



Original Research

Performance of AI Models vs. Orthopedic Residents in Turkish Specialty Training Development Exams in Orthopedics

Enver Ipek, Yusuf Sulek, Bahadır Balkanlı

Department of Orthopedics, University of Health Sciences Türkiye, Sisli Hamidiye Etfal Training and Research Hospital, Istanbul, Türkiye

Abstract

Objectives: As artificial intelligence (AI) continues to advance, its integration into medical education and clinical decision making has attracted considerable attention. Large language models, such as ChatGPT-4o, Gemini, Bing AI, and DeepSeek, have demonstrated potential in supporting healthcare professionals, particularly in specialty training examinations. However, the extent to which these models can independently match or surpass human performance in specialized medical assessments remains uncertain. This study aimed to systematically compare the performance of these AI models with orthopedic residents in the Specialty Training Development Exams (UEGS) conducted between 2010 and 2021, focusing on their accuracy, depth of explanation, and clinical applicability.

Methods: This retrospective comparative study involved presenting the UEGS questions to ChatGPT-4o, Gemini, Bing AI, and DeepSeek. Orthopedic residents who took the exams during 2010-2021 served as the control group. The responses were evaluated for accuracy, explanatory details, and clinical applicability. Statistical analysis was conducted using SPSS Version 27, with one-way ANOVA and post-hoc tests for performance comparison.

Results: All AI models outperformed orthopedic residents in terms of accuracy. Bing AI demonstrated the highest accuracy rates (64.0% to 93.0%), followed by Gemini (66.0% to 87.0%) and DeepSeek (63.5% to 81.0%). ChatGPT-4o showed the lowest accuracy among AI models (51.0% to 59.5%). Orthopedic residents consistently had the lowest accuracy (43.95% to 53.45%). Bing AI, Gemini, and DeepSeek showed knowledge levels equivalent to over 5 years of medical experience, while ChatGPT-4o ranged from 2-5 years.

Conclusion: This study showed that AI models, especially Bing AI and Gemini, perform at a high level in orthopedic specialty examinations and have potential as educational support tools. However, the lower accuracy of ChatGPT-4o reduced its suitability for assessment. Despite these limitations, AI shows promise in medical education. Future research should focus on improving the reliability, incorporating visual data interpretation, and exploring clinical integration.

Keywords: Artificial intelligence, chatbot orthopedics, orthopedic education, orthopedic exam, orthopedic surgery, traumatology

Please cite this article as "Ipek E, Sulek Y, Balkanlı B. Performance of AI Models vs. Orthopedic Residents in Turkish Specialty Training Development Exams in Orthopedics. Med Bull Sisli Etfal Hosp 2025;59(2):151-155".

Artificial intelligence (AI) is a rapidly developing technology in medicine that is used in many areas, from medical decision support systems to diagnosis.^[1] Specifically, large language models, such as ChatGPT-4o, Gemini,

Bing AI, and DeepSeek, are trained on different types of textual data (publicly available content, licensed materials, academic publications, news, etc.), thereby acquiring a broad range of topics and diverse information. All of these

Address for correspondence: Enver Ipek, MD. Department of Orthopedics, University of Health Sciences Türkiye, Sisli Hamidiye Etfal Training and Research Hospital, Istanbul, Türkiye

Phone: +90 506 929 78 04 **E-mail:** enveriipek88@gmail.com

Submitted Date: March 22, 2025 **Revised Date:** April 11, 2025 **Accepted Date:** April 16, 2025 **Available Online Date:** July 02, 2025

©Copyright 2025 by The Medical Bulletin of Sisli Etfal Hospital - Available online at www.sislietfaltip.org

OPEN ACCESS This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



models can support healthcare professionals by providing educational assistance and processing medical information to aid clinical decision-making. However, the success of AI models in medical specialty exams and whether they are comparable to human performance remains controversial.^[2-6] However, the degree to which these models can provide clinically relevant, consistent and quality answers in private medical examinations is still questioned.^[7]

Although several studies have evaluated the performance of artificial intelligence in the medical field, there is no comprehensive comparison study in which different AI models are directly subjected to specialty examinations. Existing research has generally been limited to analyzing the performance of a single model in specific exams. The differences between the performance of different AI systems in specialty exams have not been revealed.

This study aimed to evaluate and compare the performance of various AI models (ChatGPT-4o, Gemini, Bing AI, DeepSeek) with orthopedic residents in orthopedic specialty examinations to assess clinical applicability and limitations. We hypothesized that while AI models may show high performance, they are not yet comparable to orthopedic residents in clinical assessment contexts.

Methods

Study Design

This retrospective comparative study was conducted with formal approval from the Turkish Society of Orthopedics and Traumatology (TOTBID) (Document No: 89, Date: 06.03.2025). In this study, the success rates of orthopedic residents who took the UEGS between 2010 and 2021 were compared with the accuracy rates of the answers given by artificial intelligence (AI) models to the same exam questions. This study was conducted in accordance with the principles of the Declaration of Helsinki.

Study Population

The study included four state-of-the-art large language models (LLMs), each representing the most recent and advanced versions available at the time of testing. ChatGPT-4o, developed by OpenAI, was used for the May 2024 release. Gemini Pro, a product of Google, was evaluated using the March 2024 release. Bing AI, integrated into Microsoft Copilot, was tested on the April 2024 release. Finally, DeepSeek, a DeepSeek AI model, was included in the April 2024 release.

Test Procedure and Data Entry

The exam questions were manually entered into the AI models between March 2024 and 2025. Each model was tested over a 48-hour period to minimize temporal

bias due to possible model updates or contextual learning. All models were tested using private/incognito browser sessions or equivalent methods to avoid personalization or prior interaction bias.

Questions were sent to the models via their own user interfaces, and no application programming interfaces (APIs) were used. Models were accessed using official platforms (e.g., chat.openai.com for ChatGPT-4o and bard.google.com for Gemini) and without plugins, extensions, or web browsing, unless natively supported (e.g., Bing AI).

Input Standardization

Two orthopedic research assistants, each with over five years of clinical and academic experience in orthopedics, independently submitted exam questions to each AI model. To ensure consistency and reduce potential bias, a standardized input protocol was developed and strictly followed throughout the testing process. This protocol involved phrasing each question exactly as it appeared on the original exam; providing no additional context, explanations, or clues; and recording all responses immediately for documentation purposes. Before the main testing phase, both individuals conducted a pilot test with 20 questions to validate the protocol. The responses generated by the AI models were cross-validated, and no significant differences were observed between the two users, thereby confirming the reliability and consistency of the input method.

Data Source and Exclusion Criteria

The data source consists of official UEGS question sets from 2010 to 2021. Each exam from 2010 to 2012 contained 100 questions, while exams from 2013 to 2021 included 200 questions. However, questions from 2020 were excluded because of incomplete archival data and inconsistencies in the official answer key for that year. Additionally, questions that were vague, image-based, or required visual interpretation were excluded, as they could not be processed by the AI models in their current state.

Answer Evaluation

The responses generated by the AI models were independently reviewed by two orthopedic surgeons and classified as correct or incorrect according to the official UEGS answer keys. Only questions with a consistent grading by both reviewers were included in the final dataset.

Statistical Methods

Statistical analysis was performed using the SPSS Version 27 (IBM, NY, USA) software. Descriptive statistics were calculated to summarize the data, and the results were presented as mean±standard deviation and percentage distributions. The conformity of the dataset to a normal distribution was

evaluated using the Shapiro-Wilk test. One-way analysis of variance (ANOVA) was applied to statistically compare the performances of the artificial intelligence models. In cases where significant differences were detected, post hoc analyses were performed.

Results

According to the study population and demographic data, the ChatGPT-4o, Gemini, Bing AI, and DeepSeek artificial intelligence models were analyzed with the candidates who participated in the Speciality Education Development Exams (UEGS) between 2010-2021. Based on accuracy percentages, performance comparison showed that Bing AI had the highest correct answer percentage in all years, reaching accuracy rates ranging from 64.0% to 93.0% (Table 1). Gemini showed the highest performance after Bing AI with success rates ranging from 66.0% to 87.0%. DeepSeek performed moderately, with accuracy rates between 63.5% and 81.0%, whereas ChatGPT-4o had the lowest correct answer rates, ranging from 51.0% to 59.5%. The UEGS participants had the lowest percentage of correct answers in all years, ranging from 43.95% to 53.45% ($p < 0.001$) (Fig 1).

In terms of equivalent experience levels, Bing AI, Gemini, and DeepSeek were found to have knowledge levels equivalent to more than five years of experience (Table 2). ChatGPT-4o, on the other hand, showed experience levels ranging between 2-5 years. In the statistical analysis, significant differences were found between the AI models in terms of accuracy rates using one-way analysis of variance ($p < 0.001$). According to the Tukey post-hoc test results, the Bing AI model performed significantly better than the other models did. As a result, Bing AI and Gemini were the best-performing models with high accuracy rates and experience levels.

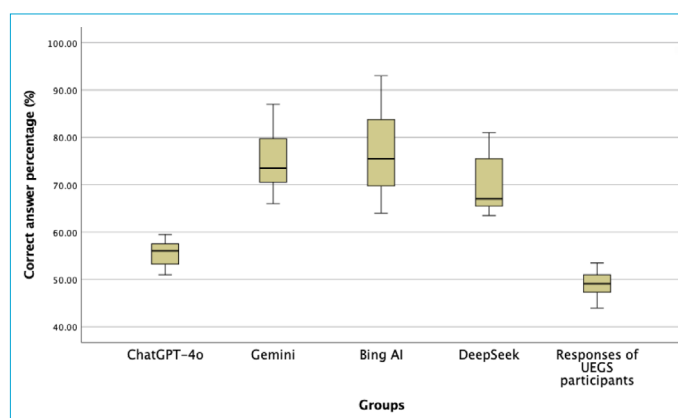


Figure 1. Correct Answer Percentages (%) by Group: ChatGPT-4o, Gemini, Bing AI, DeepSeek, and UEGS Participants.

Discussion

To the best of our knowledge, this is the first study to directly analyze the performance of the UEGS by comparing multiple AI models with real participants. The most important finding of this study shows that ChatGPT-4o, Gemini, Bing AI and DeepSeek AI models achieve higher accuracy rates in UEGS than real participants. However, significant differences were found between the correct answer percentages of the models with Bing AI, which consistently demonstrated the highest accuracy across all years. These findings support the competence of AI models alone in medical specialty exams but confirm our hypothesis that this competence may be limited.

The performance of the artificial intelligence models is consistent with that of other studies in the literature. It was reported that ChatGPT achieved a 47% accuracy rate in orthopedic examinations. However, compared with 5th-year orthopedic residents, it could not enter the top 10 percentile and, therefore, could not pass the exam.^[5] In another

Table 1. Yearly Accuracy Rates (%) of AI Models and Orthopedic Residents in UEGS

Year	Correct answer percentage %					p
	ChatGPT-4o	Gemini	Bing AI	DeepSeek	UEGS participants	
2010	57.0	87.0	64.0	72.0	51.90	p<0.001
2011	52.0	78.0	93.0	81.0	50.10	
2012	51.0	69.0	70.0	65.0	46.80	
2013	56.0	66.0	82.0	69.0	49.10	
2014	56.0	72.0	77.0	66.0	45.05	
2015	54.0	66.0	69.0	64.0	43.95	
2016	52.5	73.5	69.5	67.0	47.85	
2017	55.5	78.5	75.0	63.5	52.00	
2018	59.5	73.0	75.5	67.0	49.35	
2019	58.5	81.0	90.0	80.0	48.65	
2021	58.0	83.5	85.5	79.0	53.45	

Table 2. Estimated Clinical Experience Equivalence of AI Models by Year

Year	Equivalent Experience Level (Years)				p
	ChatGPT-4o	Gemini	Bing AI	DeepSeek	
2010	4-5 years	>5 years	>5 years	>5 years	p<0.001
2011	2-5 years	>5 years	>5 years	>5 years	
2012	3-4 years	>5 years	>5 years	>5 years	
2013	>5 years	>5 years	>5 years	>5 years	
2014	>5 years	>5 years	>5 years	>5 years	
2015	>5 years	>5 years	>5 years	>5 years	
2016	4-5 years	>5 years	>5 years	>5 years	
2017	4-5 years	>5 years	>5 years	>5 years	
2018	>5 years	>5 years	>5 years	>5 years	
2019	>5 years	>5 years	>5 years	>5 years	
2020	>5 years	>5 years	>5 years	>5 years	
2021	>5 years	>5 years	>5 years	>5 years	

study, it was emphasized that the GPT-4 performed at a level to pass the American Orthopedic Board Examination with an accuracy rate of 73.6%, but was more successful in questions that did not contain visual data and therefore had limitations in terms of its competence alone in medical specialty examinations.^[3]

It has been reported that ChatGPT performed lower than actual participants in the TOTEK exam, and that accuracy rates decreased over time.^[8] Similarly, it was reported that ChatGPT-4 outperformed ChatGPT-3.5 with an accuracy rate of 47.2% but performed significantly lower than orthopedic residents, and its performance decreased in questions requiring visual information.^[4] It was reported that ChatGPT performed lower in the orthopedic exam compared to senior residents, but the level of consistency and logic was high.^[9] The study by Guerra et al. determined that ChatGPT, Bard (Gemini), and Bing Chat are AI models that respond with accuracy comparable to that of first-year orthopedic residents.^[10]

Gemini and DeepSeek models achieved the highest accuracy rates after Bing AI and were found to have a knowledge level equivalent to more than 5 years of experience. Similar results have been observed in other studies. It was reported that ChatGPT performed lower in the orthopedic exam compared to senior residents, but the level of consistency and logic was high. The low performance of ChatGPT-4o confirmed the limitations previously reported in the literature. One study reported that although ChatGPT could answer theoretical questions correctly, it was limited to questions requiring interpretation and in multivariate situations.^[9]

Compared to the literature, the most important difference in our study is that we directly compared different artificial intelligence models and analyzed their performance in detail. Other studies have generally considered the perfor-

mance of a single model and neglected the comparisons between different models. For example, studies that only evaluated the performance of ChatGPT in the TOTEK exam did not reveal how the model was compared with other artificial intelligence systems.

This study had several limitations. First, all exam questions were presented to the AI models in a text-only format, excluding image-based or visually dependent questions commonly encountered in orthopedic assessments. This limits the ability to assess the model performance when interpreting radiographs or other visual clinical data. Second, the UEGS has not been externally validated as a comprehensive tool for assessing clinical competence, which may limit the generalizability of the results. Finally, while AI models demonstrate high accuracy, they lack the contextual understanding and clinical reasoning required for nuanced decision making in real-world settings.

Conclusion

This study showed that AI models, especially Bing AI and Gemini, perform at a high level in orthopedic specialty examinations and have potential as educational support tools. However, the lower accuracy of ChatGPT-4o reduced its suitability for assessment. Despite these limitations, AI shows promise in medical education. Future research should focus on improving the reliability, incorporating visual data interpretation, and exploring clinical integration. Future research should focus on improving the reliability of these models and exploring their integration into clinical education and decision making.

Disclosures

Acknowledgments: The authors would like to thank the Turkish Society of Orthopedics and Traumatology (TOTBİD) for providing the necessary permissions and support for this study.

Ethics Committee Approval: This retrospective comparative study was conducted with formal approval from the Turkish Society of Orthopedics and Traumatology (TOTBID) (Document No: 89, Date: 06.03.2025).

Conflict of Interest: The authors declared no conflicts of interest.

Funding: The authors declared that no financial support was received for this study.

Authorship Contributions: Concept – E.I., Y.S., B.B.; Design – E.I., Y.S., B.B.; Supervision – E.I., Y.S., B.B.; Materials – E.I., Y.S., B.B.; Data Collection and/or Processing – E.I., Y.S., B.B.; Analysis and/or Interpretation – E.I., Y.S., B.B.; Literature Review – E.I., Y.S., B.B.; Writing – E.I., Y.S., B.B.; Critical Review – E.I., Y.S., B.B.

Use of AI for Writing Assistance: The authors declared that a large language model (ChatGPT) was used exclusively for the academic language editing of the article. No AI tool were used to generate scientific content, create tables or figures, or perform data analysis or interpretation.

This study directly investigates the performance of artificial intelligence (AI) models (ChatGPT-4o, Gemini, Bing AI and DeepSeek) in the context of orthopedic speciality training examinations. These AI models were used as the subject of analysis by presenting them with real exam questions from the Specialty Education Development Examinations (UEGS) conducted between 2010 and 2021. Their responses were evaluated and statistically compared with those of human participants.

References

1. Gokbulut P, Kuskonmaz SM, Onder CE, Taskaldiran I, Koc G. Evaluation of ChatGPT-4 performance in answering patients' questions about the management of type 2 diabetes. *Med Bull Sisli Etfal Hosp* 2024;58:483–90. [\[Crossref\]](#)
2. Hofmann HL, Guerra GA, Le JL, Wong AM, Hofmann GH, Mayfield CK, et al. The rapid development of artificial intelligence: GPT-4's Performance on Orthopedic surgery board questions. *Orthopedics* 2024;47:e85–9. [\[Crossref\]](#)
3. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB 3rd. Evaluating ChatGPT performance on the orthopaedic in-training examination. *JB JS Open Access* 2023;8:e23.00056. [\[Crossref\]](#)
4. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. *Journal of the American Academy of Orthopaedic Surgeons* 2023;31:1173–9. [\[Crossref\]](#)
5. 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. [\[Crossref\]](#)
6. Lum ZC, Collins DP, Dennison S, Guntupalli L, Choudhary S, Saiz AM, et al. Generative artificial intelligence performs at a second-year orthopedic resident level. *Cureus* 2024;16:e56104. [\[Crossref\]](#)
7. Aljamaan F, Tamsah MH, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. *JMIR Med Inform* 2024;12:e54345. [\[Crossref\]](#)
8. Yigitbay A. Evaluation of ChatGPT's Performance in the Turkish board of orthopaedic surgery examination. *Med Bull Haseki* 2024;62:243–9. [\[Crossref\]](#)
9. Yaş S, Ahmadov A, Baymurat A, Tokgoz MA, Yas SC, Odlyuyurt M et al. ChatGPT vs. orthopedic residents! Who is the winner? *Gazi Med J* 2024;35:186–91. [\[Crossref\]](#)
10. Guerra GA, Hofmann HL, Le JL, Wong AM, Fathi A, Mayfield CK, et al. ChatGPT, bard, and bing chat are large language processing models that answered orthopaedic in-training examination questions with similar accuracy to first-year orthopaedic surgery residents. *Arthroscopy* 2025;41:557–62. [\[Crossref\]](#)