ORIGINAL ARTICLE
88

J Turk Spinal Surg 2025;36(2):88-91
DOI: 10.4274/jtss.galenos.2025.15870

WHO IS MORE SUCCESSFUL IN A SPINAL SURGERY EXAMINATION? CHATGPT-3.5/4.0 OR A RESIDENT DOCTOR?

Sefa Erdem Karapınar[1], Recep Dinçer[1], Hüseyin Sina Coşkun[2], Özcan Kaya[3]

[1]Süleyman Demirel University Faculty of Medicine, Department of Orthopedics and Traumatology, Isparta, Türkiye
[2]Ondokuz Mayıs University Faculty of Medicine, Department of Orthopedics and Traumatology, Samsun, Türkiye
[3]University of Health Sciences Türkiye, Kanuni Sultan Süleyman Training and Research Hospital, Clinic of Orthopedics and Traumatology, İstanbul, Türkiye

ABSTRACT

Objective: As in all work sectors, artificial intelligence (AI) is now often used and has increased especially in the field of medicine with advances in technology. The aim of this study was to compare the responses given by Chat Generative Pre-trained Transformer (ChatGPT)-4.0, ChatGPT-3.5, and orthopaedics and traumatology residents to the Turkish Orthopedics and Traumatology Education Council (TOTEK) questions about the spine.

Materials and Methods: A total of 15 residents in the orthopaedics and traumatology clinic of a tertiary-level university hospital participated in an examination consisting of questions only related to the spine. The same questions were asked to ChatGPT-3.5 and ChatGPT-4.0 on two different days. The examination consisted of true/false questions, theoretical/classical and diagram/visual sections, with each section scored from 100 points. The average score was calculated and the results were evaluated by two instructors.

Results: The mean score obtained was 72.88 for ChatGPT-3.5 (p=0.005) and 69.38 for Chat GPT-4.0 (p=0.001), showing a 5.87% difference in success. The mean score obtained by the orthopaedic residents was 69.90 (p=0.779). Both the 3.5 and 4.0 versions of ChatGPT AI were observed to have a knowledge level equivalent to that of a 3rd year resident.

Conclusion: The 4th and 5th year orthopaedic residents were able to answer more questions correctly than ChatGPT-3.5 and GPT-4 on the spine assessment questions. Both ChatGPT-3.5 and GPT-4 performed better on text-only questions than on visual questions. It is unlikely that GPT-4 or ChatGPT-3.5 would pass the TOTEK written examination.

Keywords: ChatGPT, artificial intelligence, orthopaedics and traumatology, spinal surgery

INTRODUCTION

Artificial intelligence (AI) chatbots are computer programs that have the ability to understand human language and maintain a conversation with users with detailed responses. As in all sectors, there have been very rapid technological developments in medicine. The use of online resources to access correct medical information has increased especially since the beginning of the 2000s. It has been reported that 84% of the patients of an orthopaedics and traumatology clinic have access to the internet and 64% have used online sources of orthopaedic information[1,2]. Therefore, the accuracy of this information must be examined, and it should be ensured that people do not have incorrect information. Patient access to correct information can provide benefit in respect of patient compliance with treatment and better outcomes, and it can increase patient satisfaction.

Chat Generative Pre-trained Transformer (ChatGPT) is a large language model (LLM) with increasingly widespread use. LLMs have attracted great interest, especially in the field of medicine[3]. ChatGPT was developed by OpenAI. Due to the human-like responses generated, it is increasing in popularity with more than 100 million users currently[4]. It is trained by being exposed to various reference sources and it uses this information obtained from many books and articles. By learning past data as patterns, sequences of words and sentences according to the links are presented as the output. The number of parameters is very important for GPT, as a greater number of parameters provides a greater learning capacity. This has the advantage of resolving the complex structure of human language. While ChatGPT-2 has approximately 1.5 billion

Address for Correspondence: Sefa Erdem Karapınar, Süleyman Demirel University Faculty of Medicine, Department of Orthopedics and Traumatology, Isparta, Türkiye
E-mail: sefaerdemkarapinar@gmail.com
ORCID ID: orcid.org/0000-0002-6878-2243
Received: 04.11.2024 **Accepted:** 18.03.2025 **Epub:** 20.03.2025 **Publication Date:** 15.04.2025
Cite this article as: 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(2):88-91

parameters, there are approximately 175 billion parameters in ChatGPT-3.5, which was launched in 2022[5].

In the ChatGPT-4.0 model, which was launched on 14 March 2023, it has been suggested that there are 1.76 trillion parameters[6]. AI has started to be used recently not only for reasoning skills but also for field-specific examinations. Previous studies have shown that ChatGPT-3.5 almost passed the first, second, and third stages of the United States Medical Licensing Examination (USMLE). Further studies have shown a 20% increase in the points of the three USMLE using ChatGPT-4.0[7].

Orthopaedic surgery in practice and on examinations is distinguished by the frequent need to synthesize imaging data in formulating treatment plans. There are studies in the literature to determine the clinical diagnosis and treatment process with ChatGPT-3.5. It has been attempted to determine the success rate of ChatGPT-3.5 in answering Board examination questions[7,8]. The questions in this study were taken from the examination for residents, which is organized regularly every year by the Turkish Orthopedics and Traumatology Education Council (TOTEK). The aim of the study was to present the comparative results of the responses of residents, ChatGPT-3.5 and ChatGPT-4.0 to these examination questions related to the spine.

MATERIALS AND METHODS

The questions of the in-clinic training examination, taken on 01.06.2024 by a total of 15 doctors undertaking residency as research assistants in the orthopaedics and traumatology clinic, were asked to ChatGPT-3.5 and ChatGPT-4.0. All the questions in the examination were related to spinal surgery. Ethics Committee approval was not required for this study. The examination consisted of 4 sections. The first section comprised 20 multiple-choice questions, each with 5 options and a value of 5 points. The second section comprised 20 classic questions as a theory examination and each question had a value of 5 points. In the third section, it was asked whether 40 sentences were true or false, and the correct response for each sentence was scored as 2.5 points. To determine the power of AI in visual interpretation, questions in the fourth section were related to diagrams and radiographs. There were 10 questions with a value of 10 points for each. As the examination was formed of 4 sections, with each section scored from 100 points, the average of the total points scored was recorded. The examination was repeated the next day with the same questions. The aim of this was to measure how similar the responses were that were given at different timepoints. All the examinations were evaluated by two specialist physicians. Taking the average of the total points obtained provided more objective examination results. The doctors were separated into 5 groups based on their years of seniority. The scores obtained were examined and compared between ChatGPT-3.5, ChatGPT-4.0 and with those obtained by the doctors.

Statistical Analysis

Data obtained in the study were analyzed statistically usisng SPSS vn. 29 software (IBM, Armonk, NY, USA). Categorical data such as correct or incorrect answers given by ChatGPT-3.5, GPT-4.0 and doctors were compared using chi-square analysis. Numerical data of the three groups were compared using analysis of variance with post-hoc testing using the Tukey test. Chi-square analysis was also used to compare the accuracy among the seven different subspecialties.

RESULTS

Scores were obtained according to the results of the examination, which consisted of 4 sections and was taken on two days. It was seen that in the multiple-choice section of the examination a better score was obtained by ChatGPT-3.5 than by ChatGPT-4.0 on the first day, and on the second day both versions obtained the same scores. Higher points were obtained by ChatGPT-3.5 in the true/false and theory/classic sections on both days. In the diagram/visual section of the examination, both versions obtained the same points on the first day, and on the second day ChatGPT-4.0 scored higher points than ChatGPT-3.5. According to the total mean points on the first day, the ChatGPT-3.5 version was seen to be more successful. The results of the examination on the second day were close to each other for both AI versions (Table 1).

Scoring was applied to the responses to the questions asked to both the residents and the two versions of AI. The doctors were separated into 5 groups according to their years of seniority. The mean points for each examination category were seen to be proportional to the years of seniority. The mean scores of both ChatGPT-3.5 and ChatGPT-4.0 were observed to be the equivalent of the knowledge level of 3rd year residents (Table 2).

Overall, the orthopaedic residents scored an average of 69.90 points (Table 3). ChatGPT-3.5 and ChatGPT-4 had overall scores of 72.88 and 69.38 points, respectively. The difference among the three groups in test success was statistically significant. ChatGPT-3.5 scored higher than the orthopaedic residents and ChatGPT-4 (p=0.001, p=0.779, respectively).

DISCUSSION

AI chatbot technology is trained on an abundance of information including peer-reviewed journal articles, texts, news articles, and online resources[9]. The results of this study showed that ChatGPT-3.5 outperformed orthopaedic residents and ChatGPT-4 in answering spine questions in the TOTEK question bank, although the 4th and 5th year orthopaedic residents were able to answer more questions correctly. This notable difference points to the extensive skill set required to answer orthopaedic assessment questions and can perhaps be translated into clinical practice. Unlike assessment examinations in other disciplines, orthopaedic examinations require special scrutiny

of radiographic images in conjunction with clinical assessment, which reflects the critical thinking that orthopaedic surgeons need every day and may currently be beyond the ability of these chatbots. Therefore, chatbots seem to be more successful in standard question patterns that do not require analytical thinking.

One of the most important points to be discussed in this study is that ChatGPT-3.5 was 11.12% more successful than ChatGPT-4.0 on the first day. On the second day, ChatGPT-3.5 was similarly 0.62% more successful. The crucial point here is that the same questions asked on different days received different answers. Thus, different answers and different scores in two examinations given at least 24 hours apart were compared. It was noticed that in previous studies in the literature, different answers given at different times were not taken into account or were ignored. It should be emphasized that this part remains important in terms of timing. In the current study analysis of the four different categories, it was seen that the most difficult section for AI was the diagram/visual section, which is based on interpretation. This means that AI bots such as ChatGPT need to be improved in matters of analytical thinking and interpretation. In a study by Massey et al.[10], the results of an examination using 180 questions from the ResStudy orthopaedic examination question bank were seen to be similar to the current study findings in that ChatGPT gave more correct responses to text questions than to diagram-based questions. Kung et al.[11] asked questions from the American Board of Orthopaedic Surgery part 1 examination to ChatGPT-4.0, and the AI exceeded the pass score of 67% of this examination. A dataset of 400 questions was used in a study by Lum[12], and it was reported that the ChatGPT results were similar to those of a first-year resident. In the current study, the results obtained by ChatGPT were at the knowledge level of a third-year resident.

Ali et al.[13], compared both ChatGPT-3.5 and GPT-4 on the American Board of Neurological Surgery self-assessment examination 1, and reported that ChatGPT-3.5 and ChatGPT-4 scored 73.4% and 83.4%, respectively. Those results were higher than the performance in the current study on orthopaedic assessment questions. However , 22% of the neurosurgery test questions had images in that study by Ali et al.[13], whereas at least 50% of orthopaedic examination questions have images, which could explain why ChatGPT-3.5 and ChatGPT-4 had more difficulties in the current study.

Before using AI-generated text for commercial purposes, it must be ensured that it does not violate existing copyright. According to the nature news team, ChatGPT cannot be accepted as the author of a study as it cannot take responsibility for the accuracy and legitimacy of scientific research[14].

Examination of AI in literature shows that it is useful in tasks ranging from data analysis to the formation of hypotheses and results. However, it must be accepted that there are certain

Table 1. ChatGPT examination results

Questions	1st day ChatGPT-3.5	1st day ChatGPT-4.0	2nd day Chat GPT-3.5	2nd day ChatGPT-4.0
Multiple-choice test	85.00	70.00	80.00	80.00
True/false	77.50	65.00	82.50	75.00
Theory/classic	84.00	67.00	88.00	74.00
Diagrams/visual	50.00	50.00	36.00	55.00
Total points	74.13	63.00	71.62	71.00

ChatGPT: Chat Generative Pre-trained Transformer

Table 2. Examination points of the residents

Questions	1st year resident	2nd year resident	3rd year resident	4th year resident	5th year resident
Test	50.00	65.00	70.00	85.00	90.00
True/false	45.00	57.50	72.50	85.00	90.00
Theory/classic	38.00	54.00	72.00	84.00	92.00
Diagrams/visual	40.00	58.00	70.00	86.00	94.00
Total points	43.25	58.62	71.12	85.00	91.50

Table 3. Orthopaedic assessment examination scores of residents, ChatGPT-3.5, and ChatGPT-4

	Overall scores		
	Mean	SD	p-value
ChatGPT-3.5	72.88	20.06	0.005
GPT-4.0	69.38	11.96	0.001
Orthopedic residents	69.90	3.60	0.779

ChatGPT: Chat Generative Pre-trained Transformer, SD: Standard deviation

potential difficulties and limitations related to the use of ChatGPT in orthopaedic research. Responses to the model may require specialisation and more specific information from orthopaedic specialists to avoid errors or incomplete information[15].

Study Limitations

Limitations of this study were that it was not a systematic examination and that no critical evaluation was performed. To compare categorical data versus numerical data, the sections were averaged to return an average score for each section. This resulted in a smaller sample size when comparing the averages of the sections, but notable differences were still seen. In addition, although comparisons of all 80 total questions were sufficiently powered, it should be noted that comparisons between sections with 20 and 40 questions respectively, were likely to have been underpowered.

CONCLUSION

The accuracy and reliability of the answers provided by ChatGPT in the examination in this study depended on the quality of the training data and algorithms used. The 4th and 5th year orthopaedic residents were able to answer more of the TOTEK spine assessment questions correctly than ChatGPT-3.5 and ChatGPT-4. Both ChatGPT-3.5 and ChatGPT-4 performed better on text-only questions than on visual questions. It is unlikely that either ChatGPT-4 or ChatGPT-3.5 would pass the TOTEK and spine questions written examination.

Ethics

Ethics Committee Approval: Ethics committee approval is not required as this is not a clinical trial.
Informed Consent: As this is not a clinical trial, no consent form is required.

Acknowledgement

I would like to thank Dr. Umut Can Duvarcı for his assistance in data collection.

Footnotes

Authorship Contributions

Surgical and Medical Practices: S.E.K., R.D., H.S.C., Ö.K., Concept: S.E.K., R.D., H.S.C., Ö.K., Design: S.E.K., R.D., H.S.C., Ö.K., Data Collection or Processing: S.E.K., R.D., H.S.C., Ö.K., Analysis or Interpretation: S.E.K., R.D., H.S.C., Ö.K., Literature Search: S.E.K., R.D., H.S.C., Ö.K., Writing: S.E.K., R.D., H.S.C., Ö.K.
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. Sparks CA, Fasulo SM, Windsor JT, Bankauskas V, Contrada EV, Kraeutler MJ, et al. ChatGPT is moderately accurate in providing a general overview of orthopaedic conditions. JB JS Open Access. 2024;9:e23.00129.
2. Burrus MT, Werner BC, Starman JS, Kurkis GM, Pierre JM, Diduch DR, et al. Patient perceptions and current trends in internet use by orthopedic outpatients. HSS J. 2017;13:271-5.
3. Pal S, Bhattacharya M, Lee SS, Chakraborty C. A domain-specific next-generation large language model (LLM) or ChatGPT is required for biomedical engineering and research. Ann Biomed Eng. 2024;52:451-4.
4. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, 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.
5. Herzog I, Mendiratta D, Para A, Berg A, Kaushal N, Vives M. Assessing the potential role of ChatGPT in spine surgery research. J Exp Orthop. 2024;11:e12057.
6. Kaarre J, Feldt R, Keeling LE, Dadoo S, Zsidai B, Hughes JD, et al. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. Knee Surg Sports Traumatol Arthrosc. 2023;31:5190-8.
7. 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.
8. Mert M, Vahabi A, Daştan AE, Kuyucu A, Ünal YC, Tezgel O, et al. Artificial intelligence's suggestions for level of amputation in diabetic foot ulcers are highly correlated with those of clinicians, only with exception of hindfoot amputations. Int Wound J. 2024;21:e70055.
9. Where does ChatGPT get its information from? Scribbr. Last Accessed date: 18.03.2025. Available from: https://www.scribbr.com/frequently-asked-questions/chatgpt-information
10. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. J Am Acad Orthop Surg. 2023;31:1173-9.
11. 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.
12. 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.
13. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, et al. Performance of ChatGPT, GPT-4, and Google bard on a neurosurgery oral boards preparation question bank. Neurosurgery. 2023;93:1090-8.
14. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Curr Cardiol Rep. 2014;16:441.
15. Chatterjee S, Bhattacharya M, Pal S, Lee SS, Chakraborty C. ChatGPT and large language models in orthopedics: from education and surgery to research. J Exp Orthop. 2023;10:128.