ORIGINAL ARTICLE
174

J Turk Spinal Surg 2025;36(4):174-180
DOI: 10.4274/jtss.galenos.2025.74436

ASSESSING THE ADEQUACY OF ARTIFICIAL INTELLIGENCE MODELS IN ANSWERING SPINE SURGERY QUESTIONS FROM THE ORTHOPEDIC RESIDENCY TRAINING AND DEVELOPMENT EXAMINATION

Bilge Kağan Yılmaz[1], Uğur Yüzügüldü[2]

[1]*Afyonkarahisar Health Sciences University Faculty of Medicine, Department of Orthopaedic and Traumatology, Afyonkarahisar, Türkiye*

[2]*Balıkesir Atatürk City Hospital, Clinic of Orthopedics and Traumatology, Balıkesir, Türkiye*

ABSTRACT

Objective: Artificial intelligence (AI) has undergone remarkable advancements in recent years, and its integration across various domains has been transformative. In the field of medicine, AI applications are rapidly expanding, offering novel opportunities for clinical practice, decision-making, and medical education. The present study sought to assess the performance and reliability of state-of-the-art AI models in addressing spine surgery questions from the Orthopedic Residency Training and Development Examination conducted in Türkiye between 2010 and 2023.

Materials and Methods: A total of 286 spine surgery questions were systematically analyzed. The reference standard was established using the official correct answers, which were subsequently compared with the outputs generated by three advanced AI models: Chat Generative Pre-trained Transformer-5.0 (ChatGPT-5.0), Gemini-Pro, and DeepSeek-V3. Model performance was evaluated in terms of accuracy, error rate, and non-response rate. Comparative analyses among models were performed using chi-square and McNemar tests with pairwise post-hoc comparisons. Wilson's method was employed to calculate 95% confidence intervals (CIs). In addition, subgroup analyses were conducted according to question categories and temporal strata.

Results: Gemini-Pro achieved the highest accuracy rate (85.3%), demonstrating statistically significant superiority over ChatGPT-5.0 (71.7%, p<0.001). The overall accuracy rates were as follows: Gemini-Pro, 85.3% (95% CI: 80.7-88.9; non-response 1.4%); DeepSeek-V3, 78.0% (95% CI: 72.8-82.4; non-response 3.8%); and ChatGPT-5.0, 71.7% (95% CI: 66.2-76.6; non-response 10.8%). Temporal analyses revealed that Gemini-Pro and DeepSeek-V3 performed better in earlier years, whereas Gemini-Pro consistently maintained superior and stable performance in the later periods. In contrast, ChatGPT-5.0 exhibited persistently lower accuracy across all intervals.

Conclusion: Gemini-Pro demonstrated the most consistent and robust performance across both overall and temporal analyses. These findings underscore the promising role of AI in orthopedic residency education, particularly in examination preparation. Nevertheless, integration of AI into training curricula should be approached with caution, as expert oversight remains indispensable to ensure reliability and clinical applicability.

Keywords: Artificial intelligence, ChatGPT, Gemini, DeepSeek, spine surgery

INTRODUCTION

With rapid technological advancements, the demand for instant and accessible information has increased exponentially across all domains, including healthcare. Artificial intelligence (AI) has driven a transformative shift in medicine, encompassing applications in diagnosis, surgical planning, and medical education[1]. In high-risk surgical specialties such as orthopedics and spine surgery, AI-assisted tools are increasingly utilized for radiographic interpretation, clinical decision support, and simulation-based training. Chat Generative Pre-trained Transformer (ChatGPT) has demonstrated utility in the medical field through its ability to perform case-based analyses, making it particularly valuable for literature synthesis and clinical evaluation. Its strengths lie in analyzing complex clinical cases and contributing to academic assessments[2]. Another AI model,

Address for Correspondence: Bilge Kağan Yılmaz, Afyonkarahisar Health Sciences University Faculty of Medicine, Department of Orthopaedic and Traumatology, Afyonkarahisar, Türkiye
E-mail: yilmazbk@gmail.com
ORCID ID: orcid.org/0000-0002-2765-7833
Received: 17.09.2025 **Accepted:** 25.09.2025 **Publication Date:** 16.10.2025
Cite this article as: Yılmaz BK, Yüzügüldü U. Assessing the adequacy of artificial intelligence models in answering spine surgery questions from the orthopedic residency training and development examination. J Turk Spinal Surg. 2025;36(4):174-180

Gemini, distinguishes itself with advanced reasoning capabilities and the capacity to manage complex tasks. Consequently, its integration into clinical decision-making processes has been recommended[3,4]. DeepSeek represents another widely implemented AI model. While it has been described as more dynamic and flexible in tracking developments within the medical literature, it has also been noted to lack the capability for image processing[5]. The most recent version, DeepSeek-V3, further introduces offline functionality, thereby enhancing data privacy[6]. Furthermore, comparative analyses indicate that while ChatGPT demonstrates superiority in literature synthesis, clinical reasoning, medical education, and patient communication, DeepSeek shows relative strength in surgical education, skill acquisition, patient education, and pre-operative planning[7].

Recent studies have demonstrated that large language models (LLMs) can generate clinically relevant responses to medical questions, thereby highlighting their potential role in postgraduate education and examination preparation[8]. LLMs have shown progressively improved performance on medical licensing and specialty board examinations, underscoring their potential applicability in medical education[9-11]. Prior research revealed that ChatGPT-3.5 achieved borderline-passing performance on the United States Medical Licensing Examination, whereas GPT-4 demonstrated superior outcomes on surgical knowledge assessments[12,13]. More recent reports have begun comparing emerging models such as Gemini and DeepSeek in clinical tasks[14,15].

In Türkiye, the Orthopedic Residency Training and Development Examination (UEGS), administered annually by the Turkish Society of Orthopedics and Traumatology Education Council (TOTEK), serves as a national standardized assessment of theoretical and clinical knowledge among orthopedic residents. The examination encompasses a broad spectrum of subspecialties, including trauma, arthroplasty, sports medicine, pediatric orthopedics, and spine surgery. Among these, spine surgery represents a particularly critical domain due to its technical complexity, steep learning curve, and the necessity for precise anatomical and biomechanical knowledge. Evaluating AI models on standardized board questions provides valuable insights into their capabilities, limitations, and potential integration into orthopedic training. Previous studies in other medical disciplines have explored LLM performance on certification and licensing examinations, reporting variable yet frequently promising levels of accuracy. In Türkiye, several investigations have assessed AI performance on national board examinations prepared by TOTEK, comparing model outputs against residents and/or practicing surgeons[16-18]. However, to date, no study has systematically evaluated AI performance within the context of orthopedic residency training in Türkiye, with a particular focus on the spine surgery subspecialty.

Accordingly, the present study aimed to address this gap by analyzing AI-generated responses to spine surgery questions from the UEGS administered between 2010 and 2025. Specifically, this study sought to (I) determine the adequacy of AI models in assessing spine surgery knowledge, (II) compare performance differences among distinct AI platforms, and (III) discuss the potential implications of AI integration into orthopedic residency education and assessment.

MATERIALS AND METHODS

Study Design

This study was designed as a retrospective, comparative analysis of AI model performance using a standardized national examination dataset. The investigation focused specifically on the spine surgery domain of the UEGS, administered by the TOTEK. The UEGS questions are text-based and do not include figures or tables.

Data Source and Question Selection

All UEGS questions administered between 2010 and 2025 were reviewed. Questions were obtained from official archives and verified resources accessible to orthopedic training programs. From the complete pool, questions pertaining to spine surgery were systematically identified and included. Eligible items covered anatomy, pathology, biomechanics, diagnosis, and the treatment of spinal disorders. Incomplete, or ambiguous questions were excluded. In total, 286 spine surgery questions were incorporated. The correct answer to each question, as provided by the official UEGS answer key, was used as the reference standard (gold standard) for performance evaluation. During the study period, three AI models were tested: ChatGPT-5.0 (OpenAI, San Francisco, CA, USA), Gemini-Pro (Alphabet, Mountain View, CA, USA), and DeepSeek-V3 (DeepSeek AI, Beijing, China). All models were accessed between July and August 2025 via publicly available or application programming interface-based interfaces under standardized conditions.

Testing Procedure

Each question was entered into the respective AI model in its original Turkish form. For models with limited Turkish language capabilities, parallel English translations were also used, and outputs were cross-validated for consistency. AI responses were recorded in a structured format: correct (C), incorrect (I), and no answer/unknown (N). All items were submitted individually to the models, ensuring that no duplicated entries were used. To minimize memory retention bias and potential performance inflation, each question was answered in a new session. Moreover, the entire test was repeated twice at three-day intervals for each model using the same procedure. For analysis, the mean values of responses across different trials were calculated.

Statistical Analysis

All statistical analyses were performed using IBM SPSS Statistics, version 26.0 (IBM Corp., Armonk, NY, USA). AI outputs were compared with the official answer key. Performance metrics were defined as follows: accuracy (%) = number of correct responses/total number of questions; error rate (%) = number of incorrect responses/total number of questions; [non-response rate (NR) %] = number of "n" responses/total number of questions. Comparative analyses across AI models were performed using the chi-square test for categorical outcomes. A p-value <0.05 was considered statistically significant. Subgroup analyses were additionally performed according to time intervals (2010-2015, 2016-2020, 2021-2025) and question categories (trauma, degenerative spine, deformity, oncology, infection, and general knowledge).

Ethical Approval

The study protocol was reviewed and approved by the Non-Interventional Clinical Research Ethics Committee of Afyonkarahisar Health Sciences University (approval number: 2025/11, date: 05.09.2025).

RESULTS

A total of 286 spine surgery questions from the UEGS were analyzed to determine accuracy, error, and NRs. Gemini-Pro achieved the highest accuracy (85.3%), demonstrating significantly superior performance compared with both ChatGPT-5.0 (71.7%) and DeepSeek-V3 (78.0%). The overall chi-square test indicated significant differences among the models (p<0.001). Pairwise comparisons revealed that the difference between ChatGPT-5.0 and Gemini-Pro was statistically significant (p<0.001), whereas no significant differences were observed for the other model pairs. NRs were generally low across all models, with Gemini-Pro yielding the lowest proportion of unanswered items. The performance metrics of each AI model are summarized in Table 1.

Temporal Analyses

Accuracy rates demonstrated variability across time intervals.
2010-2015: ChatGPT-5.0: 65.2% [95% confidence interval (CI): 55.1-74.2; NR: 16.3%); Gemini-Pro: 79.3% (95% CI: 70.0-86.4; NR: 2.2%); DeepSeek-V3: 79.3% (95% CI: 70.0-86.4; NR: 3.3%). Pairwise McNemar tests: ChatGPT-5.0 vs. DeepSeek-V3, p=0.0106; ChatGPT-5.0 vs. Gemini-Pro, p=0.0241; Gemini-Pro vs. DeepSeek-V3, p=1.0000.
2016-2020: ChatGPT-5.0: 74.7% (95% CI: 64.7-82.7; NR: 12.6%); Gemini-Pro: 89.7% (95% CI: 81.5-94.5; NR: 0.0%); DeepSeek-V3: 77.0% (95% CI: 67.1-84.6; NR: 3.4%). Pairwise McNemar tests: ChatGPT-5.0 vs. Gemini-Pro, p=0.0044; Gemini-Pro vs. DeepSeek-V3, p=0.0074; ChatGPT-5.0 vs. DeepSeek-V3, p=0.8318.
2021-2025: ChatGPT-5.0: 74.8% (95% CI: 65.8-82.0; NR: 4.7%); Gemini-Pro: 86.9% (95% CI: 79.2-92.0; NR: 1.9%); DeepSeek-V3: 77.6% (95% CI: 68.8-84.4; NR: 4.7%). Pairwise McNemar tests: ChatGPT-5.0 vs. Gemini-Pro, p=0.0146; Gemini-Pro vs. DeepSeek-V3, p=0.0525; ChatGPT-5.0 vs. DeepSeek-V3, p=0.6476.
These findings indicate that Gemini-Pro and DeepSeek-V3 outperformed ChatGPT-5.0 in the earlier period (2010-2015), while Gemini-Pro consistently demonstrated superior and more stable performance in subsequent years. The temporal performance trends are illustrated in Figure 1, with detailed results presented in Table 2.

Subgroup Analyses by Question Category

Subgroup analyses were conducted across six domains of spine surgery. DeepSeek-V3 achieved the highest accuracy in oncology questions, whereas Gemini-Pro outperformed the other models across all remaining categories. Specifically:
Trauma (n=42): Gemini-Pro, 83.0% (95% CI: 69.9-91.1)
Degenerative spine (n=56): Gemini-Pro, 87.5% (95% CI: 76.4-93.8)
Deformity (n=87): Gemini-Pro, 85.1% (95% CI: 76.1-91.1)
Oncology (n=42): DeepSeek-V3, 92.3% (95% CI: 66.7-98.6)
Infection (n=21): Gemini-Pro, 81.0% (95% CI: 60.0-92.3)
General knowledge (n=62): Gemini-Pro, 87.1% (95% CI: 76.6-93.3)
A comprehensive summary of category-specific performances is provided in Table 3.

DISCUSSION

This study represents the first systematic evaluation of AI model performance on spine surgery questions from the UEGS, a standardized national examination in Türkiye. The findings demonstrate that Gemini-Pro achieved a notably higher accuracy rate compared with ChatGPT-5.0 and DeepSeek-V3, suggesting that advanced LLMs may serve as a complementary tool in orthopedic education.
Across the complete dataset of 286 spine surgery questions, Gemini-Pro consistently outperformed the other models, attaining both the highest accuracy and the lowest NR.

Table 1. Accuracy, error, and non-response rates of AI models on spine surgery questions from the UEGS between 2010 and 2025

Model	Total (n)	Correct (n)	Correct (%)	Incorrect (n)	Incorrect (%)	Uncertain (n)	Uncertain (%)
ChatGPT-5	286	205	71.7	50	17.5	31	10.8
Gemini	286	244	85.3	38	13.3	4	1.4
DeepSeek	286	223	78.0	52	18.2	11	3.8

AI: Artificial intelligence, UEGS: Orthopedic residency training and development examination

These results are consistent with the growing body of literature demonstrating that LLMs are approaching passing-level performance on high-stakes examinations and surgical knowledge assessments[12,13]. Global reviews of exam performance have further underscored substantial heterogeneity among model families[5], and emerging reports suggest that DeepSeek may achieve performance comparable to other systems in certain clinical decision-support tasks[14,15]. In the present study, the 71.7% accuracy of ChatGPT-5.0 aligns with findings from other disciplines evaluating LLM performance on specialty board examinations[19]. Gemini-Pro's higher accuracy and DeepSeek-V3's acceptable, albeit lower, accuracy rates reflect the performance variability across AI architectures, in line with previous reports[20].

Several prior studies have assessed AI performance on Turkish orthopedic examinations. Yağar et al. [21] reported that ChatGPT-4o performed favorably on the Turkish Orthopedics and Traumatology Board Examination, particularly in basic science questions. Pamuk et al. [16] found that ChatGPT not only performed with high accuracy but also surpassed the majority of human examinees, outperforming 98.7% of candidates. Conversely, Yiğitbay[18] observed relatively limited performance of ChatGPT in the same context. Ayik et al. [22] compared multiple models and showed that ChatGPT-4 achieved the highest accuracy compared with ChatGPT-3.5 and Gemini on Turkish orthopedic progress examinations. Similarly, Lum[23] reported that ChatGPT exhibited low likelihood of success in the American Board of Orthopaedic Surgery Examination when benchmarked against residents.

Beyond examination settings, AI tools have also been investigated in clinical contexts. Demir and Kültür[24] compared ChatGPT-4o, DeepSeek-V3, and Gemini-Pro with orthopedic surgeons in patient assessment and decision-making, reporting that AI systems performed significantly worse on case-based scenarios but demonstrated comparable accuracy on knowledge-based questions. Karapınar et al. [17] specifically examined spine-related questions from orthopedic residency examinations and found that ChatGPT-3.5 and ChatGPT-4.0 provided answers equivalent to the knowledge level of a third-year resident.

The low NRs observed across all models suggest a general tendency to provide definitive answers. However, the presence of incorrect responses highlights the risk of misleading outputs. Thus, while AI tools may provide valuable support in exam preparation, interpretation of results should remain under expert supervision.

From an educational perspective, the integration of AI-based platforms into residency curricula could foster self-directed learning, enable immediate feedback, and promote standardization in exam preparation. Future investigations should incorporate larger datasets, extend analyses across different subspecialties, and explore interactive, real-time

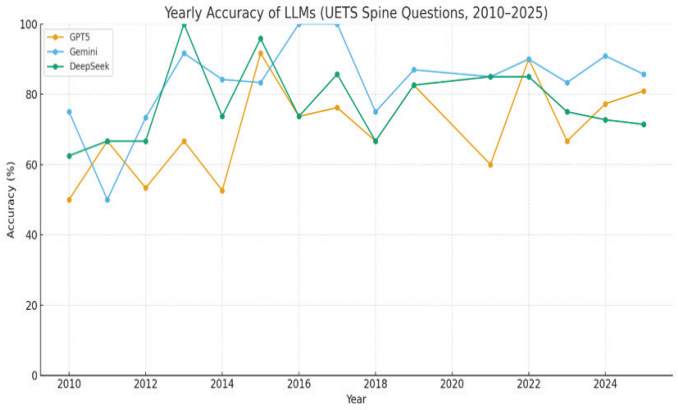

Figure 1. Yearly accuracy of LLMs (UETS Spine Questions, 2010-2025). LLMs: Large language models, UETS: Unified European Training Syllabus

Table 2. Binary McNemar comparisons by time periods

Period	Comparison	A wrong/B right (b01)	A right/B wrong (b10)	Discordant (n)	McNemar p-value (exact)
2010-2025 (Overall)	GPT5 vs DeepSeek	41	23	64	**0.0328**
2010-2025 (Overall)	GPT5 vs Gemini	56	17	73	**<0.0001**
2010-2025 (Overall)	Gemini vs DeepSeek	16	37	53	**0.0055**
2010-2015	GPT5 vs DeepSeek	18	5	23	**0.0106**
2010-2015	GPT5 vs Gemini	21	8	29	**0.0241**
2010-2015	Gemini vs DeepSeek	8	8	16	1.0000
2016-2020	GPT5 vs DeepSeek	12	10	22	0.8318
2016-2020	GPT5 vs Gemini	16	3	19	**0.0044**
2016-2020	Gemini vs DeepSeek	2	13	15	**0.0074**
2021-2025	GPT5 vs DeepSeek	11	8	19	0.6476
2021-2025	GPT5 vs Gemini	19	6	25	**0.0146**
2021-2025	Gemini vs DeepSeek	6	16	22	0.0525

GPT: Generative pre-trained transformer

Table 3. Subgroup analysis by question categories

Subtype/ period	Model	Items (n)	Correct (n)	Accuracy (%)	95% CI (Accuracy)	Unknown (%)	Answered (n)	Coverage-adjusted accuracy (%)	95% CI (CA)	McNemar p-value: GPT-5 vs Gemini	McNemar p-value: GPT-5 vs DeepSeek	McNemar p-value: Gemini vs DeepSeek
Overall (2010-2025)	GPT-5	286	205	71.7%	66.2%-76.6%	10.8%	255	80.4%	75.1%-84.8%	**<0.0001**	**0.0328**	**0.0055**
	Gemini	286	244	85.3%	80.7%-88.9%	1.4%	282	86.5%	82.0%-90.0%			
	DeepSeek	286	223	78.0%	72.8%-82.4%	3.8%	275	81.1%	76.0%-85.3%			
Trauma	GPT-5	47	33	70.2%	56.0%-81.3%	8.5%	43	76.7%	62.3%-86.8%	0.1460	0.7539	0.2891
	Gemini	47	39	83.0%	69.9%-91.1%	0.0%	47	83.0%	69.9%-91.1%			
	DeepSeek	47	35	74.5%	60.5%-84.7%	2.1%	46	76.1%	62.1%-86.1%			
Degenerative	GPT-5	56	41	73.2%	60.4%-83.0%	8.9%	51	80.4%	67.5%-89.0%	**0.0386**	0.7905	0.1460
	Gemini	56	49	87.5%	76.4%-93.8%	0.0%	56	87.5%	76.4%-93.8%			
	DeepSeek	56	43	76.8%	64.2%-85.9%	3.6%	54	79.6%	67.1%-88.2%			
Deformity	GPT-5	87	58	66.7%	56.2%-75.7%	13.8%	75	77.3%	66.7%-85.3%	**0.0025**	**0.0072**	0.5488
	Gemini	87	74	85.1%	76.1%-91.1%	2.3%	85	87.1%	78.3%-92.6%			
	DeepSeek	87	71	81.6%	72.2%-88.4%	4.6%	83	85.5%	76.4%-91.5%			
Oncology	GPT-5	13	8	61.5%	35.5%-82.3%	23.1%	10	80.0%	49.0%-94.3%	0.3750	0.1250	1.0000
	Gemini	13	11	84.6%	57.8%-95.7%	0.0%	13	84.6%	57.8%-95.7%			
	DeepSeek	13	12	92.3%	66.7%-98.6%	0.0%	13	92.3%	66.7%-98.6%			
Infection	GPT-5	21	16	76.2%	54.9%-89.4%	4.8%	20	80.0%	58.4%-91.9%	1.0000	1.0000	0.6250
	Gemini	21	17	81.0%	60.0%-92.3%	4.8%	20	85.0%	64.0%-94.8%			
	DeepSeek	21	15	71.4%	50.0%-86.2%	9.5%	19	78.9%	56.7%-91.5%			
General knowledge	GPT-5	62	49	79.0%	67.4%-87.3%	9.7%	56	87.5%	76.4%-93.8%	0.2668	0.7744	0.1435
	Gemini	62	54	87.1%	76.6%-93.3%	1.6%	61	88.5%	78.2%-94.3%			
	DeepSeek	62	47	75.8%	63.8%-84.8%	3.2%	60	78.3%	66.4%-86.9%			

CI: Confidence interval, CA: California, GPT: Generative pre-trained transformer

assessments with residents.

When evaluating AI performance, it is important to consider differences in question formats. Prior studies have demonstrated that the performance of LLMs may vary depending on whether the assessment involves multiple-choice questions (MCQ) or true/false questions. Isleem et al.[25] reported that ChatGPT's accuracy differed according to question type. In our study, the UEGS exam format was limited exclusively to true/false items. While this binary structure simplifies decision-making for AI and may yield higher accuracy compared to more complex MCQ, it simultaneously restricts the depth of reasoning and clinical judgment that can be assessed. Therefore, the findings should be interpreted within the context of this inherent limitation of the exam format.

Study Limitations

The limitations of this study include its retrospective design, lack of qualitative assessment of AI-generated responses, and potential heterogeneity in model versions over the study period. Moreover, given that the study focuses exclusively on spine surgery questions and employs a simple true/false format, the findings may not fully capture the breadth of medical knowledge or the complexity of clinical judgment. These findings establish an important foundation for the integration of AI into orthopedic residency education and underscore the need for multicenter, prospective studies to validate these results.

CONCLUSION

This study demonstrates that AI models can serve as supportive tools in orthopedic residency education and examination preparation. Among the evaluated systems, Gemini-Pro achieved significantly higher accuracy compared with ChatGPT-5.0 and DeepSeek-V3. The observed variability in performance across time underscores the dynamic evolution of AI capabilities. Larger, multicenter studies incorporating broader datasets and interactive educational modules will be essential to fully elucidate the role of AI in orthopedic training.

Ethics

Ethics Committee Approval: The study protocol was reviewed and approved by the Non-Interventional Clinical Research Ethics Committee of Afyonkarahisar Health Sciences University (approval number: 2025/11, date: 05.09.2025).

Informed Consent: This study was designed as a retrospective.

Footnotes

Authorship Contributions

Surgical and Medical Practices: B.K.Y., U.Y., Concept: B.K.Y., U.Y., Design: B.K.Y., U.Y., Data Collection or Processing: B.K.Y., Analysis or Interpretation: B.K.Y., Literature Search: B.K.Y., U.Y., Writing: B.K.Y.

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. Zhou B, Yang G, Shi Z, Ma S. Natural language processing for smart healthcare. IEEE Rev Biomed Eng. 2024;17:4-18.
2. Charles YP, Lamas V, Ntilikina Y. Artificial intelligence and treatment algorithms in spine surgery. Orthop Traumatol Surg Res. 2023;109:103456.
3. Wong CR, Zhu A, Baltzer HL. The accuracy of artificial intelligence models in hand/wrist fracture and dislocation diagnosis: a systematic review and meta-analysis. JBJS Rev. 2024;12.
4. Seth I, Marcaccini G, Lim K, Castrechini M, Cuomo R, Ng SK, et al. Management of dupuytren's disease: a multi-centric comparative analysis between experienced hand surgeons versus artificial intelligence. Diagnostics (Basel). 2025;15:587.
5. Kaygisiz ÖF, Teke MT. Can deepseek and ChatGPT be used in the diagnosis of oral pathologies? BMC Oral Health. 2025;25:638.
6. Temsah A, Alhasan K, Altamimi I, Jamal A, Al-Eyadhy A, Malki KH, et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. Cureus. 2025;17:e79221.
7. Bhattacharya K, Bhattacharya S, Bhattacharya N, Bhattacharya N. DeepSeek versus ChatGPT in surgical practice. Indian J Surg. 2025.
8. Saad A, Iyengar KP, Kurisunkal V, Botchu R. Assessing ChatGPT's ability to pass the FRCS orthopaedic part a exam: a critical analysis. Surgeon. 2023;21:263-6.
9. 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;3:141.
10. Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ. 2024;17:926-31.
11. Zong H, Wu R, Cha J, Wang J, Wu E, Li J, et al. Large language models in worldwide medical exams: platform development and comprehensive analysis. J Med Internet Res. 2024;26:e66114.
12. 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:e0000198.
13. Beaulieu-Jones BR, Berrigan MT, Shah S, Marwaha JS, Lai SL, Brat GA. Evaluating capabilities of large language models: performance of GPT-4 on surgical knowledge assessments. Surgery. 2024;175:936-42.
14. Sandmann S, Hegselmann S, Fujarski M, Bickmann L, Wild B, Eils R, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. Nat Med. 2025;31:2546-9.
15. Tordjman M, Liu Z, Yuce M, Fauveau V, Mei Y, Hadjadj J, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. Nat Med. 2025;31:2550-5.
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. Karapınar SE, Dinçer R, Coşkun HS, Kaya Ö. Who is more successful in a spinal surgery examination? CHATGPT-3.5/4.0 or a resident doctor? J Turk Spinal Surg. 2025;36:88-91.
18. Yigitbay A. Evaluation of ChatGPT's performance in the Turkish board of orthopaedic surgery examination. Med Bull Haseki. 2024;62:243-9.
19. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA. 2023;329:842-4.

20. Sylolypavan A, Sleeman D, Wu H, Sim M. The impact of inconsistent human annotations on AI driven clinical decision making. NPJ Digit Med. 2023;6:26.
21. Yağar H, Gümüşoğlu E, Mert Asfuroğlu Z. Assessing the performance of ChatGPT-4o on the Turkish Orthopedics and Traumatology Board Examination. Jt Dis Relat Surg. 2025;36:304-10.
22. Ayik G, Kolac UC, Aksoy T, Yilmaz A, Sili MV, Tokgozoglu M, et al. Exploring the role of artificial intelligence in Turkish orthopedic progression exams. Acta Orthop Traumatol Turc. 2025;59:18-26.
23. Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery Examination? Orthopaedic residents versus ChatGPT. Clin Orthop Relat Res. 2023;481:1623-30.
24. Demir MT, Kültür Y. A comparative study of orthopedic surgeons and AI models in the clinical evaluation of spinal surgery. J Turk Spinal Surg. 2025;36:125-9.
25. Isleem UN, Zaidat B, Ren R, Geng EA, Burapachaisri A, Tang JE, et al. Can generative artificial intelligence pass the orthopaedic board examination? J Orthop. 2023;53:27-33.