Araştırma Doi: 10.25282/ ted.1716591

Comparison of The Performance of ChatGPT 4.0 and Gemini in Anatomy Questions Asked in Turkey National Medical Specialization Exams

Türkiye Ulusal Tıp Uzmanlık Sınavlarında Sorulan Anatomi Sorularında ChatGPT 4.0 ve Gemini'nin Performansının Karşılaştırılması

Arif Keskin[1]
Orcid: 0000-0002-1634-1091
Tayfun Aygün[2]
Orcid: 0000-0001-5058-3513
[1]Assistant professor, Department of Anatomy, Faculty of Medicine, Giresun University, Giresun, Türkiye
[2]Lecturer Dr., Department of Anatomy, Faculty of Medicine, Giresun University, Giresun, Türkiye

Sorumlu Yazar:
Arif Keskin
E-posta:
arif.keskin@giresun.edu.tr

Keywords:
Anatomy Education, ChatGPT 4.0, Clinical Anatomy, Medical Specialization Examination, Large Language Models, AI in Education
Anahtar Sözcükler:
Anatomi Eğitimi, ChatGPT 4.0, Klinik Anatomi, Tıp Uzmanlık Sınavı, Büyük Dil Modelleri, Eğitimde Yapay Zeka

Gönderilme Tarihi / Submitted:
10.06.2025
Kabul Tarihi / Accepted:
03.11.2025

Künye:
Keskin A, Aygun T. Comparison of The Performance of ChatGPT 4.0 and Gemini in Anatomy Questions Asked in Turkey National Medical Specialization Exams. World of Medical Education, 2025;24(74): 127-134

Abstract

Background: The scientific validity of using artificial intelligence-based applications for studying anatomy and medical specialization exams has begun to be discussed. The aim of this study is to evaluate the performance of ChatGPT 4.0 and Google Gemini in answering anatomy questions from the national medical specialty examination administered in Turkey.

Methods: For this study, anatomy course questions were extracted from examinations held twice annually between 2006-2021 and made openly available on the institutional website. We selected 384 appropriate questions from a total of 400 questions and asked both chatbots simultaneously in an open-ended manner to receive their responses. Questions were classified according to their topics, types, and content. Questions containing clinical information were recorded. Questions with more than 40 words were considered long questions. Questions were divided under systematic anatomy headings according to their subjects (neuroanatomy, locomotor, digestive, respiratory, urogenital, circulatory, and endocrine). ChatGPT 4.0 and Google Gemini 1.5 Pro models were used. Questions were directed to both platforms simultaneously in a single session and in open-ended format. Statistical analysis was performed using IBM SPSS 23 program, with chi-square and Fisher exact tests used to compare independent group rates. Statistical significance level was accepted as $p<0.05$.

Results: 384 anatomy course questions were included in the study. Of these questions, 56 (14.6%) were questions requiring

clinical inference, and 69 questions (18%) were long questions. Overall success rate was found to be 80.7% in ChatGPT 4.0 and 69.3% in Gemini (p<0.001). ChatGPT 4.0 was found to provide more correct answers to questions requiring clinical knowledge and inference than Gemini (ChatGPT 4.0: 91.1%, Gemini: 71.4%) (p=0.007). ChatGPT 4.0 had a statistically significant rate of correct answers to clinically based questions (p=0.021), while Gemini did not show statistical significance in the accuracy of answers given to clinical questions. When examined systematically and topographically, the correct and incorrect answers given were not statistically significant. ChatGPT answered 310 questions correctly (80.7%) and 74 incorrectly (19.3%), while Gemini answered 266 correctly (69.3%) and 118 incorrectly (30.7%). Effect size measurements (Cramér's V) were calculated as 0.127 for overall performance difference and 0.253 for clinical question differences.

Conclusions: The use of ChatGPT 4.0 may be considered more reliable than Gemini in anatomy education and specialization preparation processes. ChatGPT 4.0's superiority is particularly evident in clinical anatomy-based questions. However, it should always be recommended to utilize diverse sources in coordination with reliable literature sources. Considering the changes in anatomy education in recent years based on clinical anatomy and the changes in question styles, it is possible to encounter longer anatomy questions that require clinical inferences and solutions. These findings suggest that AI tools such as ChatGPT 4.0 can be incorporated into clinical anatomy curricula for high-stakes exam preparation. However, human expertise is essential for complex medical scenarios. These data are important for future anatomy education strategies.

Özet

Amaç: Yapay zeka tabanlı uygulamaların anatomi ve tıp uzmanlık sınavlarına hazırlık sürecinde kullanımının bilimsel geçerliliği tartışılmaya başlanmıştır. Bu çalışmanın amacı, Türkiye'de uygulanan ulusal tıp uzmanlık sınavındaki anatomi sorularını yanıtlamada ChatGPT 4.0 ve Google Gemini'nin performansını değerlendirmektir.

Gereç ve Yöntem: Çalışma için 2006-2021 yılları arasında yılda iki kez düzenlenen sınavlardan anatomi dersi sorularını çıkararak kurumsal web sitesinde açık erişim olarak sunulmuş sorular kullanıldı. Toplam 400 sorudan 384 uygun soru seçildi ve her iki chatbot'a aynı anda açık uçlu olarak sorularak yanıtları alındı. Sorular konularına, türlerine ve içeriklerine göre sınıflandırıldı. Klinik bilgi içeren sorular kaydedildi. 40 kelimeden fazla olan sorular uzun sorular olarak kabul edildi. Sorular konularına göre sistematik anatomi başlıkları altında (nöroanatomi, lokomotor, sindirim, solunum, ürogenital, dolaşım ve endokrin) ayrıldı. ChatGPT 4.0 ve Google Gemini 1.5 Pro modelleri kullanılmıştır. Sorular her iki platforma aynı anda, tek bir oturumda ve açık uçlu formatta yönlendirilmiştir. İstatistiksel analizde IBM SPSS 23 programı kullanılarak, bağımsız iki grup oranlarının karşılaştırılmasında ki-kare ve Fisher kesin testleri uygulanmıştır. İstatistiksel anlamlılık düzeyi p<0.05 olarak kabul edilmiştir.

Bulgular: Çalışmaya 384 anatomi dersi sorusu dahil edildi. Bu soruların 56'sı (%14,6) klinik çıkarım gerektiren sorulardı. Soruların 69'u (%18) uzun sorulardı. Genel başarı oranı ChatGPT 4.0'da %80,7, Gemini'de %69,3 olarak bulundu (p<0,001). ChatGPT 4.0'ın klinik bilgi ve çıkarım gerektiren sorulara Gemini'den daha fazla doğru yanıt verdiği görüldü (ChatGPT 4.0: %91,1, Gemini: %71,4) (p=0,007). ChatGPT 4.0, klinik temelli sorulara istatistiksel olarak anlamlı oranda doğru yanıt verdi (p=0,021). Gemini ise klinik sorulara verilen yanıtların doğruluğunda istatistiksel anlamlılık göstermedi. Sistematik ve topografik olarak incelendiğinde, verilen doğru ve yanlış yanıtlar istatistiksel olarak anlamlı değildi. ChatGPT 310 soruyu doğru (%80,7) ve 74'ünü yanlış (%19,3) yanıtlarken, Gemini 266'sını doğru (%69,3) ve 118'ini yanlış (%30,7) yanıtlamıştır. Effect size ölçümleri (Cramér's V) genel performans farkı için 0.127, klinik sorulardaki fark için 0.253 olarak hesaplanmıştır.

Sonuç: Anatomi eğitimi ve uzmanlık hazırlık süreçlerinde ChatGPT 4.0 kullanımı Gemini'ye göre daha güvenilir olarak değerlendirilebilir. Özellikle klinik anatomi temelli sorularda ChatGPT 4.0'ın üstünlüğü belirgindir. Ancak her zaman güvenilir literatür kaynakları ile koordineli olarak farklı kaynaklardan yararlanılması önerilmelidir. Son

yıllarda anatomi eğitiminde klinik anatomiye dayalı değişim ve soru stillerindeki değişim göz önüne alındığında, klinik çıkarım ve çözüm gerektiren daha uzun anatomi sorularıyla karşılaşmak mümkündür. Bu bulgular, ChatGPT 4.0 gibi yapay zeka araçlarının yüksek riskli sınavlara hazırlıkta klinik anatomi müfredatına dahil edilebileceğini göstermektedir. Ancak karmaşık tıbbi senaryolar için insan uzmanlığı gereklidir. Gelecekteki anatomi eğitimi stratejileri için bu veriler önemlidir.

INTRODUCTION

Artificial intelligence (AI) technologies, unlike natural intelligence, are learned models that emerge through programming and have caused significant changes in the field of education in recent years (1). ChatGPT, one of the AI models by OpenAI, has received intense interest worldwide since its introduction in November 2022. The system can perform article production, translation services, and question-answer interactions with its natural language processing capacity (2). These systems, called large language models, also show significant success in processing, analyzing, and presenting medical information.

There are numerous studies evaluating the effectiveness of AI applications in medical education. These studies focus on areas such as clinical problem solving, academic performance, exam success, patient evaluation processes, and patient education (3-8). However, the literature lacks critical discussion regarding the ability of AI models to generate and interpret anatomical content specifically in the context of medical specialty examinations.

The Medical Specialty Examination in Turkey is a strategic evaluation exam that regulates the transition of medical school graduates to specialty training. In this examination, administered twice a year by the Measurement, Selection and Placement Center (ÖSYM), there are questions in the fields of basic medical sciences and clinical sciences. The number of anatomy questions in exams has changed after 2011. These questions are organized under systematic anatomy categories: neuroanatomy, musculoskeletal system, digestive system, respiratory system, urogenital system, cardiovascular system, and endocrine system headings (9).

The aim of this study is to compare the performances of ChatGPT 4.0 and Google Gemini in answering anatomy questions from the national medical specialty examination administered in Turkey according to question content and characteristics. The research hypothesis is that there are significant performance differences between these systems in answering anatomy questions, and these differences will be more pronounced especially in questions requiring clinical inference. This study seeks to answer the following research question: Does ChatGPT 4.0 significantly outperform Gemini in answering clinically-oriented anatomy questions from the Turkish National Medical Specialty Examination?

MATERIALS AND METHODS

In this retrospective comparative study, Medical Specialty Examination anatomy questions prepared by the ÖSYM between 2006-2021 and presented as open access on the institutional website were utilized. Ethical committee approval was not obtained for the research as it did not involve human or animal subjects and was not used for commercial purposes.

From a total of 400 anatomy questions constituting the study universe, 4 cancelled questions and 12 questions with visual content were excluded, and 384 appropriate questions were included in the research. All questions were in multiple-choice format and contained a single correct answer. In the 2006-2011 period, two exams were administered annually with 10 anatomy questions in each exam. After 2011, two exams per year continued, but the number of anatomy questions in each exam was increased to 14.

Questions were systematically classified according to their topical content, type, and characteristics. Questions requiring clinical knowledge and inference were evaluated in a separate category. Questions exceeding 40 words were defined as long questions. In terms of topical distribution, questions were divided into systematic anatomy headings: neuroanatomy, musculoskeletal system, digestive system, respiratory system, urogenital system, cardiovascular system, and endocrine system categories.

Prompt Design and AI Model Specifications

The AI models used in this study were ChatGPT 4.0 (accessed May 2024) and Google Gemini

1.5 Pro (accessed May 2024). All questions were directed to both platforms simultaneously, in one session, and in open-ended format. Sample prompt structure used for both AI platforms: "I will present multiple-choice anatomy questions from the Turkish National Medical Specialization Exam (TUS). Each question has five options (A-E) and only one correct answer. Please indicate the correct answer (A-E)".
The responses provided by the AI platforms to the questions were recorded as correct or incorrect by comparing them with the ÖSYM answer key.
Statistical analyses were performed using IBM SPSS 23 program. Mean ± standard deviation was used for continuous variables, and frequency and percentage values were used for categorical variables. Chi-square and Fisher exact tests were applied for comparing independent two-group proportions. Effect size measures (Cramér's V) were calculated for statistically significant associations to provide deeper statistical interpretation. Statistical significance level was accepted as $p<0.05$.

RESULTS

Of the 384 anatomy questions included in the research, 56 (14.6%) were of a nature requiring clinical inference, and 69 (18.0%) were in the long question category. The topical distribution of questions is presented in Table 1.

Table 1. Numerical Distribution of TUS Questions by Systematic Anatomy (2006-2021).

System Category	Frequency	Percent	Cumulative Percent
Neuroanatomy	132	34.4	34.4
Locomotor	114	29.7	64.1
Circulatory	70	18.2	82.3
Respiratory	12	3.1	85.4
Digestive	42	10.9	96.4
Urogenital	13	3.4	99.7
Endocrine	1	0.3	100.0
Total	384	100.0	

General performance analysis of ChatGPT 4.0 showed that it gave correct answers to 310 of 384 questions (80.7%) and produced incorrect answers in 74 (19.3%). Gemini gave correct answers to 266 questions (69.3%) and incorrect answers to 118 questions (30.7%). This performance difference between the two programs was found to be statistically significant ($p<0.001$). Cramér's V was calculated as 0.127, indicating a small to moderate effect size.

Analysis of answers given to clinically based questions revealed that ChatGPT 4.0 exhibited distinct superiority in this category ($p=0.021$). While it answered 51 of 56 questions requiring clinical inference correctly (91.1%), Gemini could give correct answers to 40 questions (71.4%) in the same category. This difference is statistically significant ($p=0.007$) with a Cramér's V of 0.253, indicating a moderate effect size (Table 2).

Table 2. Distribution of Correct Answers to Clinical Questions by ChatGPT 4.0 and Gemini 1.5 Pro.

Clinical Question	ChatGPT 4.0		Total	p	Gemini 1.5 Pro		Total	p
	True	False			True	False		
No	259 (79%)	69 (21%)	328 (100%)	**0.021**	226 (68.9%)	102 (31.1%)	328 (100%)	0.418
Yes	51 (91.1%)	5 (8.9%)	56 (100%)		40 (71.4%)	16 (28.6%)	56 (100%)	
Total	310 (80.7%)	74 (19.3%)	384 (100%)		266 (69.3%)	118 (30.7%)	384 (100%)	

When the effect of question length on answer accuracy was evaluated, no statistically significant change was observed in ChatGPT 4.0's performance. However, it was determined that Gemini showed a tendency to increase error rate in long questions, but this increase had limited significance (error rate for long questions: 39.1%, error rate for short questions: 28.9%, p=0.065) (Table 3). In the analysis conducted according to systematic and topographic anatomy categories, it was determined that both platforms exhibited similar distribution patterns and showed no statistically significant difference. The distribution of answers is visualized in Figure 1.

Table 3. Performance Comparison of ChatGPT 4.0 and Gemini by Question Length.

Question Length	ChatGPT 4.0		Total	p	Gemini 1.5 Pro		Total	p
	True	False			True	False		
Long	56 (81.2%)	13 (18.8%)	69	0.536	42 (60.9%)	27 (39.1%)	69	0.065
Short	254 (80.6%)	61 (19.4%)	315		224 (71.1%)	91 (28.9%)	315	
Total	310	74	384		266	118	384	

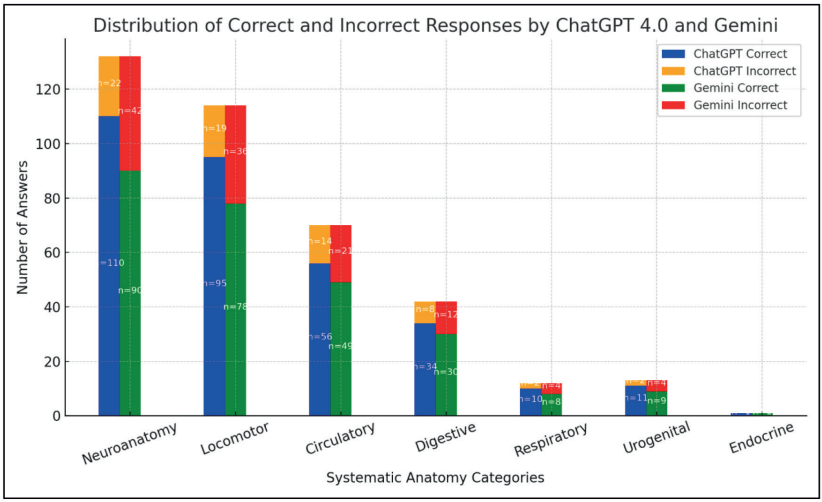

Figure 1. Distribution of Correct and Incorrect Responses Provided by ChatGPT 4.0 and Gemini According to Systematic Anatomy Categories (n=384). (X-axis represents systematic anatomy categories; Y-axis shows number of answers. Blue bars: ChatGPT 4.0 correct answers, Red bars: ChatGPT 4.0 incorrect answers, Green bars: Gemini correct answers, Orange bars: Gemini incorrect answers).

Detailed subgroup analyses showed that ChatGPT 4.0 exhibited more consistent performance especially in questions requiring complex anatomical relationships. The platform showed an 84.2% success rate in questions requiring multi-system integration, while Gemini achieved 71.8% success in this category.

DISCUSSION

This study is the first comprehensive research comparing the performance of different AI programs in national medical specialty examination anatomy questions. The findings obtained show that

ChatGPT 4.0 exhibits distinct superiority compared to Gemini, and this difference becomes even more pronounced especially in questions requiring clinical inference.

Numerous studies evaluating ChatGPT's performance in different areas of medicine have been conducted in the last two years. In addition to clinical science examinations such as dermatology (10-12), orthopedics and traumatology (13,14), urology (15), specialty and national license exam performances in different countries have also been systematically examined. The common finding of these studies is that ChatGPT generally exhibits adequate performance but shows variable success rates in specific areas (16-20).

In the comparative analysis conducted by Alessandri-Bonetti and colleagues, it was reported that ChatGPT 4.0 showed significant superiority compared to Gemini in the burn support examination (90% vs. 70%) (21). In the evaluation conducted by Fowler and colleagues on ophthalmology exam questions, it was also documented that ChatGPT performed more effectively than Gemini (22). The findings of the current study support the superiority of ChatGPT 4.0 in anatomy questions, consistent with this literature.

Considering the central role of anatomy in medical education, the competencies of AI platforms in this field are of critical importance. The high performance shown by ChatGPT 4.0 in clinical questions (91.1%) in our study demonstrates that this platform can successfully achieve clinical anatomy integration. This finding parallels the study by Kung and colleagues, who emphasized ChatGPT's potential in USMLE questions (23).

However, some limitations regarding the use of AI platforms in medical education should also be considered. Totlis and colleagues stated that these systems may not always produce correct answers on the first attempt and their use in anatomy education may be limited, because anatomy is usually taught before students gain experience in patient interaction, diagnostic reasoning, and examining patient records (24). In the study by Mantzou and colleagues evaluating the consistency of ChatGPT 3.5 in musculoskeletal system anatomy examination, it was suggested that responses showed variability and could not be an independent, reliable source, therefore information should be verified with anatomy literature (25).

Factors such as prompt engineering and user-dependent variability may significantly influence AI performance in educational contexts. The way questions are formulated, the level of detail provided in prompts, and individual user interaction patterns could affect the accuracy and consistency of responses provided by these AI systems. Future studies should investigate the impact of different prompting strategies and user interaction styles on AI performance in medical education.

In the comparative study by Ilgaz and Çelik testing question generation and article writing skills in anatomy, it was shown that accuracy rates were quite variable (26). These findings indicate that AI platforms can be used as supportive tools but are not sufficient alone.

Similar studies conducted in the field of dentistry have shown that the performance of generative pre-trained transformers in answering anatomy questions can be variable (27). Studies examining the evolution of anatomy questions in medical specialty examinations in Turkey over the years have revealed that changes in question types can affect AI performance (9).

The strengths of the current study include the use of long-term data sets, standardized evaluation criteria, and objective evaluation using the answer key of the ÖSYM exam questions. However, the study also has some limitations. Questions were asked only once and first responses were evaluated. Exclusion of questions with visual content and use of only Turkish questions may affect the generalizability of results. Additionally, since AI programs are continuously trained by updating their data sets, the performance of programs may change in subsequent studies.

Future research should evaluate the performance of AI platforms with questions in different languages and with visual content, and also investigate the effect of multiple-attempt strategies. Furthermore, comparing student performance with AI performance will be valuable for understanding the real impact of these systems on education quality. Educational institutions should consider developing guidelines for the appropriate integration of AI tools in medical curricula, particularly focusing on how these technologies can complement traditional teaching methods in clinical anatomy education.

CONCLUSION

This research has shown that ChatGPT 4.0 exhibits distinct superiority compared to Google Gemini in national medical specialty examination anatomy questions. Particularly, ChatGPT 4.0's performance in questions requiring clinical inference indicates that this platform is a more reliable tool in anatomy education and specialty preparation processes. However, it is of critical importance that AI platforms be used as supportive tools in medical education and that different sources always be utilized in coordination with reliable literature sources.

Considering the paradigm shift based on clinical anatomy in anatomy education in recent years and the evolution in question styles, the possibility of encountering more complex anatomy questions requiring clinical inference and analytical thinking is increasing. In this context, ChatGPT 4.0's more successful performance in both situations suggests that AI-based tools like ChatGPT 4.0 may be effectively incorporated into clinical anatomy curricula, particularly in preparing for high-stakes exams. These findings provide important data for determining future anatomy education strategies and highlight the potential for AI integration in medical education while emphasizing the continued importance of human expertise and critical evaluation.

References

1. Meroueh C, Chen ZE. Artificial intelligence in anatomical pathology: building a strong foundation for precision medicine. Hum Pathol. 2023;132:31-8.
2. Mogali SR. Initial impressions of ChatGPT for anatomy education. Anatomical sciences education. 2024;17(3):444-7.
3. Pirkle S, Yang J, Blumberg TJ. Do ChatGPT and Gemini Provide Appropriate Recommendations for Pediatric Orthopaedic Conditions? J Pediatr Orthop. 2024;44(8):e123-9.
4. Peters M, Leclercq M, Yanni A, Eynden XV, Martin L, Haute NV, et al. ChatGPT and Trainee performances in the management of maxillofacial patients. J Stomatol Oral Maxillofac Surg. 2024;125(4):102090.
5. Al-Sharif EM, Penteado RC, Dib El Jalbout N, Topilow NJ, Shoji MK, Kikkawa DO, et al. Evaluating the Accuracy of ChatGPT and Google BARD in Fielding Oculoplastic Patient Queries: A Comparative Study on Artificial versus Human Intelligence. Ophthal Plast Reconstr Surg. 2024;40(3):303-11.
6. Mayo-Yáñez M, Lechien JR, Maria-Saibene A, Vaira LA, Maniaci A, Chiesa-Estomba CM. Examining the Performance of ChatGPT 3.5 and Microsoft Copilot in Otolaryngology: A Comparative Study with Otolaryngologists' Evaluation. Indian J Otolaryngol Head Neck Surg. 2024;76(4):3465-9.
7. Meral G, Ateş S, Günay S, Öztürk A, Kuşdoğan M. Comparative analysis of ChatGPT, Gemini and emergency medicine specialist in ESI triage assessment. Am J Emerg Med. 2024;81:146-50.
8. Qin S, Chislett B, Ischia J, Ranasinghe W, de Silva D, Coles-Black J, et al. ChatGPT and generative AI in urology and surgery-A narrative review. BJUI Compass. 2024;5(9):813-21.
9. Aygün, T., Keskin, A., & Yücel, N. Changes in the types of anatomy questions asked in the medical specialization exam over the years, Türkiye example. BMC Medical Education. 2025;25(1): 607.
10. 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.
11. Reddy S, Schwartzman G, Flowers RH. ChatGPT in Dermatology Clinical Practice: Potential Uses and Pitfalls. Cutis. 2023;112(2):E15-7.
12. 12. D'Agostino M, Feo F, Martora F, Genco L, Megna M, Cacciapuoti S, et al. ChatGPT and dermatology. Ital J Dermatol Venereol. 2024;159(4):234-40.
13. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations. J Am Acad Orthop Surg. 2023;31(23):1173-9.
14. Jain N, Gottlich C, Fisher J, Campano D, Winston T. Assessing ChatGPT's orthopedic in-service training exam performance and applicability in the field. J Orthop Surg Res. 2024;19(1):27.

15. Schoch J, Schmelz HU, Strauch A, Borgmann H, Nestler T. Performance of ChatGPT-3.5 and ChatGPT-4 on the European Board of Urology (EBU) exams: a comparative analysis. World J Urol. 2024;42(1):445.

16. 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.

17. Ishida K, Hanada E. Potential of ChatGPT to Pass the Japanese Medical and Healthcare Professional National Licenses: A Literature Review. Cureus. 2024;16(8):e66324.

18. Takagi S, Koda M, Watari T. The Performance of ChatGPT-4V in Interpreting Images and Tables in the Japanese Medical Licensing Exam. JMIR Med Educ. 2024;10:e54283.

19. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ. 2024;24(1):143.

20. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: An observational study. Medicine (Baltimore). 2023;102(32):e34673.

21. Alessandri-Bonetti M, Liu HY, Donovan JM, Ziembicki JA, Egro FM. A Comparative Analysis of ChatGPT, ChatGPT-4, and Google Bard Performances at the Advanced Burn Life Support Exam. J Burn Care Res. 2024;45(4):945-8.

22. Fowler T, Pullen S, Birkett L. Performance of ChatGPT and Bard on the official part 1 FRCOphth practice questions. Br J Ophthalmol. 2024;108(10):1379-83.

23. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.

24. Totlis T, Natsis K, Filos D, Ediaroglou V, Mantzou N, Duparc F, et al. The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. Surg Radiol Anat. 2023;45(10):1321-9.

25. Mantzou N, Ediaroglou V, Drakonaki E, Syggelos SA, Karageorgos FF, Totlis T. ChatGPT efficacy for answering musculoskeletal anatomy questions: a study evaluating quality and consistency between raters and timepoints. Surg Radiol Anat. 2024;46(9):1455-64.

26. İlgaz HB, Çelik Z. The importance of artificial intelligence platforms in anatomy education: ChatGPT and Google Bard experience. Turkish Clinics Journal of Anatomy. 2023;15(3):45-52.

27. Keskin A., Aygun, T. A Performance of Generative Pre-Trained Transformers (GPT) in Answering Questions on Anatomy in The Turkish Dentistry Specialization Exam. JITSI: Jurnal Ilmiah Teknologi Sistem Informasi. (2024); 5(4): 188-192.