Medical Journal of Bakırköy

Med J Bakirkoy 2026;22(Suppl 1):10-16
DOI: 10.4274/BMJ.galenos.2024.2024.7-12

Research

Evaluation of ChatGPT's Performance in the Turkish Orthopedic Speciality Education Development Examination (UEGS)

ChatGPT'nin Türk Ortopedi Uzmanlık Eğitimi Gelişim Sınavı'ndaki (UEGS) Performansının Değerlendirilmesi

Ahmet Yiğitbay

İzmir Tepecik Training and Research Hospital, Department of Orthopedics and Traumatology, İzmir, Türkiye

ABSTRACT

Objective: Artificial intelligence (AI) technologies are revolutionizing many fields, including health sciences, engineering, and art education. This article aims to evaluate the Speciality Education Development Examination (UEGS) for orthopedics and traumatology residents in Türkiye, in light of Chat Generative Pre-Trained Transformer's (ChatGPT) potential contributions, and to compare ChatGPT's performance with the actual exam results.

Methods: To evaluate ChatGPT's performance, the questions from the Turkish Orthopedics and Traumatology Education Council UEGS from the last three years were entered into the dataset. The results of ChatGPT were classified as correct or incorrect and compared with the actual exam results. The results were analyzed statistically.

Results: Of the 600 questions, 599 were used as data in this study. ChatGPT received 21%, 9.3%, and 5.6% in the 2021, 2022, and 2023 UEGS, respectively. Over the past three years, the average UEGS performance of ChatGPT was 11.9%. No significant difference was found between ChatGPT and actual exam results (p=0.109).

Conclusion: This study showed that ChatGPT's exam performance exceeded that of assistants with first-year residents and was comparable to that of assistants with second-year residents. The findings highlight the importance of continuously updating and customizing these tools while exploring the potential of AI technologies in education.

Keywords: ChatGPT, general orthopedics, Turkish orthopedics exam, UEGS

ÖZ

Amaç: Yapay zeka (YZ) teknolojileri, sağlık bilimlerinden mühendisliğe ve sanat eğitimine kadar birçok alanda devrim yaratmaktadır. Bu makale, Türkiye'de ortopedi ve travmatoloji uzmanlık eğitimi kapsamında uygulanan Uzmanlık Eğitimi Gelişim Sınavı'nın (UEGS) değerlendirilmesinde üretken önceden eğitilmiş dönüştürücü ChatGPT'nin potansiyel katkılarını incelemeyi ve performansını gerçek sınav sonuçlarıyla karşılaştırmayı amaçlamaktadır.

Gereç ve Yöntem: ChatGPT'nin performansını değerlendirmek için, son üç yılın Türk Ortopedi ve Travmatoloji Eğitim Konseyi UEGS soruları veri olarak kullanıldı. ChatGPT'nin sonuçları doğru ve yanlış olmak üzere iki parametre altında değerlendirildi ve gerçek sınav sonuçlarıyla karşılaştırıldı. Sonuçlar istatistiksel olarak analiz edildi.

Bulgular: Toplamda 600 soru bulunmaktaydı, bunlardan 599'u bu çalışmada veri olarak kullanıldı. ChatGPT, 2021, 2022 ve 2023 UEGS'lerinden sırasıyla %21, %9,3 ve %5,6 puan aldı. Son üç yılın ortalama UEGS performansı %11,9 olarak belirlendi. ChatGPT sonuçları ve gerçek sınav sonuçları istatistiksel olarak analiz edildiğinde, anlamlı bir fark bulunmadı (p=0,109).

Sonuç: Bu çalışma, ChatGPT'nin sınav performansının, 1. yıl asistanlarının sınav sonuçlarından daha yüksek ve 2. yıl asistanlarının sınav sonuçlarına benzer olduğunu göstermiştir. Bulgular, bu araçların sürekli olarak güncellenmesinin ve özelleştirilmesinin önemini vurgularken, eğitimde YZ teknolojilerinin potansiyelini keşfetmenin önemini de ortaya koymaktadır.

Anahtar Kelimeler: ChatGPT, genel ortopedi, Türk ortopedi sınavı, UEGS

Address for Correspondence: Ahmet Yiğitbay, MD, İzmir Tepecik Training and Research Hospital, Department of Orthopedics and Traumatology, İzmir, Türkiye
E-mail: ahmetyigitbay@gmail.com **ORCID ID:** orcid.org/0000-0002-7845-1974

Cite as: Yiğitbay A. Evaluation of ChatGPT's performance in the Turkish Orthopedic Speciality Education Development Examination (UEGS). Med J Bakirkoy. 2026;22(Suppl 1):10-16

Received: 31.07.2024
Accepted: 04.10.2024
Publication Date: 25.09.2026

INTRODUCTION

Artificial intelligence (AI) technologies are revolutionizing many fields, from health sciences to engineering and art education. The opportunities offered by these innovative technologies have significant potential, especially in transforming education and training methods. While medical education constitutes one of the most critical stages of this transformation, the effectiveness of educational processes in specialties such as orthopedics and traumatology plays a crucial role in training future health professionals. Specialty training for orthopedics and traumatology residents in Türkiye requires an intensive, comprehensive curriculum, and successful management of this process directly affects both the quality of education and the standards of patient care. In this context, the Specialty Education Development Examination (UEGS), developed by the Turkish Orthopedics and Traumatology Education Council (TOTEK), is an essential tool for assessing residents' knowledge and skills and for measuring the effectiveness of training programs. However, the objectivity and comprehensiveness of such assessments and the effectiveness of feedback processes have always been subject to critical scrutiny. AI technologies, especially Chat Generative Pre-Trained Transformer (ChatGPT), are prominent in natural language processing and provide a novel perspective on these evaluation processes (1-4).

UEGS has been conducted by the TOTEK within the Turkish Orthopedics and Traumatology Association (TOTBID) since 2010 to evaluate the personal and institutional development of resident physicians trained in orthopedics and traumatology. The true-false question format is used instead of the multiple-choice question format. To ensure that test takers are more careful in selecting answers, each wrong answer cancels one correct answer (5).

ChatGPT's ability to analyze complex data sets, personalize learning materials, and create interactive learning experiences can make valuable contributions to existing methods in medical education. This article aims to evaluate the UEGS organized for orthopedics and traumatology residents in Türkiye in light of ChatGPT's potential contributions and to compare ChatGPT's performance with the actual exam results. It will analyse how ChatGPT's ability to comprehend and summarise the training materials and to answer questions correlates with the exam results.

METHODS

This study aimed to compare ChatGPT's performance with the results of the UEGS taken by orthopedics and traumatology residents in Türkiye. The study consists of two main phases: first, the accuracy of ChatGPT's answers to the UEGS sample questions was evaluated; second, these answers were compared with the actual exam results.

Data Collection

The UEGS questions published by TOTEK and the actual exam results formed the study's primary dataset. These questions covered a variety of clinical scenarios and theoretical knowledge. During interactions with ChatGPT, the model was asked to answer these questions. ChatGPT's first response to each question was recorded. Sample answers given by ChatGPT are shown in Figures 1 and 2.

Method of Analysis

The answers given by ChatGPT were evaluated by comparing them with the correct answers to the questions. For each question, the model's answer was scored for accuracy, and an overall success rate was calculated by summing these scores. This method allowed us to determine the extent to which ChatGPT could correctly answer the UEGS exam questions.

Comparison

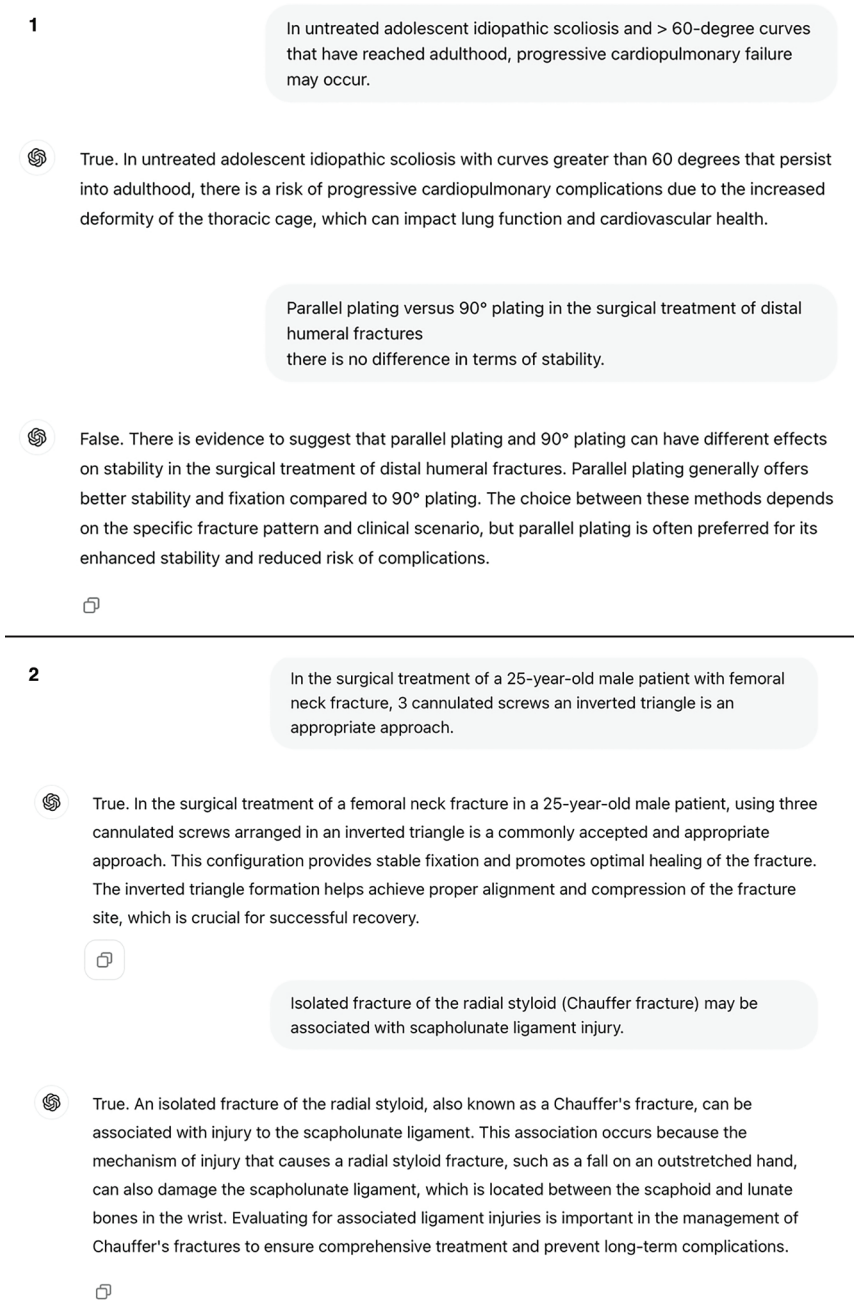
The actual UEGS results of the resident physicians were obtained from the anonymized dataset provided by TOTEK. The model's performance was evaluated by comparing ChatGPT's overall success rate to these accurate exam results. In addition, the model's performance for each question type was analyzed separately, thereby identifying the model's strengths and weaknesses in specific subject areas.

Evaluation

Evaluation of the results yielded inferences regarding how ChatGPT's UEGS performance compares with that of actual physicians. These conclusions are intended to provide information about the applicability and potential of ChatGPT in the field. Versions of ChatGPT-3.5 and GPT-4 were used in this study, and average values were calculated. The answers given by ChatGPT were evaluated in two groups, and the variables were statistically analyzed. Data were analyzed using a Python software package. $p<0.05$ was considered statistically significant.

Ethical Considerations

All data used in the study were anonymized to exclude personal information. The research was designed and conducted in accordance with ethical guidelines. ChatGPT was incorporated into the educational process under licenses permitting both open-source and commercial use. This study was conducted with the written approval of the TOTBID board of directors, as documented in their official correspondence (approval no: 159, date: 26.04.2024).



Figures 1 and 2. ChatGPT's sample answers to UEGS exam questions
ChatGPT: Chat Generative Pre-Trained Transformer, UEGS: Speciality Education Development Examination

This study was conducted using publicly available examination questions and did not involve human participants or patient data. Therefore, informed consent was not required.

Statistical Analysis

The data were analyzed using the Python software package. The normality of the quantitative data distribution was assessed using the Shapiro-Wilk test and graphical methods. The Mann-Whitney U test was used to compare two independent, non-normally distributed samples. The Pearson chi-square test was used to compare qualitative data. Statistical significance was evaluated at $p<0.05$.

RESULTS

Questions that were canceled or that contained visual images were excluded, leaving 599 of 600 questions for

analysis. ChatGPT scored 21%, 9.3%, and 5.6% in the 2021, 2022, and 2023 TOTEK UEGS, respectively. When the TOTBID period books were examined, the average UEGS results for 2021, 2022, and 2023 were 30.4%, 18.7%, and 22.5%, respectively (6,7). These results indicate that ChatGPT's performance on examinations has declined over time. This decrease may be due to an increase in exam difficulty, changes in assessment criteria, or changes in ChatGPT's performance in specific subjects.

When analyzed by subject area, ChatGPT demonstrated the highest success rates in basic orthopedics and trauma (26.67% each) and the lowest in foot and ankle surgery (3.33%) and spine surgery (13.33%). Further year-by-year analysis revealed performance fluctuations across different topics. For instance, ChatGPT's accuracy in pediatric orthopedics increased from 40% in 2021 to 50% in 2022, but dropped to 20% in 2023, whereas its performance in trauma declined to 4% in 2022 before improving to 32% in 2023. These results indicate that ChatGPT's ability to answer UEGS-style questions varies by subject and exam structure (Figures 3-5).

Figure 6 shows annual changes in the performance of both groups, comparing ChatGPT results with actual exam results. The graph shows how the scores of both groups change from 2021 to 2023. While ChatGPT's scores decrease over time, the actual exam results fluctuate. The graph shows trends and changes in performance over time.

No significant difference was found when ChatGPT and the actual exam results were compared using the Mann-Whitney U test (p=0.109). This indicates that the differences between the results of the two groups may be due to random variation and that neither group performed statistically significantly better or worse.

Analysis of the actual exam results from the last three years in the TOTBID 10th- and 11th-term books shows that the average exam success score for residents with one-year seniority was 11.3%, and for residents with two-year seniority was 12.1% (6,7). The average UEGS performance of ChatGPT for the last three years was 11.9%. When these data are analyzed, they show that ChatGPT's

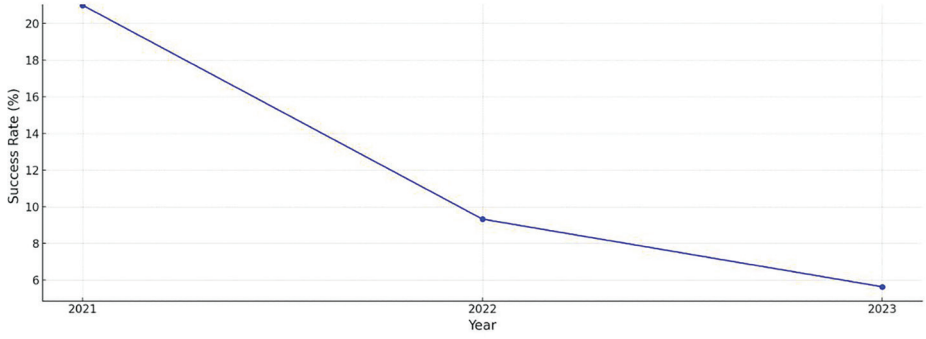


Figure 3. ChatGPT's success graph by years

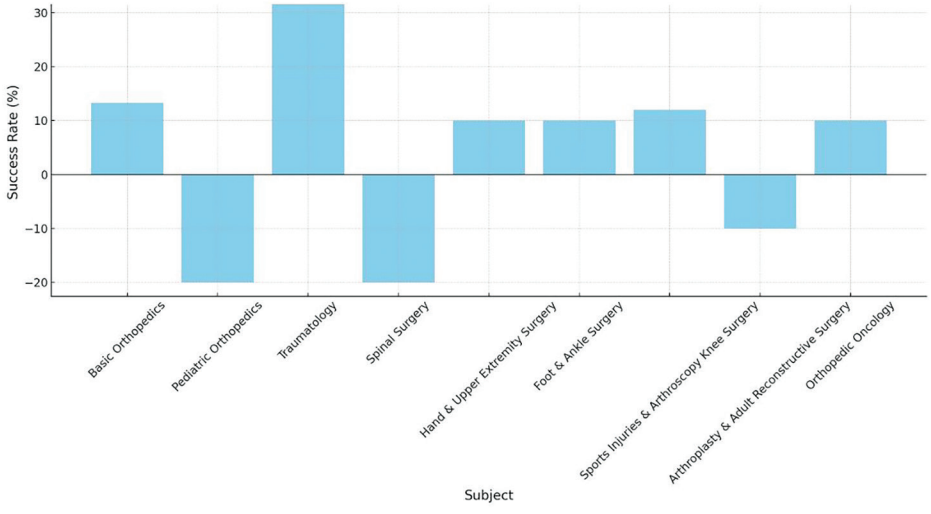


Figure 4. ChatGPT's average success percentage by subjects

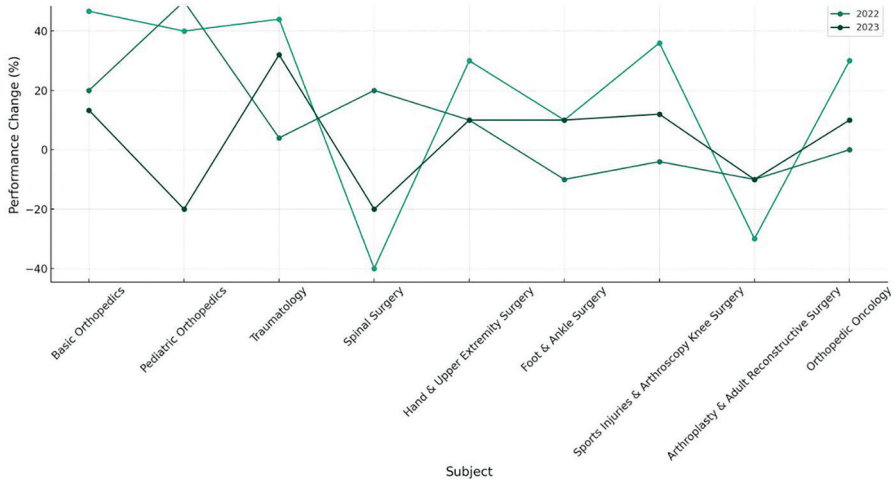


Figure 5. Graph showing the annual subject-based performance change of ChatGPT

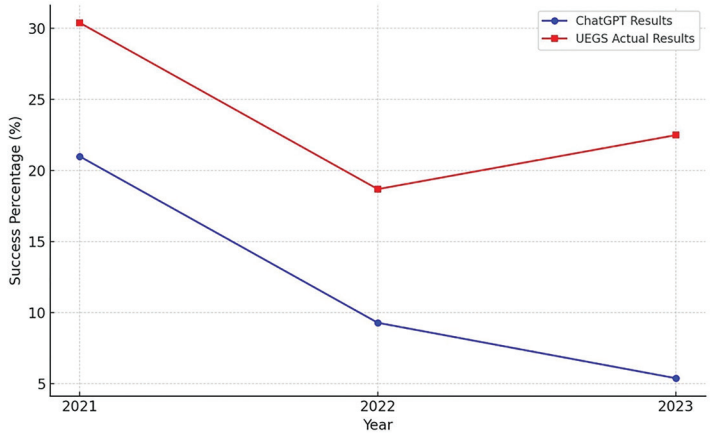


Figure 6. Comparison of "ChatGPT results" and "real exam results" by years
ChatGPT: Chat Generative Pre-Trained Transformer's, UEGS: Speciality Education Development Examination

exam performance is higher than the average results for residents with one year of seniority and comparable to those for residents with two years of seniority.

DISCUSSION

This study analyzed ChatGPT's annual examination results and their relationship to the actual examination results for the corresponding years. Our analyses revealed that ChatGPT's performance varied significantly over the years and in various subject areas. In particular, ChatGPT achieved high accuracy rates in some subjects but performed poorly in others. This suggests that the knowledge and skills of AI-based systems, as human learners do, may vary across subject areas.

In 2021, a moderately positive correlation was observed between ChatGPT and accurate exam results, while in 2022 this relationship was weakly negative. In 2023, the relationship turned positive again, but was less intense than in 2021. These fluctuations may be due to changes in exam content, updates to ChatGPT's training datasets, or specificities of subject areas.

The high success rates of ChatGPT in subjects such as basic orthopedics and trauma indicate that it understands and processes knowledge and terminology in these areas effectively. In contrast, the low performance in other areas, such as spine surgery and paediatric orthopedics, may be due to the complexity of these topics or ChatGPT's insufficient training in them. These findings offer important lessons for integrating AI-based learning and assessment tools into educational processes. Tools such as ChatGPT can enrich students' learning experiences by leveraging their strengths. However, it is essential to be aware of the

limitations of these tools and to integrate guidance from human teachers and other educational materials to provide students with a comprehensive learning experience.

A review of the literature indicates that few studies have evaluated ChatGPT's performance in orthopedics and traumatology examinations. Cuthbert and Simpson (8) evaluated whether ChatGPT could pass part 1 of the Royal College of Surgeons' Fellowship in Trauma and Orthopedic Surgery (FRCS) examination. They stated that ChatGPT could not perform the high-level and multistep logical reasoning required to pass the FRCS examination. Lum (9) evaluated whether ChatGPT could pass the American Board of Orthopedic Surgery Examination and found that ChatGPT was unlikely to pass the exam, but reported that ChatGPT's performance was comparable to that of a first-year orthopedic surgery resident. Saad et al. (10) evaluated ChatGPT's performance on the Orthopedic FRCS (FRCS Orth) part A exam, but reported that ChatGPT did not achieve a passing score. In a multinational study, Alfertshofer et al. (11) manually entered 300 questions from the medical licensing examinations in the United States of America, Italy, France, Spain, the United Kingdom, and India to evaluate the performance of ChatGPT and compare the accuracy of the answers. The study reported significant differences in ChatGPT test accuracy across countries. The highest performance was reported in the Italian exam (73% correct), and the lowest in the French exam (22% correct). ChatGPT's performance on the Examen Nacional de Medicina in Peru was evaluated, and it achieved expert-level performance (12). A study conducted in Japan reported that ChatGPT met the passing standards of the National Medical Licensing Examination (13). ChatGPT's performance on Anesthesiology Board-Style Examination Questions was evaluated in another study, which reported that ChatGPT needed to improve to pass the exam (14).

The use of true-false questions rather than multiple-choice questions in the UEGS exam presents both advantages and disadvantages, which are reflected in ChatGPT's performance. One key advantage of the true-false format is its ability to assess fundamental knowledge efficiently, requiring examinees to make definitive judgments about statements. Additionally, this format discourages random guessing through negative marking. However, a major limitation is its reduced capacity to evaluate complex reasoning and problem-solving skills, which are better assessed by multiple-choice questions that include nuanced answer choices. In our study, ChatGPT's performance varied significantly across different subject areas; its relatively lower performance on the UEGS may be partly attributable to the true-false format, as large language models perform better with contextualized multiple-choice questions. Furthermore, the 50% chance of guessing correctly on true-false questions may have introduced variability in ChatGPT's accuracy. Future evaluations could explore whether a hybrid approach incorporating both question types would enhance the exam's ability to measure a broader range of competencies and to provide more meaningful insights into AI-assisted medical education.

CONCLUSION

ChatGPT's mean reported score on the 2021-2023 UEGS examinations was 11.9%, numerically above the mean for first-year residents (11.3%) and close to that for second-year residents (12.1%). Its scores were lower than the overall resident mean in each examination year, although the reported comparison did not reach statistical significance (p=0.109). This finding should not be interpreted as evidence of equivalent performance.

Performance varied across examination years and subject areas, indicating limitations in the consistency of ChatGPT's responses to orthopedics and traumatology questions. These findings support further investigation of ChatGPT as a supplementary educational resource, but do not establish its effectiveness in improving learning or examination outcomes. Future studies should evaluate individual model versions separately, clearly define scoring procedures, and assess educational outcomes before drawing conclusions about its role in residency training.

ETHICS

Ethics Committee Approval: This study was conducted with the written approval of the TOTBID board of directors, as documented in their official correspondence (approval no: 159, date: 26.04.2024).

Informed Consent: This study was conducted using publicly available examination questions and did not involve human participants or patient data. Therefore, informed consent was not required.

FOOTNOTES

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

Financial Disclosure: The author declares that this study received no financial support.

REFERENCES

1. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595.
2. Sonntagbauer M, Haar M, Kluge S. Künstliche Intelligenz: wie werden ChatGPT und andere KI-anwendungen unseren ärztlichen

alltag verändern? [Artificial intelligence: how will ChatGPT and other AI applications change our everyday medical practice?]. Med Klin Intensivmed Notfmed. 2023;118:366-71. German.

3. Tustumi F, Andreollo NA, Aguilar-Nascimento JE. Future of the language models in healthcare: the role of ChatGPT. Arq Bras Cir Dig. 2023;36:e1727.
4. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ. 2024;17:926-31. Erratum in: Anat Sci Educ. 2024;17:1779.
5. Aydoğdu S. Turkish Orthopaedics and Traumatology Education Council (TOTEK) Fifth Study Period Book. 2009-2011. p. 86-7. Available from: https://totek.totbid.org.tr/uploads/files/totek_5_donem_kitap.pdf
6. Özdemir G, Elçin M. Turkish Orthopaedics and Traumatology Education Council (TOTEK) Tenth Term Book. 2019-2021. p. 16-25. Available from: https://totek.totbid.org.tr/uploads/totek_10.donemkitabi.pdf
7. Yıldız HY. Turkish Orthopaedics and Traumatology Training Council (TOTEK) Eleventh Term Book. 2021-2023. p. 16-37. Available from: https://totek.totbid.org.tr/uploads/11.donemkitabi_totek.pdf
8. Cuthbert R, Simpson AI. Artificial intelligence in orthopaedics: can Chat Generative Pre-trained Transformer (ChatGPT) pass section 1 of the Fellowship of the Royal College of Surgeons (trauma & orthopaedics) examination? Postgrad Med J. 2023;99:1110-4.
9. 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.
10. 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.
11. Alfertshofer M, Hoch CC, Funk PF, Hollmann K, Wollenberg B, Knoedler S, et al. Sailing the Seven Seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. Ann Biomed Eng. 2024;52:1542-5.
12. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, De la Cruz-Galán JP, Gutiérrez-Arratia JD, Quiroga Torres BG, et al. Performance of ChatGPT on the Peruvian National Licensing Medical Examination: cross-sectional study. JMIR Med Educ. 2023;9:e48039.
13. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the National Medical Licensing Examination in Japan: evaluation study. JMIR Form Res. 2023;7:e48023.
14. Khan AA, Yunus R, Sohail M, Rehman TA, Saeed S, Bu Y, et al. Artificial intelligence for Anesthesiology Board-Style Examination questions: role of large language models. J Cardiothorac Vasc Anesth. 2024;38:1251-9.