gemini %75,9 doğru cevap oranı gösterdi. Temel bilimlerde LLM'lerin performansı benzer iken (p=0.134), klinik bilimlerde ve tüm sorularda LLM'lerin performansları arasında istatistiksel açıdan anlamlı farklılık görüldü (sırasıyla p=0.007 ve p=0.005).

Sonuç: Tüm sorularda Claude 3.5 Haiku en iyi performansı gösterirken, Gemini 1.5 en kötü performansı gösterdi, GPT 3.5 ve Co pilot'un performansı benzer bulundu. İncelenen 4 LLM modeli temel bilimlerde klinik bilimlere göre daha yüksek bir başarı oranı gösterdi. Sonuçlar, yapay zeka tabanlı LLM'lerin temel bilimler gibi bilgiye dayalı sorularda iyi performans sergileyebileceğini ancak klinik bilimler gibi bilgi ile birlikte klinik muhakeme, tartışma ve yorum gerektiren sorularda daha düşük performans sergilediğini gösterdi.

Anahtar Kelimeler: Yapay zekâ, Diş Hekimliği, Diş Hekimliği uzmanlık eğitimi, Büyük dil modeli.

INTRODUCTION

Artificial intelligence (AI) refers to the ability of computers to mimic human intelligence to perform tasks that typically require human abilities, such as understanding, reasoning, and decision making.[1] Many tasks in everyday life have become easier thanks to the use of AI. AI software has the potential to provide a productive avenue for education and learning.[2] Large language models (LLMs) have a variety of applications in medicine and dentistry, representing a significant advance in the field of AI.[3] Potential uses for LLMs in dentistry include clinical decision support, patient and dental education, scientific writing, and scientific education.[4] In dentistry education, LLMs have been shown to help students create a "personalized learning experience," write dental essays, and help instructors create test questions.[5,6]

It is well known that Chat GPT and other LLMs are rapidly innovating and evolving. Commonly known LLMs include ChatGPT, Gemini, Claude and Co-pilot developed by OpenAI, Google, Anthropic, and Microsoft. These models have advanced conversational capabilities that resemble human-like interactions. This feature offers great potential in educational environments such as virtual assistants, chatbots, and online learning support systems.[7] These LLMs are revolutionizing medical education and practice by processing and producing human language at an unprecedented scale.[8] The capacity of LLMs to analyze and interpret large amounts of data has made them indispensable in finding solutions to complex problems.

LLMs are increasingly popular in dentistry, with potential applications ranging from improving diagnostic accuracy to patient education.[9,10] The increasing availability of artificial intelligence and LLMs offers the advantage of providing comprehensive information to healthcare professionals and patients in the fields of medicine and dentistry.[11] However, since they cover critical issues related to human health, it is extremely important that the information provided is accurate and reliable. It is important to understand the practical use of LLMAs in dentistry for exam preparation and information acquisition, and their potential risks. LLMAs can be used by students for information acquisition, but LLMAs are prone to hallucination, meaning that LLMs can persuasively convey false information, and should therefore be used with caution.[12] Furthermore, the performance of different LLMAs is not fully known, and it remains to be discovered whether one is more suitable than another for professional use. LLMs have shown promising results in medical licensing examinations around the world, demonstrating their potential to understand complex medical information. Studies have highlighted that these models have the ability to pass the United States Medical Licensing Examination (USMLE) and other national medical examinations in Japan and China. The findings suggest that LLMs may be a valuable tool in vocational training.[13] However, ChatGPT-3.5 reportedly failed the Iranian Endodontics Specialist Board.[14] No study has been found in the literature that focuses on investigating the performance of LLMs in answering the questions asked in the Dentistry Specialist Education entrance exam (DSE) in Türkiye. The aim of the study was to evaluate the performance

Gönderilme Tarihi/Received: 11 Nisan, 2025
Kabul Tarihi/Accepted: 30 Haziran, 2025
Yayınlanma Tarihi/Published: 19 Eylül, 2025
Atıf Bilgisi/Cite this article as: Ekici Ö. Comparative Evaluation of Four Large Language Models in Turkish Dentistry Specialization Exam. Selcuk Dent J 2025; Udeg 3. Uluslararası Diş Hekimliği Eğitimi Kongresi Özel Sayı: 6-10 Doi: 10.15311/ selcukdentj.1674113

Sorumlu yazar/Corresponding Author: Ömer EKİCİ
E-mail: dromerekici@hotmail.com
Doi: 10.15311/ selcukdentj.1674113

of four leading LLMs in the 2021 DSE and their potential application in dentistry education and practice.

MATERIAL AND METHODS

Since this study used only publicly available internet data and did not involve human participants, ethics committee approval was not required.

The following four LLMs were evaluated in this study: 1) Claude 3.5 Haiku (San Francisco, California, USA); 2) Chat GPT-3.5 (OpenAI, San Francisco, California, USA); 3) Gemini1.5 (Google LLC, Mountain View, California, USA); 4) Co-pilot (Microsoft, Redmond, Washington)

The latest exam published by the Student Selection and Placement Center (SSPC), the 2021 DSE, was included in the study. Two questions that were canceled by SSPC and six questions that included figures and graphics were not included in the study. Therefore, 112 questions were used, 39 in basic sciences and 73 in clinical sciences.

On January 15, 2025, new accounts were created for each LLM evaluated in this study (Claude 3.5 Haiku, Chat GPT-3.5, Gemini1.5, and Copilot. The models were tested using their latest publicly available versions. Cookies and internet history were deleted before queries were made. All questions were entered into the respective LLMs by a single researcher(ÖE). Before each question, the following instruction was entered in Turkish:

"You are a student taking the DUS exam. The exam consists of multiple choice questions and you must select only the most appropriate answer. Which of the following best represents the most appropriate answer?"

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

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. Chi-square analysis was used to compare correct response rates among LLMs. Pairwise comparisons between LLMs were performed using Post-Hoc Chi-Square analysis with Bonferroni correction. For Pearson Chi-Square tests, $p<0.05$ was considered statistically significant, and for Post-Hoc comparisons, the significance level was set at $p<0.0062$ after Bonferroni correction. $P<0.05$ was considered statistically significant.

RESULTS

112 questions, excluding the canceled questions and questions containing figures, were analyzed. The response accuracy of four LLMs, namely Claude 3.5 Haiku, Chat GPT-3.5, Gemini 1.5 and Co-pilot, was compared to the questions asked in the 2021 DSE exam.

In basic sciences, Claude-3.5 Haiku and GPT-3.5 answered all questions 100% correctly, while Gemini 1.5 answered 94.9% and Copilot answered 92.3%. In basic sciences, Gemini 1.5 answered 80% of the Anatomy questions and 83% of the Medical Biochemistry questions correctly. Co-pilot answered 75% of the histology and embryology and medical genetics questions and 83.3% of the Medical Microbiology questions correctly (**Table 1**).

Table 1. Comparison of correct answers given by LLMs to questions asked in the basic sciences test.

Basic Sciences	n	Cloude 3.5 Haiku n(%)	Chat GPT 3.5 n(%)	Co-pilot n(%)	Gemini 1.5 n(%)
Anatomy	5	5(100)	5(100)	5(100)	4(80)
Histology and embryology	4	4(100)	4(100)	3(75)	4(100)
Physiology	6	6(100)	6(100)	6(100)	6(100)
Medical biochemistry	6	6(100)	6(100)	6(100)	5(83.3)
Medical microbiology	6	6(100)	6(100)	5(83.3)	6(100)
Medical pathology	4	4(100)	4(100)	4(100)	4(100)
Medical pharmacology	4	4(100)	4(100)	4(100)	4(100)
Medical biology and genetics	4	4(100)	4(100)	3(75)	4(100)
TOTAL	39	39(100)	39(100)	36(92.3)	37(94.9)

LLMs: Large language models; n:number

In clinical sciences, Claude 3.5 Haiku showed 89% correct answer rate, Copilot 80.9%, GPT-3.5 79.7% and Gemini 1.5 65.7%. The minimum and maximum success of the 4 LLM models in clinical sciences were as follows: Claude 3.5 (70-100%), Chat GPT 3.5 (50-100%), Copilot (44.4%-100%), Gemini 1.5 (40-88.9%). Claude 3.5 Haiku answered all orthodontics questions, Chat GPT 3.5 Oral and Maxillofacial Radiology and Endodontics questions and Copilot answered all periodontics questions correctly. Claude 3.5 Haiku showed the lowest success in prosthetic dentistry (70%), Chat GPT 3.5 in periodontics (50%), Copilot in pedodontics (44.4%), Gemini 1.5 in prosthetic dentistry (40%) (**Table 2**).

Claude 3.5 Haiku showed 92.85% correct answer rate in all questions, GPT-3.5 86.6%, Copilot 84.8% and Gemini 75.9% (**Figure 1**).

Table 2. Comparison of correct answers given by LLMs to questions asked in the clinical sciences test

Clinical Sciences	n	Cloude 3.5 Haiku n(%)	Chat GPT 3.5 n(%)	Co-pilot n(%)	Gemini 1.5 n(%)
Restorative dentistry	9	8(88.9)	7(77.8)	8(88.9)	8(88.9)
Prosthetic dentistry	10	7(70)	7(70)	7(70)	4(40)
Oral and maxillofacial surgery	10	9(90)	9(90)	9(90)	7(70)
Oral and maxillofacial radiology	9	8(88.9)	9(100)	7(77.8)	6(66.7)
Periodontology	10	9(90)	5(50)	10(100)	7(70)
Orthodontics	9	9(100)	8(88.9)	8(88.9)	6(66.7)
Endodontics	7	5(71.4)	7(100)	6(85.7)	6(85.7)
Pediatric dentistry	9	8(88.9)	6(66.7)	4(44.4)	4(44.4)
TOTAL	73	65(89)	58(79.7)	59(80.9)	48(65.7)

LLMs: Large language models; n:number

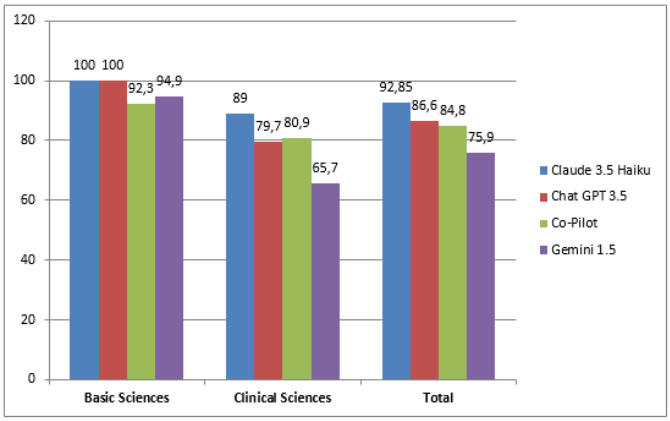

Figure 1. Comparison of correct response rates of LLMs in basic sciences, clinical sciences and all questions (%)

There was no statistically significant difference between the correct answer rates of the four LLMs in basic sciences (p=0.134). Statistically significant difference was observed between LLMs in clinical sciences and all questions (p=0.007 and p=0.005, respectively). Cloude 3.5 Haiku showed the best performance, Gemini 1.5 showed the worst performance. Chat GPT 3.5 and Co pilot showed similar success rates (**Table 3**). According to the pairwise comparison results between LLMs in Clinical Sciences questions and all questions, a statistically significant difference was observed only between Claude 3.5 Haiku and Gemini 1.5 ($p<0.001$). After Benforrri correction, no statistically significant difference was observed between the performances of Claude 3.5 Haiku, Chat GPT3.5 and Co-pilot.(**Table 4** and **Table 5**).

Table 3. Comparison of correct response rates of LLMs in basic sciences, clinical sciences and all questions

	n	Claude 3.5 Haiku n(%)	Chat GPT 3.5 n(%)	Co-pilot n(%)	Gemini 1.5 n(%)	P value
Basic Sciences	39	39(100)	39(100)	36(92.3)	37(94.9)	0.134
Clinical Sciences	73	65(89)	58(79.7)	59(80.9)	48(65.7)	0.007*
TOTAL	112	104 (92.85)	97(86.6)	95(84.8)	85(75.9)	0.005*

*LLMs: Large language models; n:number; *:p<0.05*

Table 4. Pairwise comparisons of LLMs' performance on clinical sciences questions

	Claude 3.5 Haiku	Chat GPT 3.5	Co-pilot	Gemini 1.5
Claude 3.5 Haiku	-	0.086	0.124	<0.001**
Chat GPT 3.5	0.086	-	0.500	0.047
Co-pilot	0.124	0.500	-	0.030
Gemini 1.5	<0.001**	0.047	0.030	-

*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<0.0062, bonferroni correction)*

Table 5. Pairwise comparisons of LLMs' performance on all questions

	Claude 3.5 Haiku	Chat GPT 3.5	Co-pilot	Gemini 1.5
Claude 3.5 Haiku	-	0.093	0.044	<0.001**
Chat GPT 3.5	0.093	-	0.424	0.029
Co-pilot	0.044	0.424	-	0.065
Gemini 1.5	<0.001**	0.029	0.065	-

*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<0.0062, bonferroni correction)*

DISCUSSION

In this study, we evaluated the correct response rates of Claude 3.5 Haiku, Chat GPT 3.5, Gemini 1.5 and Copilot in answering questions that do not contain figures and graphics asked in the 2021 DSE exam. In the study, the highest performance in all questions was obtained with Claude 3.5 Haiku (92.85%), followed by Chat GPT 3.5 (86.6%), Copilot (84.8%) and Gemini 1.5 (75.9%), respectively. A statistical difference emerged between the correct response rates of the LLM models examined in all questions. Cloude 3.5 Haiku showed the best performance, while Gemini showed the worst performance.

The success of LLMs has been researched in various national dental examinations. In the 2023 Japan National Dentistry Examination (JNDE) using 185 questions, Chat GPT 4 performed best on all questions (73.5%), followed by Bard (66.5%) and Chat GPT 3.5 (51.9%). Performance of GPT 4 and Bard was significantly higher than GPT 3.5. LLMs performed worse on Professional Dentistry questions, but the order of performance was similar and no statistically significant difference was observed: GPT-4 (51.6%), Bard (45.3%), GPT-3.5 (35.9%).[15] In the 2023 Japanese national dental hygienist exam, where 73 questions were analyzed, it was reported that the highest correct answer rates were seen in GPT-4 (75.3%), followed by Bing (68.5%) and GPT-3.5 (63.0%), but there was no statistically significant difference between the correct answers of the LLMs.[16] GPT-4 has been successful in the Korean National Dental Hygienist Exam Questions between 63.6% and 70.3% over the last 5 years, outperforming Gemini (49.4% vs. 58.2%) and GPT-3.5 (39.2% vs. 45.5%).[17] In our study, unlike these studies, we used Claude 3.5 Haiku. Claude 3.5 Haiku showed high performance like GPT-4 and the best performance among the LLMs examined. In our study, Copilot (formerly Bing) and Chat GPT 3.5 showed higher performance than Gemini (formerly Bard). In this study conducted in Türkiye, there was a significant difference between the LLMs in clinical sciences and all questions, while there was no statistically significant difference between the performances of the LLMs examined, similar to the studies conducted in Japan in the field of basic sciences. In our study, it is seen that the accuracy rates in all LLMs we examined are higher than similar studies in the literature. It should be taken into account that the results may vary from country to country or from exam to exam depending on the time the research was conducted, the content and language of the data set used.

In our study, the LLM models examined showed lower success in clinical sciences than in basic sciences. Similar to our research results, in a study conducted in Poland, 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).[18] The lower success of LLMs in clinical sciences is due to both the nature of medical knowledge and the limitations of current artificial intelligence models. Basic sciences have more systematic, formulated, rule-based information. However, clinical decisions include a large number of variables (patient medical history, clinical examination findings, laboratory/radiography findings). Differential diagnosis in most diseases requires clinical intuition, experience, and contextual evaluation. LLMs are still limited in understanding the context; they cannot fully mimic real clinical reasoning. In addition, there may be conflicting information in the literature, so situations such as level of evidence, risk/benefit analysis are important in clinical practices, and such evaluations are difficult for LLMs. In addition, clinical guidelines may change frequently and are updated at certain intervals. While content related to basic sciences is available on the internet in a more standard and accessible form, clinical information is more included in medical practice and is based on patient data, data is specific, limited, and restricted for ethical reasons. LLMs can be improved in clinical sciences by increasing their ability to make sense of the patient context, continuous integration of clinical guidelines, training with real case-based educational data, and feedback-based learning methods, but they can never fully replace clinical experience and physician judgment.

The success of LLMs may also vary among different clinical sciences. In our study, Claude 3.5 Haiku answered all orthodontics questions correctly, while the lowest response rate was seen in prosthodontics (70%). Copilot answered all periodontics questions correctly, while Chat GPT 3.5 could only answer half of them correctly. However, the lowest response rates were seen in prosthodontics and pedodontics in Copilot and Gemini 1.5. However, in the Polish Final Dentistry Exam, Chat GPT-4 performed better in areas such as Endodontics and Restorative Dentistry (71.74%) and Prosthodontics (80%), but showed lower accuracy in oral surgery (64%), pediatric dentistry (62.07%) and orthodontics (52.63%). The researchers explained the decrease in AI accuracy in more clinically challenging areas such as pediatric dentistry and oral surgery as AI models may have difficulty with questions that require clinical reasoning.[18] In a study analyzing the performance of LLMAs in solving restorative dentistry and endodontics student assessment questions, 151 questions were used for analysis, of which ChatGPT-4.0o showed the highest success (72%), followed by ChatGPT-4.0 (62%), Google Gemini 1.0 (44%) and ChatGPT-3.5 (25%).[19] Similarly, Suárez reported in his study that the percentage of correct answers for GPT-4 in endodontics questions was only 57.3%, and that these models cannot currently replace dentists in clinical decision-making processes in dentistry.[20]

Since the LLM models we examined do not have image analysis features, we did not include questions containing figures and pictures. However, a study reported that ChatGPT-4V had an overall correct response rate of 35% for image-based Japanese National Dental Examination questions. The researchers noted that ChatGPT-4V's image recognition feature is not yet reliable or comprehensive enough to be used as an educational tool in the medical and dental fields. [21] In addition, since this study did not provide data on the response rates of questions asked in the DUS 2021 exam, we were unable to compare the success of LLM models with candidates who took the exam. In the Polish Final Dentistry Exam, GPT-4o achieved a success rate of 70.85%, but fell short of the highest human success rate of 94.97%.[18] In the Korean National Dental Hygienist Examination, GPT-4 achieved an average success rate of 65.9%, but similarly fell short of the success of human candidates of over 74%.[17] These studies suggest that although LLMs have the potential as a complementary educational tool in dentistry, their clinical reasoning capabilities are limited and do not yet reach the critical thinking and clinical judgment demonstrated by human candidates. Despite the promising progress in AI-based LLMs, their limitations in clinical reasoning indicate the need for continued development and improvement.

There are several limitations to this study. First, LLMs were tested only once in this study. Multiple trials may provide a more accurate assessment of response consistency. Second, LLMs were not tested on imaging questions. As image analysis capabilities improve, a new

assessment that includes these questions will be necessary. Third, rapid advances in LLM technology mean that responses to DUS questions can vary, so testing was conducted on a single day to minimize this problem. Fourth, the study assessed the correct response rates of LLM models, but not the quality and adequacy of the information and interpretations provided in the responses. Despite these limitations, this is the first study to assess the performance of leading LLM models in answering questions asked on DSE.

CONCLUSION

Claude 3.5 Haiku performed best on all questions, while Gemini 1.5 performed worst. The 4 LLM models examined showed a higher success rate in basic sciences than in clinical sciences. The results showed that AI-based LLMs can perform well on knowledge-based questions such as basic sciences, but perform poorly on questions that require interpretation along with knowledge, such as clinical sciences. It should be kept in mind in clinical education and practice that AI has limitations in skills related to clinical experience, such as combining medical history with clinical examination, clinical reasoning, discussing and interpreting ambiguous situations, etc.

Değerlendirme / Peer-Review

İki Dış Hakem / Çift Taraflı Körleme

Etik Beyan / Ethical statement

Bu çalışma 20-22 Şubat 2025 tarihlerinde Bolu Abant İzzet Baysal Üniversitesi'nde düzenlenen UDEG 3. Uluslararası Diş Hekimliği Eğitimi Kongresi'nde "sözlü bildiri" olarak sunulmuştur.

Bu çalışmanın hazırlanma sürecinde bilimsel ve etik ilkelere uyulduğu ve yararlanılan tüm çalışmaların kaynakçada belirtildiği beyan olunur.

This study was presented as an "oral presentation" at the UDEG 3rd International Dentistry Education Congress held at Bolu Abant İzzet Baysal University on 20-22 February 2025.

It is declared that during the preparation process of this study, scientific and ethical principles were followed and all the studies benefited are stated in the bibliography.

Benzerlik Taraması / Similarity scan

Yapıldı – ithenticate

Etik Bildirim / Ethical statement

dishekimligidergisi@selcuk.edu.tr

Çıkar Çatışması / Conflict of interest

Çıkar çatışması beyan edilmemiştir.

Finansman / Grant Support

Yazarlar bu çalışma için finansal destek almadığını beyan etmiştir. | The authors declared that this study has received no financial support.

Yazar Katkıları / Author Contributions

Çalışmanın Tasarlanması | Design of Study: ÖE (%100)
Veri Toplanması | Data Acquisition: ÖE (%100)
Veri Analizi | Data Analysis: ÖE (%100)
Makalenin Yazımı | Writing up: ÖE (%100)
Makale Gönderimi ve Revizyonu | Submission and Revision: ÖE (%100)

REFERENCES

1. Dashti M, Londono J, Ghasemi S, et al. Attitudes, knowledge, and perceptions of dentists and dental students toward artificial intelligence: a systematic review. J Taibah Univ Med Sci. 2024;19(2):327-337. doi:10.1016/j.jtumed.2023.12.010
2. Chakravorty S, Aulakh BK, Shil M, Nepale M, Puthenkandathil R, Syed W. Role of Artificial Intelligence (AI) in Dentistry: A Literature Review. J Pharm Bioallied Sci. 2024;16(Suppl 1):S14-S16. doi:10.4103/JPBS. JPBS_466_23,
3. Sur J, Bose S, Khan F, Dewangan D, Sawriya E, Roul A. Knowledge, attitudes, and perceptions regarding the future of artificial intelligence in oral radiology in India: A survey. Imaging Sci Dent. 2020;50(3):193-198. doi:10.5624/ISD.2020.50.3.193
4. 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):1098-1102. doi:10.1111/JERD.13046
5. Shrivastava PK, Uppal S, Kumar G, Jha P. Role of ChatGPT in Academia: Dental Students' Perspectives. Prim Dent J. 2024;13(1):89-90. doi:10.1177/20501684241230191,
6. Rahad K, Martin K, Amugo I, et al. ChatGPT to Enhance Learning in Dental Education at a Historically Black Medical College. Dent Res oral Heal. 2024;7(1). doi:10.26502/DROH.0069
7. Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ. 2023;103:102274. doi:10.1016/J.LINDIF.2023.102274
8. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-1940. doi:10.1038/S41591-023-02448-8
9. Chau RCW, Thu KM, Yu OY, Hsung RTC, Lo ECM, Lam WYH. Performance of Generative Artificial Intelligence in Dental Licensing Examinations. Int Dent J. 2024;74(3):616-621. doi:10.1016/j.identj.2023.12.007
10. Alhaidry HM, Fatani B, Alrayes JO, Almana AM, Alfhaed NK. ChatGPT in Dentistry: A Comprehensive Review. Cureus. 2023;15(4):e38317. doi:10.7759/CUREUS.38317
11. Huang H, Zheng O, Wang D, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. Int J Oral Sci. 2023;15(1):1-13. doi:10.1038/S41368-023-00239-Y;SUBJMETA=139,1449,3032,692,700;KWRD=DENTISTRY, ELECTRODIAGNOSIS
12. Ji Z, Lee N, Frieske R, et al. Survey of Hallucination in Natural Language Generation. ACM Comput Surv. 2022;55(12). doi:10.1145/3571730
13. 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. doi:10.2196/48002
14. Farajollahi M, Modaberi A. Can ChatGPT pass the "Iranian Endodontics Specialist Board" exam? Iran Endod J. 2023;18(3):192. doi:10.22037/iej.v18i3.42154
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). doi:10.7759/CUREUS.50369
16. Yamaguchi S, Morishita M, Fukuda H, 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
17. 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. Published online 2024. doi:10.1111/IDH.12848
18. Jaworski A, Jasiński D, Sławińska B, et al. GPT-4o vs. Human Candidates: Performance Analysis in the Polish Final Dentistry Examination. Cureus. 2024;16(9). doi:10.7759/CUREUS.68813
19. 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
20. 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):108-113. doi:10.1111/IEJ.13985
21. Morishita M, Fukuda H, Muraoka K, et al. Evaluating GPT-4V's performance in the Japanese national dental examination: A challenge explored. *J Dent Sci*. 2024;19(3):1595-1600. doi:10.1016/J.JDS.2023.12.007