Original Investigation / Orijinal Araştırma

DOI: 10.4274/tod.galenos.2025.04875
Turk J Osteoporos

Performance Comparison of Gemini, DeepSeek, and ChatGPT-4o on American Board of Physical Medicine and Rehabilitation Board Exam Practice Questions

Amerikan Fiziksel Tıp ve Rehabilitasyon Kurulu Sınavı Deneme Sorularında Gemini, DeepSeek ve ChatGPT-4o'nun Performans Karşılaştırması

Gonca Sağlam Akkaya, Hanife Baykal Şahin

Karadeniz Technical University Faculty of Medicine, Department of Physical Medicine and Rehabilitation, Trabzon, Türkiye

Abstract

Objective: The rapid advancement of large language models (LLMs) has demonstrated their important potential in medical education and assessment. This study aimed to evaluate the performance of three prominent LLMs (Gemini, DeepSeek, and ChatGPT-4o) on practice questions designed to be representative of the American Board of Physical Medicine and Rehabilitation (ABPMR) certification examination. By comparing their accuracy across various medical domains, we sought to understand their current capabilities as supplementary tools for medical trainees.

Materials and Methods: We used a comprehensive set of 100 publicly available ABPMR practice questions from 2015, ensuring a consistent benchmark for comparison. These questions, which cover a wide range of topics and clinical scenarios, were systematically fed into Gemini, DeepSeek, and ChatGPT-4o via their web interfaces. The responses were then independently analyzed by a blinded physical medicine and rehabilitation specialist to ensure an unbiased evaluation.

Results: DeepSeek achieved the highest overall accuracy at 88%, significantly outperforming Gemini (81%, p=0.022) but not showing a statistically significant difference compared to ChatGPT-4o (86%, p=0.238). The models displayed varying strengths across different specialty areas. ChatGPT-4o performed best in Neurologic disorders (90%) and electrodiagnosis (87%). In contrast, DeepSeek led in musculoskeletal medicine (88%), patient management (97%), and amputation (100%). Gemini performed comparably to DeepSeek in equipment/assistive technology (90%). No significant inter-model differences were found in domains such as rehabilitation problems (93%), basic sciences (80%), and applied sciences (83%).

Conclusion: Our findings suggest that while DeepSeek demonstrated superior aggregate performance, all three LLMs possess unique, complementary strengths across different domains of physical medicine and rehabilitation. The lack of significant differences in domain-stratified analyses points to the task-specific nature of LLM efficacy. These results indicate that LLMs are promising supplementary educational tools, but their persistent limitations in complex clinical reasoning necessitate continued human oversight and validation

Keywords: Large language models, physical medicine and rehabilitation, medical education

Öz

Amaç: Büyük dil modellerinin (BDM'ler) hızlı gelişimi, tıp eğitimi ve değerlendirmesinde önemli bir potansiyel göstermiştir. Bu çalışmanın amacı, önde gelen üç BDM olan Gemini, DeepSeek ve ChatGPT-4o'nun, Amerikan Fiziksel Tıp ve Rehabilitasyon Kurulu (ABPMR) sertifika sınavını temsil eden deneme sorularını yanıtlama performansını değerlendirmekti. Bu modellerin tıp öğrencileri için yardımcı araçlar olarak mevcut yeteneklerini anlamak için farklı tıbbi alanlardaki doğruluklarını karşılaştırma hedeflendi.

Gereç ve Yöntem: 2015 yılında erişime sunulmuş olan 100 adet ABPMR deneme sorusundan oluşan kapsamlı bir set kullandıldı. Bu sorular, geniş konu çeşitliliği ve klinik senaryoları kapsamakta olup, Gemini, DeepSeek ve ChatGPT-4o'nun web arayüzlerine sistematik bir şekilde girildi. Yanıtlar, tarafsız bir değerlendirme sağlamak amacıyla, hangi BDM tarafından üretildiği bilinmeyen (körleme yöntemi) bağımsız bir fiziksel tıp ve rehabilitasyon uzmanı tarafından analiz edildi.

Corresponding Author/Sorumlu Yazar: Assoc. Prof. Gonca Sağlam Akkaya MD, Karadeniz Technical University Faculty of Medicine, Department of Physical Medicine and Rehabilitation, Trabzon, Türkiye
E-mail: goncasaglam@hotmail.com **ORCID ID:** orcid.org/0000-0001-7713-4435
Received/Geliş Tarihi: 14.08.2025 **Accepted/Kabul Tarihi:** 22.09.2025 **Epub:** 08.10.2025

Cite this article as/Atıf: Sağlam Akkaya G, Baykal Şahin H. Performance comparison of Gemini, DeepSeek, and ChatGPT-4o on American board of physical medicine and rehabilitation board exam practice questions. Turk J Osteoporos. [Epub Ahead of Print]

Öz

Bulgular: DeepSeek, %88 ile en yüksek genel doğruluğa ulaştı. Gemini'den (%81, p=0,022) önemli ölçüde daha iyi performans göstermiş, ancak ChatGPT-4o'dan (%86, p=0,238) istatistiksel olarak anlamlı bir farkla ayrılmamıştı. Modeller, farklı uzmanlık alanlarında değişen güçlü yönler sergiledi. ChatGPT-4o, nörolojik bozukluklar (%90) ve elektrodiyagnoz (%87) alanlarında en yüksek performansı gösterdi. Buna karşılık, DeepSeek kas-iskelet tıbbı (%88), hasta yönetimi (%97) ve ampütasyon (%100) alanlarında lider oldu. Gemini ise ekipman/yardımcı teknoloji (%90) alanında DeepSeek ile benzer bir performans sergiledi. Rehabilitasyon sorunları (%93), temel bilimler (%80) ve uygulamalı bilimler (%83) gibi alanlarda ise modeller arasında anlamlı bir fark bulunmadı.

Sonuç: Bulgularımız, DeepSeek'in genel performansta üstünlük gösterse de, her üç BDM'nin de fiziksel tıp ve rehabilitasyonun farklı alanlarında benzersiz ve tamamlayıcı güçlü yönlere sahip olduğunu düşündürmektedir. Alana göre yapılan analizlerde istatistiksel olarak anlamlı farklılıkların bulunmaması, BDM etkinliğinin göreve özgü değişkenliğini vurgulamaktadır. Bu sonuçlar, BDM'lerin tıp eğitiminde umut verici ek araçlar olduğunu göstermekle birlikte, karmaşık klinik muhakemedeki kalıcı sınırlamaları nedeniyle insan gözetiminin ve doğrulamasının kritik önemini koruduğunu vurgulamaktadır.

Anahtar kelimeler: Büyük dil modelleri, fiziksel tıp ve rehabilitasyon, tıp eğitimi

Introduction

The field of artificial intelligence (AI) has seen rapid advancements in recent years, particularly in large language models (LLMs). They have demonstrated their potential to revolutionize various fields, including healthcare and medical education (1,2). These models, trained on vast datasets of text and code, can generate human-like text, translate languages, write different kinds of creative content, and answer questions in an informative way (3). In the medical domain, LLMs are being explored for applications ranging from assisting with clinical decision-making (4) and summarizing medical records (5) to generating patient education materials (6) and potentially aiding in exam preparation.

The American Board of Physical Medicine and Rehabilitation (ABPMR) examination serves as a crucial benchmark for physicians specializing in this field, assessing their knowledge and clinical reasoning skills (7). Success on this exam is essential for board certification and signifies competency in the specialty. This study aimed to evaluate the performance of three prominent LLMs; Gemini, DeepSeek, and ChatGPT-4, in their ability to answer questions representative of ABPMR practice questions.

Materials and Methods

Data Collection

We obtained a comprehensive set of 100 practice questions from the ABPMR. These questions were sourced from publicly available practice materials and previous examination sets, ensuring a representative sample of the board's assessment style and content. They were released by the ABPMR in June 2015 as a study tool and have been permanently removed from their active examination item banks. They present a wide range of topics relevant to physical medicine and rehabilitation (PMR). These questions spanned all core domains of PMR, distributed as follows: Neurologic disorders (30 questions), musculoskeletal medicine (32 questions), electrodiagnosis (15 questions), amputation (5 questions), rehabilitation problems (15 questions), basic sciences (15 questions), and applied sciences (15 questions). Additionally, questions were categorized by focus: Patient management (32 questions) and equipment/assistive technology (10 questions). The questions are in a multiple-choice format, often presenting clinical scenarios, designed to assess foundational knowledge and clinical reasoning relevant to the certification examination. The original document includes an answer key. The static nature and prior public release of these questions ensure a consistent and accessible benchmark for comparing the performance of LLM.

AI Models and Question Processing

Three state-of-the-art LLMs were selected for this study: Gemini, DeepSeek, and ChatGPT-4. Each model was accessed through its web interface, using the most recent available versions. The ABPMR questions were formatted and input into each AI model without modification. We ensured that the input format was consistent across all three models to maintain fairness in the comparison.

Response Generation

Each AI model was prompted with the ABPMR questions individually. The models were instructed to provide their best attempt at answering each question without any additional context or information beyond what was provided in the question itself. The analysis was performed after a total of three attemps. A PMR specialist, blinded to which AI model generated each response, independently scored the answers. The performance was assessed by another specialist who calculated the overall accuracy, percentage and number of correctly answered questions, for each LLM. Additionally, the performances were assessed across different question topics to identify potential strengths and weaknesses of each model in specific areas of PMR.

This study was conducted in compliance with ethical guidelines for AI research. No patient data or confidential examination materials were used. The study focused solely on the AI models' performance on publicly available practice questions. This study did not require ethical approval as it was based on publicly

available data and did not involve human participants or private medical information.

Statistical Analysis

The statistical analysis was performed using SPSS version 29 (IBM, Armonk, NY). Chi-squared tests of independence were used for categorical comparisons of correct versus incorrect responses within each question category. Fisher's exact test was used in categories with small sample sizes, such as "amputation". To compare the mean accuracy percentages across the three models, a One-Way ANOVA was conducted for overall performance, as well as for questions categorized by organ system and by question focus. When a significant result was found in the ANOVA for overall performance, a Tukey HSD post-hoc analysis was performed to identify specific pairwise differences between the models, with adjustments made to control for multiple comparisons. Key assumptions for the ANOVA, including homogeneity of variances (verified by Levene's test) and normality of residuals (confirmed by Shapiro-Wilk tests), were checked and met.

Results

All models showed incremental improvement from first to third attempts, with DeepSeek maintaining the highest accuracy at each stage (86%, 86%, 88%). DeepSeek significantly outperformed Gemini (p=0.022), though no significant differences existed between DeepSeek and ChatGPT-4o (p=0.238) or ChatGPT-4o and Gemini (p=0.100) (Table 1). Table 2 presents the summary of key findings.

When stratified by organ system, model performance varied by specialty. In neurologic disorders (30 questions), ChatGPT-4o achieved the highest accuracy (90%), exceeding Gemini (80%) and DeepSeek (87%). For musculoskeletal medicine (32 questions), DeepSeek led (88%) over Gemini (81%) and ChatGPT-4o (84%). DeepSeek achieved perfect accuracy in amputation (5 questions, 100%), while Gemini and ChatGPT-4o both scored 80%. All models performed equally in rehabilitation problems (93%) and basic sciences (80%). No significant differences across models for organ system-based performance was found (F=1.12, p=0.350), consistent with individual categories (all p>0.05) (Figure 1).

Analysis by question focus revealed additional task-specific strengths. In electrodiagnosis (15 questions), ChatGPT-4o excelled (87%) over DeepSeek (80%) and Gemini (67%). DeepSeek dominated patient management (32 questions, 97%) compared to ChatGPT-4o (91%) and Gemini (88%). For equipment/assistive technology (10 questions), Gemini and DeepSeek tied (90%), outperforming ChatGPT-4o (80%). No model differences emerged in applied sciences (all 83%). No significant difference was observed for inter-model variation (F=0.63, p=0.548), aligning with non-significant chi-squared tests for all focus categories (Figure 2).

Discussion

LLMs are creating a major change in medical education, allowing for new uses in knowledge sharing, customized self-assessment tools, and simulated practice of clinical reasoning. As these systems are increasingly used for high-stakes tasks like preparing for board exams and assisting with clinical decisions, it is crucial to rigorously evaluate their accuracy, limitations, and potential biases to ensure safe implementation (8). This imperative is underscored by documented instances of LLMs generating plausible yet incorrect medical information, highlighting critical gaps in reliability (1,9).

Table 1. Comparison of overall performances of AI models (Gemini, DeepSeek and ChatGPT-4o) for ABPMR board practice questions

		Gemini (%)	Deepseek (%)	ChatGPT-4o (%)	p
1st attemtp	Incorrect	20	14	16	0.21
	Correct	80	86	84	
2nd attemp	Incorrect	19	14	16	0.41
	Correct	81	86	84	
3rd attempt	Incorrect	16	12	13	0.17
	Correct	84	88	87	

AI: Artificial intelligence, ABPMR: American Board of Physical Medicine and Rehabilitation

Table 2. Summary of findings

Metric	Gemini	DeepSeek	ChatGPT-4o
Highest overall accuracy	81.7%	86.7%	85.0%
Top specialty area	Equipment and assistive tecnology, medical rehabilitation	Musculoskeletal medicine, patient management, amputation	Neurologic disorders, electrodiagnosis
Statistical advantage	None	Outperformed Gemini (p=0.022)	None

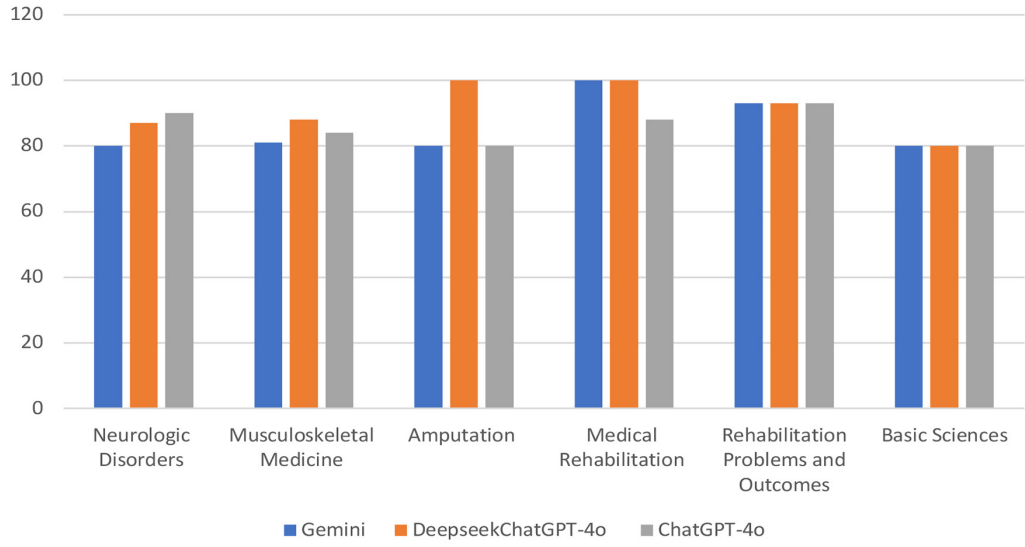

Figure 1. Comparative performance percentages of AI models (Gemini, DeepSeek, and ChatGPT-4o) for "type of problem/organ system" practice questions
AI: Artificial intelligence

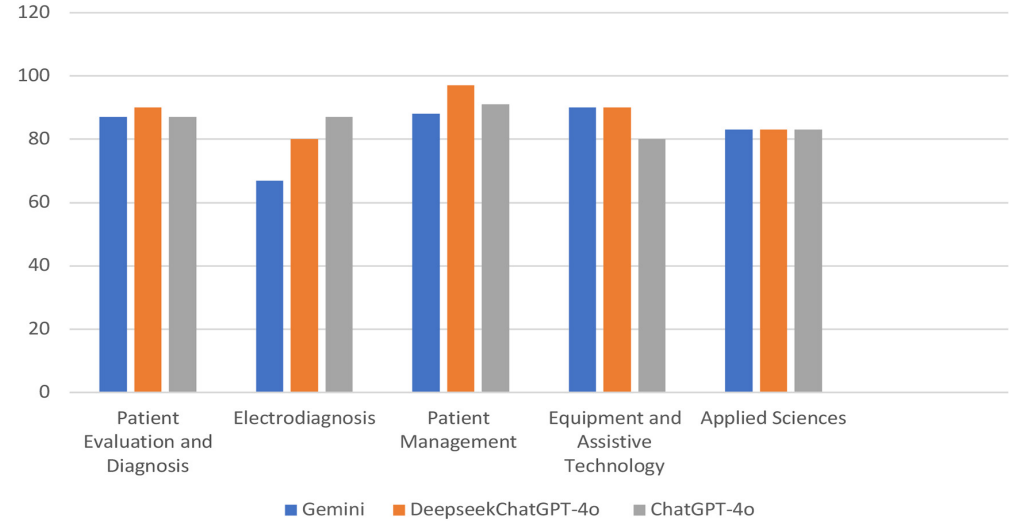

Figure 2. Comparative performance percentages of AI models (Gemini, DeepSeek, and ChatGPT-4o) for "focus of question and patient management" practice questions
AI: Artificial intelligence

Our analysis indicates that while all three LLMs demonstrated a certain level of proficiency in answering ABPMR-style questions, their accuracy varied significantly across different domains. DeepSeek exhibited the highest overall accuracy, suggesting that its training data and model architecture may be better optimized for medical reasoning tasks. This aligns with prior research demonstrating that advanced LLMs can achieve near-expert performance in certain medical domains. Gemini and DeepSeek, while also performing well, showed slightly lower accuracy rates, potentially due to differences in training methodologies and dataset composition (2).

While our study focused on the field of PMR, the utility of LLMs in other medical specialties has also been explored, with promising results in rheumatology (10). Recent studies have specifically evaluated the performance of various LLMs on rheumatology board-level questions, finding that they can achieve high accuracy. A comparative study using questions from the American College of Rheumatology's CARE-2022 Question Bank found that GPT-4 demonstrated a 78% accuracy rate, outperforming Claude 3: Opus (63%) and Gemini Advanced (53%) (11). Another study, which assessed the performance of LLMs on the Spanish access exam to specialized medical training (MIR), reported that GPT-4 achieved a remarkable accuracy of 93.71% on rheumatology questions, significantly higher than ChatGPT's 66.43% (12).

Similar to its utility in other medical specialties, AI and LLMs show significant promise for orthopedic study and board exam preparation. Recent studies have assessed the performance of various LLMs, including GPT-4 and Google Gemini, on standardized tests like the orthopaedic in-training examination (OITE) and the Turkish orthopedics and traumatology board examination (13,14). One study found that GPT-4 performed at the level of a third-year resident on the 2021 OITE (15). Another study using the 2022 OITE found that Google Gemini was the most accurate model, correctly answering 69.9% of questions, a performance level approaching that of fourth- and fifth-year residents (16).

While LLMs generally perform well on text-based, knowledge-recall questions, their accuracy can be significantly lower on questions that include images (17). Moreover, a key limitation highlighted by these studies is the tendency for LLMs to provide inaccurate in response to complex or fact-based questions (14). This suggests that while these tools are becoming valuable for their ability to provide explanations and enhance learning, they are not yet a substitute for human clinical judgment and must be used with caution and careful verification of their output.

LLMs are also showing considerable potential in neurology, with studies indicating their effectiveness in answering board-style questions and assisting with clinical reasoning (18,19). A study evaluating several LLMs on questions from the self-assessment in neurological surgery American Board of Neurological Surgery Primary Board Examination Review found that all models exceeded the passing threshold. The highest accuracy was achieved by OpenAI o1 (87.6%), followed by Claude 3.5 Sonnet (83.2%) and Gemini 2.0 (81.0%) (20). Another study, which specifically evaluated GPT-4 on neurology board-style questions, reported an accuracy rate of 75.0%, outperforming the average human test-taker score of 69% and the passing score of 70%. This study also highlighted that GPT-4 performed particularly well in subspecialties like neuromuscular disorders, pharmacology, and cognitive and behavioral disorders (21). While these results are promising for medical education and exam preparation, the studies also underscore limitations, such as lower performance on questions involving images and a need for continued physician supervision to ensure accuracy and reliability.

According to our results, DeepSeek excelled in musculoskeletal medicine and patient management, ChatGPT-4o led in neurologic disorders and electrodiagnosis, and Gemini performed strongly in equipment/assistive technology and medical rehabilitation (100%). All models achieved parity in rehabilitation problems and applied sciences. These findings suggest that while DeepSeek holds an aggregate advantage, task-specific expertise varies across models.

Our observed higher accuracy (e.g., DeepSeek: 88%, ChatGPT-4o: 86%) compared to prior studies in specialties like rheumatology (GPT-4: 78%) or orthopedics (Gemini: 69.9%) likely reflects rapid advancements in LLM capabilities rather than methodological differences alone. Key drivers include iterative model evolution (e.g., optimizations from GPT-4 to GPT-4o), architectural refinements improving clinical reasoning, and potential task-specific fine-tuning enhancing performance on medical benchmarks. This trajectory of technical progress suggests LLMs are steadily narrowing the accuracy gap with human expertise across medical domains, though persistent limitations in handling novel scenarios warrant ongoing validation.

Study Limitations

Despite these promising results, our study also underscores several limitations in the current generation of LLMs. Notably, none of the models achieved perfect accuracy, indicating that they are still prone to errors in medical reasoning and knowledge retrieval. Previous studies have also pointed out that LLMs can occasionally generate incorrect or misleading information, particularly when dealing with complex clinical scenarios (4). The ability of these models to handle complex and specialized knowledge makes them valuable tools for clinicians and trainees, potentially serving as a supplementary aid for exam preparation and continuing medical education (22).

Another key limitation is the evolving nature of AI models. As LLMs continue to be updated and refined, their performance on medical assessments may improve, necessitating ongoing evaluation and benchmarking. Moreover, while our study utilized publicly available ABPMR-style questions, it is possible that these questions do not fully capture the depth and breadth of the actual board examination. Further studies incorporating a larger and more diverse set of questions, as well as real-world clinical case evaluations, would provide a more comprehensive understanding of these models' capabilities.

Understanding the capabilities and limitations of these LLMs in this context is crucial for determining their potential role in medical education and assessment, while also highlighting areas where human oversight and expertise remain essential. We acknowledge potential limitations, including the evolving nature of AI models and the possibility that the questions used may not fully represent the current ABPMR examination. Additionally, we recognize that AI models' performance may not directly translate to clinical competence or decision-making ability.

Conclusion

This study highlights the growing potential of LLMs such as Gemini, DeepSeek, and ChatGPT-4 in medical education and assessment. While these models demonstrate impressive accuracy in answering board-style questions, their limitations emphasize the need for human oversight and further refinements to ensure reliability in clinical decision-making contexts. Future research should focus on improving LLM interpretability, optimizing training datasets, and developing hybrid AI-human systems to enhance the effectiveness of AI-assisted medical education.

Ethics

Ethics Committee Approval: This study did not require ethical approval as it was based on publicly available data.

Informed Consent: Informed consent was not required for this study as it did not involve human participants or private medical information.

Footnotes

Authorship Contributions

Concept: G.S.A., Design: G.S.A., Data Collection or Processing: G.S.A, H.B.Ş., Analysis or Interpretation: G.S.A, H.B.Ş., Literature Search: G.S.A., Writing: G.S.A.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2:230-43.
2. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115-8. Erratum in: Nature. 2017;546:686.
3. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017;30:1-11.
4. Topol E. Deep medicine: how artificial intelligence can make healthcare human again. New York: Basic Books; 2019.
5. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620:172-80. Erratum in: Nature. 2023;620:E19.
6. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne). 2024;11:1477898.
7. American Board of Physical Medicine and Rehabilitation. About the ABPMR. Available from: https://www.abpmr.org/About
8. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388:1233-9.
9. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023;11:887.
10. Is EE, Menekseoglu AK. Comparative performance of artificial intelligence models in rheumatology board-level questions: evaluating Google Gemini and ChatGPT-4o. Clin Rheumatol. 2024;43:3507-13.
11. Flores-Gouyonnet J, Cuéllar-Gutiérrez MC, Figueroa-Parra G, Kimbrough B, Joerns EK, Navarro-Mendoza E, et al. Performance of large language models in rheumatology board-like questions: accuracy, quality, and safety. Lancet Rheumatol. 2025;7:e152-4.
12. Madrid-García A, Rosales-Rosado Z, Freites-Nuñez D, Pérez-Sancristóbal I, Pato-Cour E, Plasencia-Rodríguez C, et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. Sci Rep. 2023;13:22129.
13. Ghanem D, Covarrubias O, Raad M, LaPorte D, Shafiq B. ChatGPT Performs at the level of a third-year orthopaedic surgery resident on the orthopaedic in-training examination. JB JS Open Access. 2023;8:e23.00103.
14. Yaş S, Ahmadov A, Baymurat AC, Tokgöz MA, Coşkun Yaş S, Odluyurt M, et al. ChatGPT vs. orthopedic residents! Who is the winner? GMJ. 2024;35:185-91.
15. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB. Evaluating ChatGPT performance on the orthopaedic in-training examination. JB JS Open Access. 2023;8:e23.00056.
16. Pamuk Ç, Uyanık AF, Kuyucu E, Uğurlar M. Can ChatGPT pass the Turkish Orthopedics and Traumatology Board Examination? Turkish orthopedic surgeons versus artificial intelligence. Ulus Travma Acil Cerrahi Derg. 2025;31:310-5.
17. Nawari A, Zahir J, Kumar S, Ocampo L, Opara O, Ahmad H, et al. Artificial intelligence large language models are nearly equivalent to fourth-year orthopaedic residents on the orthopaedic in-training examination: a cause for concern or excitement? J Orthop Educ Inst. 2025;6.
18. Schubert MC, Wick W, Venkataramani V. Performance of large language models on a neurology board-style examination. JAMA Netw Open. 2023;6:e2346721. Erratum in: JAMA Netw Open. 2024;7:e240194.
19. Romano MF, Shih LC, Paschalidis IC, Au R, Kolachalama VB. Large language models in neurology research and future practice. Neurology. 2023;101:1058-67.
20. Andrade NS, Donty S. Comparison of large language models' performance on neurosurgical board examination questions. medRxiv. 2025. [Preprint].
21. Shu E, Sharma A, Nanda P, Kothari S, Wang T. Large language model performance in neurology board questions (S33.001). Neurology. 2024;102(Suppl 17):S33.001.
22. Tarabanis C, Zahid S, Mamalis M, Zhang K, Kalampokis E, Jankelson L. Performance of publicly available large language models on internal medicine board-style questions. PLOS Digit Health. 2024;3:e0000604.