Original Article

DOI: 10.4274/haseki.galenos.2024.10038
Med Bull Haseki 2024;62:243-249

# Evaluation of ChatGPT's Performance in the Turkish Board of Orthopaedic Surgery Examination

Ahmet Yigitbay

*Siverek State Hospital, Clinic of Orthopedics and Traumatology, Sanliurfa, Turkey*

**Abstract**

**Aim:** Technological advances lead to significant changes in education and evaluation processes in medicine. In particular, artificial intelligence and natural language processing developments offer new opportunities in the health sector. This article evaluates Chat Generative Pre-Trained Transformer's (ChatGPT) performance in the Turkish Orthopaedics and Traumatology Education Council (TOTEK) Qualifying Written Examination and its applicability.

**Methods:** To evaluate ChatGPT's performance, TOTEK Qualifying Written Examination questions from the last five years were entered as data. The results of ChatGPT were assessed under four parameters and compared with the actual exam results. The results were analyzed statistically.

**Results:** Of the 500 questions, 458 were used as data in this study. Chat Generative Pre-Trained Transformer scored 40.2%, 26.3%, 37.3%, 32.9%, and 35.8% in the 2019, 2020, 2021, 2022, and 2023 TOTEK Qualifying Written Examination, respectively. When the correct answer percentages of ChatGPT according to years and the simple linear regression model applied to these data were analyzed, it was determined that there was a slightly decreasing trend in the correct answer rates as the years progressed. ChatGPT's TOTEK Qualifying Written Examination performance showed a statistically significant difference from the actual exam results. It was observed that the correct answer percentage of ChatGPT was below the general average success scores of the exam for each year.

**Conclusions:** This analysis of artificial intelligence's applicability in the field and its role in training processes is essential to assess ChatGPT's potential uses and limitations. Chat Generative Pre-Trained Transformer can be a training tool, especially for knowledge-based and logical questions on specific topics. Still, its current performance is not at a level that can replace human decision-making in specialized medical fields.

**Keywords:** Artificial intelligence, humans, orthopedics, specialty boards

## Introduction

Artificial intelligence (AI) and natural language processing (NLP) technologies drive significant transformations across many fields. The use of these technologies in the healthcare sector has been impactful across a broad spectrum, from medical education to patient care. The primary purpose of AI is to improve patient experience, enhance the reliability of clinicians, and provide more information for the clinical decision-making process. Instead of replacing healthcare workers, these goals aim to enhance their experience (1-4). Language modeling systems like Chat Generative Pre-Trained Transformer (ChatGPT) support health professionals in various areas, from education to clinical applications. Chat Generative Pre-Trained Transformer, developed by OpenAI, is an AI model incorporating several language modeling and comprehension techniques, allowing users to communicate in their native language (5,6).

Technological advancements cause significant changes in education and assessment processes within medicine. Developments in AI and NLP, in particular, are introducing new possibilities in the healthcare sector (7-9). In this context, large language models like ChatGPT can play a significant role in medical education and exam evaluations.

This article aims to assess the performance of ChatGPT in the Turkish Orthopaedics and Traumatology Education Council (TOTEK) Qualification Written Exam and its applicability in the field. The role of ChatGPT in evaluating knowledge and skills in this area and its advantages and

**Address for Correspondence:** Ahmet Yigitbay, Siverek State Hospital, Clinic of Orthopedics and Traumatology, Sanliurfa, Turkey
**Phone:** +90 543 639 61 62 **E-mail:** ahmetyigitbay@gmail.com **ORCID:** orcid.org/0000-0002-7845-1974
**Received:** 28.08.2024 **Accepted:** 30.10.2024

disadvantages compared to traditional exam formats were examined. Additionally, based on practical field experience and feedback from medical experts, the usability of ChatGPT in orthopedics and traumatology education and practice was evaluated. The qualification written exam has been conducted by TOTEK under the Turkish Orthopaedics and Traumatology Association (TOTBID) since 2003, in two stages. The first stage consists of a written exam with "objective structured multiple-choice questions", and the second stage consists of an oral exam that includes an objective structured clinical examination (10,11).

This study is conducted to understand the potential of AI-supported language models in medical education and evaluation processes and to provide a framework for future directions. In the following sections of the article, the performance and applicability of ChatGPT will be analyzed in depth.

## Methods

The primary objective of this study is to evaluate ChatGPT's performance in the TOTEK Qualification Written Exam. To this end, ChatGPT's ability to solve exam questions has been compared with the exam performance of physicians who have previously taken the exam.

### Data Collection

Data from the last five years of TOTEK exam questions was used. Each year's questions were asked individually to ChatGPT, and the answers provided were recorded. To evaluate the performance of real physicians, the average exam results of physicians who had previously taken the exam over the past five years (2019-2023) were used. These data were obtained from the "period books" published by TOTBID, which contain past exam results (12-14).

### Performance Evaluation

Chat Generative Pre-Trained Transformer's performance was measured by its ability to solve questions in the dataset. Chat Generative Pre-Trained Transformer's accuracy rate was used to determine how correctly it answered the exam questions. Only multiple-choice questions containing text were included in the assessment. Due to limitations, questions containing images, tables, figures, and canceled questions were excluded from the evaluation. Only first-stage exam questions were included in the assessment. Questions were asked to ChatGPT only once, and the responses were recorded.

The physicians' performance was determined by taking the average results of physicians who had previously taken the exam.

### Comparison

The performance of ChatGPT and the performance of physicians who have taken the exam have been compared. Differences between the two groups were statistically analyzed, and the results were compared.

### Assessment

The results obtained have been used to compare ChatGPT's performance in the TOTEK Qualification Written Exam with physicians' performance. These inferences have been evaluated to provide information about ChatGPT's applicability and potential in the field. Chat Generative Pre-Trained Transformer-4's version was used in all parts of this project.

The responses given by ChatGPT were evaluated under four categories, and these variables were statistically analyzed.

1: Correct answer, consistent logic
2: Correct answer, inconsistent logic
3: Incorrect answer, consistent logic
4: Incorrect answer, inconsistent logic

### Ethical Considerations

All data used in the study were anonymized to avoid including personal information. The research was designed and conducted according to ethical rules. Chat Generative Pre-Trained Transformer was included in the educational process under licenses suitable for open-source and commercial use. Written permission was obtained from the TOTBID board of directors for this study (document no.: 159, dated: 26.04.2024).

## Results

Of the 500 questions, 458 were used as data in this study. Since two questions were canceled according to the answer keys, these questions were excluded from the study. Chat Generative Pre-Trained Transformer scored 40.2%, 26.3%, 37.3%, 32.9%, and 35.8% in the 2019, 2020, 2021, 2022, and 2023 TOTEK Qualifying Written Examination, respectively (Table 1). 47.2% of the candidates were successful in the TOTEK qualifying exam held in 2023, 37% of the candidates in the exam held in 2022, 55.3% of the candidates in the exam held in 2021, 46.4% of the candidates in the exam held in 2020, and

**Table 1. ChatGPT exam results by year**

| Year | Correct answer percentage % | Wrong answer percentage % |
|---|---|---|
| **2019** | 41.05 | 58.95 |
| **2020** | 26.32 | 73.68 |
| **2021** | 37.35 | 62.65 |
| **2020** | 32.9 | 67.03 |
| **2023** | 36.96 | 63.04 |

ChatGPT: Chat Generative Pre-Trained Transformer

70.5% of the candidates in the exam held in 2019. In the TOTEK qualifying exam, the average success score for 2019 was 60, the average success score for 2020 was 55, the average success score for 2021 was 60, the average success score for 2022 was 60, and finally, the average success score for 2023 was 60 (12-14). Figure 1 presents a comparison between ChatGPT and real exam results.

After analyzing the numerical analysis of the answers given by ChatGPT over the years, the number of answers with correct and consistent logic has remained relatively constant. There are very few answers under the category of the correct answer, and inconsistent logic shows that ChatGPT generally gives logical answers. Although there is an increase in incorrect and consistent logic answers in 2022, the number of ChatGPT's answers with incorrect but consistent logic varies. Wrong answer, inconsistent logic: The area in which ChatGPT struggled the most was the answers with incorrect and inconsistent logic. Especially in 2020, there was a significant increase in the number of such answers (Table 2, Figure 2).

When the percentages of correct answers given by ChatGPT are analyzed by year, fluctuations are observed in its performance over time. The highest percentage of correct answers was achieved in 2019, while the lowest was recorded in 2020. Although ChatGPT's annual performance shows some variation, these fluctuations remain within a relatively limited range. This variability may stem from changes in the datasets used to train the model, updates to the model itself, or differences in the complexity of exam questions across the years.

When the correct and incorrect answer rates change over time and are visualized, differences between both rates are observed in specific years (Figure 3). In particular, the correct answer rate fluctuates over time, while the incorrect answer rate follows a similar pattern. This allows us to understand better the possible effects of yearly changes in ChatGPT's performance and how the

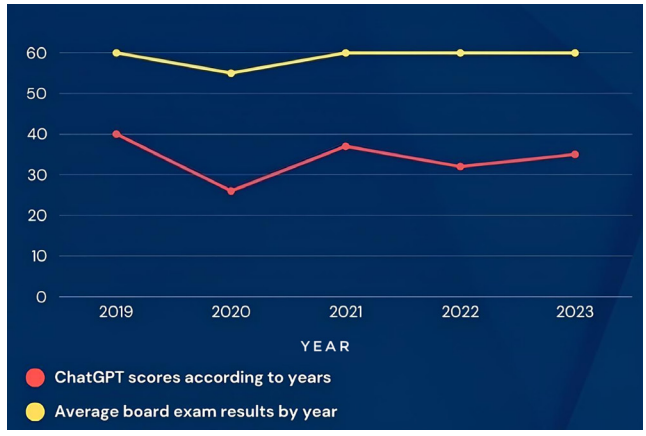


**Figure 1.** Comparison of ChatGPT and real exam results
*ChatGPT: Chat Generative Pre-Trained Transformer*

**Table 2. Numerical analysis of the answers given by ChatGPT by years**

| Answers | 2019 | 2020 | 2021 | 2022 | 2023 |
|---|---|---|---|---|---|
| **Correct answer, consistent logic** | 38 | 25 | 31 | 30 | 33 |
| **Correct answer, inconsistent logic** | 1 | 0 | 0 | 0 | 1 |
| **Incorrect answer, consistent logic** | 7 | 1 | 5 | 27 | 22 |
| **Incorrect answer, inconsistent logic** | 49 | 69 | 47 | 34 | 36 |
| **Question that could not be evaluated** | 3 | 5 | 17 | 9 | 8 |

ChatGPT: Chat Generative Pre-Trained Transformer

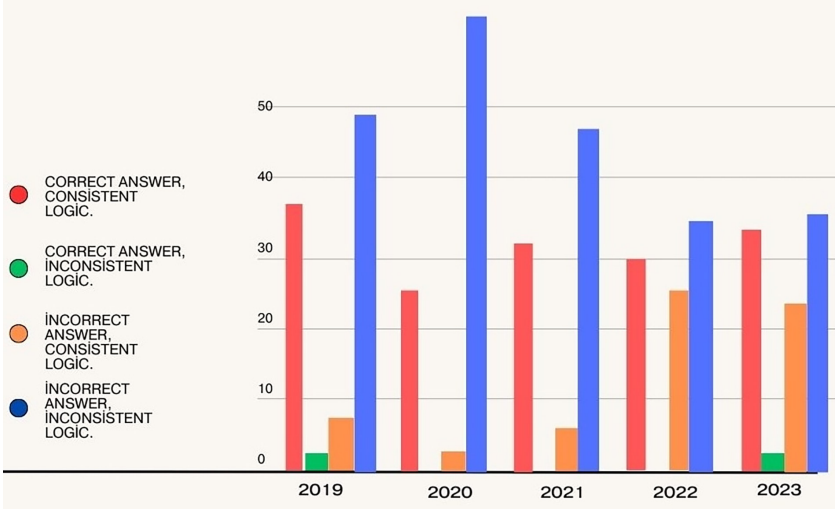


**Figure 2.** Detailed analysis of the answers given by ChatGPT
*ChatGPT: Chat Generative Pre-Trained Transformer*

model responds to certain types of questions in specific years. The Mann-Whitney U test to assess whether the differences between the "correct answer, coherent logic" and "incorrect answer, incoherent logic" categories are statistically significant shows a statistically significant difference between the distributions of the two groups (p=0.032). This result indicates that the medians of the two groups are not the same at the 5% significance level. There is a statistically significant difference between the distributions of the answers in the categories "correct answer, consistent logic" and "incorrect answer, inconsistent logic". This analysis shows that ChatGPT's tendency to give correct answers using consistent logic is statistically significantly different from its tendency to provide incorrect answers using inconsistent logic. These results can be considered when developing strategies to improve ChatGPT's performance and accuracy. For example, the focus could be increasing the correct answer rate by strengthening the model's consistent logic.

When the correct answer percentages by years and the simple linear regression model applied to these data are analyzed, a slightly decreasing trend is observed in the correct answer rates as the years progress (Figure 4). The model's slope is negative, indicating a decrease in the correct answer rates as the years progress. However, due to the low R-square ($R^2$) value, the model only partially explains the variability in the correct answer rates. This indicates that other factors may influence the change over the years ($R^2$ value 0.0366).

The t-statistic obtained from the paired sample t-test between the results of ChatGPT and the actual exam results was calculated as -15.52 and p=0.0001. This

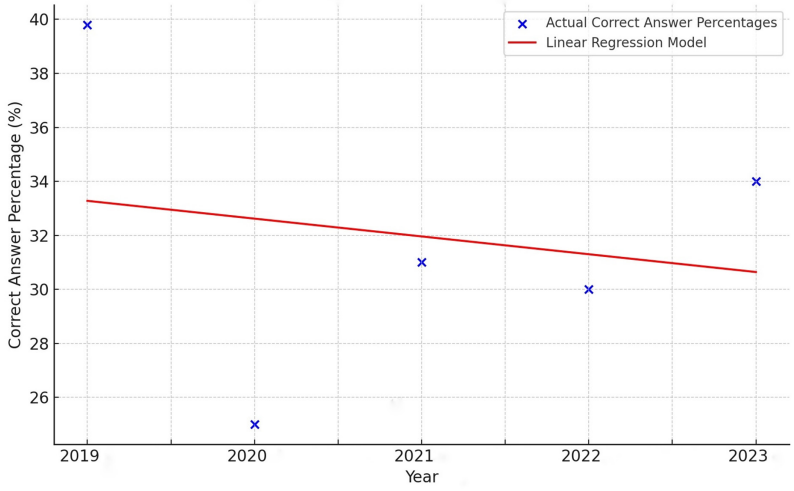


**Figure 3.** Percentage of correct answers by years and linear regression model

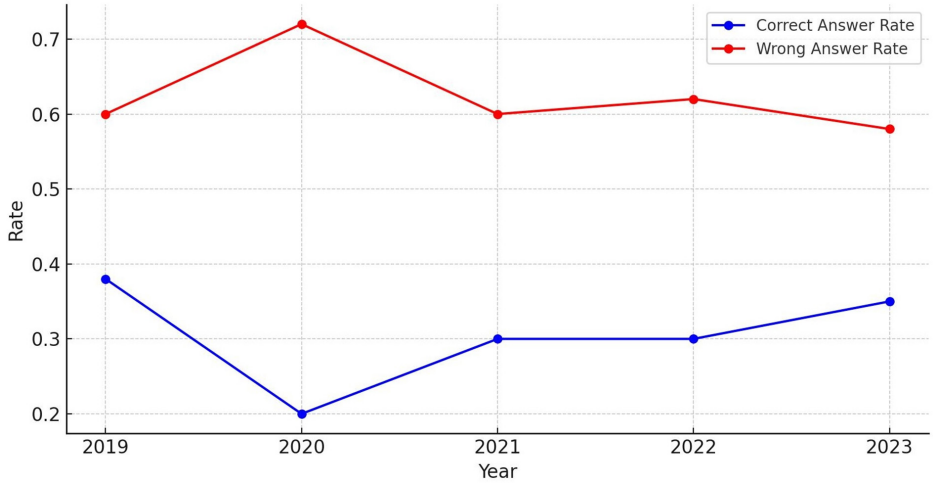


**Figure 4.** Simple linear regression analysis on the percentage of correct answers by year

result shows a statistically significant difference between the results of ChatGPT and the actual exam results at a 5% significance level. The low p-value indicates that this difference is not random and a significant difference exists in the general population. This analysis shows that ChatGPT's performance on the TOTEK Qualifying Written Examination significantly differs from the exam's overall pass rates. Chat Generative Pre-Trained Transformer's scores are below the average passing scores of the actual exam, indicating that the model has limitations in solving such exam questions and that ChatGPT needs to address some aspects of the exam fully. These differences point to potential areas for improvement in ChatGPT training and its ability to understand test questions.

## Discussion

Chat Generative Pre-Trained Transformer in the clinical field has demonstrated its potential to revolutionize healthcare by providing accurate and understandable information on orthopedic issues. Creating interactive quizzes and educational tools for students supports learning and provides instant feedback (15). This AI-powered technology holds great promise for the future of orthopedics, as it is expected to enhance patient care, surgical planning, and medical education (16). Upon evaluating the results of this study, we first observed that ChatGPT's ability to solve exam questions offers certain advantages compared to the performance of actual physicians. Chat Generative Pre-Trained Transformer can effectively address exam questions by quickly accessing and analyzing a vast pool of medical knowledge, which is especially crucial in complex and fast-paced medical scenarios. However, there are areas where improvement is needed in terms of practical applicability and reliability. The lack of citations in the information provided by ChatGPT hinders users from verifying its accuracy, which can limit the use of AI, particularly in healthcare (17).

Nevertheless, some studies indicate that AI can help provide advice and recommendations based on medical history, symptoms, and clinical data (18,19). Chat Generative Pre-Trained Transformer's performance varies depending on specific exam formats and the characteristics of the questions. Additionally, when compared to actual physicians' clinical experience and human skills, ChatGPT's accuracy and reliability may require further improvement. Further research is needed to determine how ChatGPT can be optimally used in educational and assessment processes and identify the areas where it will be most effective. Nonetheless, AI technologies like ChatGPT are expected to play an increasingly significant role in healthcare. These technologies offer educational support to healthcare professionals and contribute to the improvement of diagnosis and treatment processes. However, careful management is necessary to ensure these technologies' effective integration and reliability (17,20-22).

When the existing literature is examined, there are very few articles comparing the board exam results of different countries with ChatGPT's performance. Jain et al. (23) evaluated ChatGPT's decision-making process to assess the performance of the ChatGPT-3.5 version on the Orthopaedic In-Service Training Examination (OITE), conducted by the American Academy of Orthopaedic Surgeons and covering 11 topics, and to determine whether it is practical to adopt it as a resource in this field. At the end of the study, they found that ChatGPT-3.5 performed at the level of a first-year postgraduate (PGY-1) based on annual OITE technical reports for residents. They reported that ChatGPT performed better in basic science and sports. However, when the whole study was evaluated, they noted that ChatGPT in its current form lacks the essential capabilities to be a comprehensive tool in orthopaedic surgery (23).

Kung et al. (4) examined ChatGPT's performance in all three sections of the USMLE directly using publicly available questions on the official website. They reported that ChatGPT performed at or near the passing threshold in all three exams without any special training or support. Furthermore, ChatGPT showed high levels of cohesion and insight in its annotations. These results suggest that large language models may assist medical education and clinical decision-making (4). Gilson et al. (24) evaluated questions from the United States Medical Licensing Examination (USMLE) using ChatGPT. They reported that ChatGPT achieved accuracy rates of 44% (44/100), 42% (42/100), 64.4% (56/87), and 57.8% (59/102) in four data sets: AMBOSS-Step1, AMBOSS-Step2, NBME-Free-Step1, and NBME-Free-Step2, respectively. They also noted that ChatGPT performed 8.15% better than InstructGPT on average across all data sets, while GPT-3 performed similarly to random chance (24). In a study conducted in Peru, the Peruvian National Licensing Medical Examination [Examen Nacional de Medicina (ENAM)] was analyzed using GPT-3.5 and GPT-4. They found that ChatGPT (GPT-3.5 and GPT-4) was able to achieve expert-level performance on ENAM and outperformed most of the examinees (25).

### Study Limitations

The findings indicate that ChatGPT can play an essential role in exploring its potential in medical examinations. However, the study also highlighted several critical points and limitations of ChatGPT. First, the dataset used to evaluate ChatGPT's performance is limited in scope. It remains unclear whether the dataset on which ChatGPT is trained is comprehensive enough in the medical domain, which impacts its effectiveness in real-world scenarios (16,26).

Second, the evaluation of ChatGPT's exam performance has limitations when compared to physicians' performance. Chat Generative Pre-Trained Transformer's natural language processing capabilities differ from physicians' clinical experience and expertise. Thus, further research is needed to reach definitive conclusions regarding the real-world applicability of ChatGPT's exam performance (3,27).

Third, ethical and security issues surrounding ChatGPT should also be considered. Using AI systems like ChatGPT in medical training and assessment processes may raise patient privacy and security concerns. Therefore, it is crucial to address these concerns during the implementation of ChatGPT.

Fourth, ChatGPT's ability to process visual questions requires improvement. Compared to human perception and interpretation, ChatGPT currently has a limited capacity to understand visual information. This limitation affects its ability to accurately answer complex or specific questions requiring visual detail (26,28,29).

## Conclusion

This study provided a comprehensive evaluation of the performance of AI, specifically ChatGPT, in the TOTEK Qualifying Written Examination and showed that ChatGPT-4 correctly answered less than half of the TOTEK Qualifying Written Exam questions. Our analyses revealed that ChatGPT's ability to understand exam questions and produce appropriate answers was significantly lower compared to the average exam performance of human candidates. While these findings demonstrate ChatGPT's potential as a supportive tool for learning and exam preparation, they also emphasize that it cannot replace human guidance in areas requiring expertise and in-depth knowledge.

In conclusion, we should view ChatGPT and similar AI tools in medical education as aids, not as technologies meant to replace educators. Their role should be as supportive elements in learning processes. Future developments in this technology may allow AI to take a more active role in exam preparation and training; however, this requires ongoing evaluation and a human-centered approach.

**Footnote**

**Ethics Committee Approval:** Written permission was obtained from the TOTBID board of directors for this study (document no.: 159, dated: 26.04.2024).

**Informed Consent:** Not required.

**Financial Disclosure:** This study received no financial support.

## References

1. Atik OŞ. Artificial intelligence, machine learning, and deep learning in orthopedic surgery. Jt Dis Relat Surg. 2022;33:484-5.
2. Beyaz S. A brief history of artificial intelligence and robotic surgery in orthopedics & traumatology and future expectations. Jt Dis Relat Surg. 2020;31:653-5.
3. Sedaghat S. Early applications of ChatGPT in medical practice, education and research. Clin Med (Lond). 2023;23:278-9.
4. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198.
5. OpenAI. ChatGPT: optimizing language models for dialogue [Internet]. OpenAI; 2022 [cited 2024 July 24 1]. Available from: https://chatgpt.r4wand.eu.org/.
6. OpenAI Platform. Developer quickstart. Get up and running with the OpenAI API Available at: https://platform.openai.com/docs/quickstart / [Accessed: 24.07.2024].
7. Voytovich L, Greenberg C. Natural Language Processing: Practical Applications in Medicine and Investigation of Contextual Autocomplete. Acta Neurochir Suppl. 2022;134:207-14.
8. Zhou B, Yang G, Shi Z, Ma S. Natural Language Processing for Smart Healthcare. IEEE Rev Biomed Eng. 2024;17:4-18.
9. Névéol A, Zweigenbaum P. Clinical Natural Language Processing in 2014: Foundational Methods Supporting Efficient Healthcare. Yearb Med Inform. 2015;10:194-8.
10. Tözün R. The First Term Book of the Turkish Orthopaedics and Traumatology Education Council (TOTEK) (2001-2003), p:29
11. Gönen E, Berk H. Comparison of TOTEK (Turkish Orthopaedics and Traumatology Education Council) and EBOT (European Board of Orthopaedics and Traumatology Fellowship) qualification exams. TOTBİD J. 2014;13:488-99.
12. Güçlü B. The Ninth Term Book of the Turkish Orthopaedics and Traumatology Education Council (TOTEK) (2017-2019), p:43
13. Özdemir G. The Tenth Term Book of the Turkish Orthopaedics and Traumatology Education Council (TOTEK) (2019-2021), pp:26-27
14. Huri G. The Eleventh Term Book of the Turkish Orthopaedics and Traumatology Education Council (TOTEK) (2021-2023), pp:39-41
15. Huang J, Tan M. The role of ChatGPT in scientific communication: writing better scientific review articles. Am J Cancer Res. 2023;13:1148-54.
16. Uz C, Umay E. "Dr ChatGPT": Is it a reliable and useful source for common rheumatic diseases? Int J Rheum Dis. 2023;26:1343-9.

17. Morya VK, Lee HW, Shahid H, et al. Application of ChatGPT for Orthopedic Surgeries and Patient Care. Clin Orthop Surg. 2024;16:347-56.

18. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature. 2023;614:224-6.

19. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized Clinical Trials of Machine Learning Interventions in Health Care: A Systematic Review. JAMA Netw Open. 2022;5:e2233946.

20. Tomar L, Govil G, Dhawan P. Closed Negative Suction Drain Entrapment in Total Knee Arthroplasty: A Report on the Implications of a Broken Drain Based on the ChatGPT Outlook. Cureus. 2023;15:e36290.

21. Seth I, Rodwell A, Tso R, Valles J, Bulloch G, Seth N. A Conversation with an open artificial intelligence platform on osteoarthritis of the hip and treatment. J Orthop Sports Med. 2023;5:112-20.

22. Bernstein J. Not the Last Word: ChatGPT Can't Perform Orthopaedic Surgery. Clin Orthop Relat Res. 2023;481:651-5.

23. Jain N, Gottlich C, Fisher J, Campano D, Winston T. Assessing ChatGPT's orthopedic in-service training exam performance and applicability in the field. J Orthop Surg Res. 2024;3;19:27.

24. Gilson A, Safranek CW, Huang T, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312.

25. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, et al. Performance of ChatGPT on the Peruvian National Licensing Medical Examination: Cross-Sectional Study. JMIR Med Educ. 2023;9:e48039.

26. Liu PR, Zhang JY, Xue MD, et al. Artificial intelligence to diagnose tibial plateau fractures: an intelligent assistant for orthopedic physicians. Curr Med Sci. 2021;41:1158-64.

27. 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.

28. 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.

29. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. Pak J Med Sci. 2023;39:605-7.