

DOI: <http://dx.doi.org/10.12996/gmj.2024.4067>

ChatGPT vs. Orthopedic Residents! Who is the Winner?

ChatGPT Ortopedi Asistanlarına Karşı! Kazanan Kim?

© Semih Yaş¹, © Asim Ahmadov¹, © Alim Can Baymurat¹, © Mehmet Ali Tokgöz¹, © Secdegül Coşkun Yaş², © Mustafa Odlyurt³, © Tolga Tolunay¹

¹Gazi University Faculty of Medicine, Department of Orthopedics and Traumatology, Ankara, Türkiye

²Ankara Training and Research Hospital, Clinic of Emergency Medicine, Ankara, Türkiye

³Zonguldak Çaycuma State Hospital, Clinic of Orthopedics and Traumatology, Zonguldak, Türkiye

ABSTRACT

Objective: In recent advancements in artificial intelligence, ChatGPT by OpenAI has emerged as a versatile tool capable of performing various tasks; however, its application in medicine is challenged by complexities and limitations in accuracy. This article aims to compare ChatGPT's performance with orthopedic residents at Gazi University in a multiple-choice exam to assess its applicability and reliability in the field of orthopedics.

Methods: In this observational study conducted at Gazi University, 31 orthopedic residents were stratified by experience level and assessed using a 50-question multiple-choice test on various orthopedic topics. The study also evaluated ChatGPT 3.5's responses to the same questions, focusing on both the correctness and reasoning behind the answers.

Results: Orthopedic residents tested, ranging from 6 months to 5 years in experience, scored between 23 and 40 out of 50 in a multiple-choice exam, with a mean score of 30.81, varying by seniority. ChatGPT provided correct answers for 25 out of 50 questions, showing consistency in different languages and times, but also exhibited limitations by giving incorrect responses or stating that the correct answer was not among the choices for some questions.

Conclusion: While ChatGPT can accurately answer some theoretical questions, its effectiveness is limited in interpretive scenarios and in situations with multiple variables, although its accuracy may improve with updates over time.

Keywords: ChatGPT, artificial intelligence, orthopedics, traumatology

ÖZ

Amaç: Yapay zeka alanındaki son gelişmelerde, OpenAI tarafından geliştirilen ChatGPT çeşitli görevleri yerine getirebilen çok yönlü bir araç olarak ortaya çıkmıştır, ancak tıp alanındaki uygulaması karmaşıklıklar ve doğruluk sınırlamaları nedeniyle zorlanmaktadır. Bu makale, ChatGPT'nin performansını Gazi Üniversitesi'ndeki ortopedi asistanları ile çoktan seçmeli bir sınavda karşılaştırarak ortopedi alanında uygulanabilirliğini ve güvenilirliğini değerlendirmeyi amaçlamaktadır.

Yöntemler: Gazi Üniversitesi'nde yapılan bu gözlemsel çalışmada, 31 ortopedi asistanı deneyim düzeyine göre sınıflandırılmış ve çeşitli ortopedik konularda 50 soruluk çoktan seçmeli bir test kullanılarak değerlendirilmiştir. Çalışmada ayrıca ChatGPT 3.5'in aynı sorulara verdiği yanıtlar, yanıtların hem doğruluğuna hem de arkasındaki mantığa odaklanılmıştır.

Bulgular: Deneyimleri 6 ay ile 5 yıl arasında değişen ortopedi asistanları, çoktan seçmeli bir sınavda 50 üzerinden 23 ile 40 arasında puan aldı ve kıdeme göre değişmekle birlikte ortalama puan 30,81 oldu. ChatGPT, farklı dillerde ve zamanlarda tutarlılık göstererek 50 sorudan 25'ine doğru yanıt vermiş, ancak yanlış yanıtlar vererek veya bazı sorular için doğru yanıtın seçenekler arasında olmadığını belirterek sınırlamalar da sergilemiştir.

Sonuç: Çalışma, ChatGPT'nin bazı teorik soruları doğru bir şekilde yanıtlayabilse de, yorumlayıcı senaryolarda ve çok değişkenli durumlarda etkinliğinin sınırlı olduğu, ancak zaman içinde kişiselleştirilmiş güncellemelerle doğruluğunun artabileceği sonucuna varmaktadır.

Anahtar Sözcükler: ChatGPT, yapay zeka, ortopedi, travmatoloji

Address for Correspondence/Yazışma Adresi: Semih Yaş MD, Gazi University Faculty of Medicine, Department of Orthopedics and Traumatology, Ankara, Türkiye

E-mail / E-posta: drsemihyas@gmail.com

ORCID ID: orcid.org/0000-0001-7823-3400

Received/Geliş Tarihi: 05.12.2023

Accepted/Kabul Tarihi: 24.02.2024



©Copyright 2024 The Author. Published by Galenos Publishing House on behalf of Gazi University Faculty of Medicine. Licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND) International License.

©Telif Hakkı 2024 Yazar. Gazi Üniversitesi Tıp Fakültesi adına Galenos Yayınevi tarafından yayımlanmaktadır. Creative Commons Atıf-GayriTicari-Türetilemez 4.0 (CC BY-NC-ND) Uluslararası Lisansı ile lisanslanmaktadır.

INTRODUCTION

In recent years, advancements in the field of artificial intelligence have experienced an upsurge in the scientific community. Of particular note, Chat Generative Pre-trained Transformer (ChatGPT) developed by OpenAI and endowed with a chatbot capability has been described as a groundbreaking development in this domain. Launched in November 2022, ChatGPT, unlike other chatbots, can respond to questions very quickly and can be used for different purposes (1). For example, it can write code for computer software, create a film script or a story, and answer multiple-choice questions in written form (2). However, despite all these impressive features, the use of artificial intelligence programs in the field of medicine can be limited compared with other areas because of the large number of variables involved. The impact of these developments on academic life is still a topic of study that has not yet been clearly defined.

In some journals, publications have begun to emerge where ChatGPT is recognized as a co-author (3,4). In contrast to the journals that have recognized ChatGPT as a co-author, some publications have raised concerns over the ethical implications of attributing authorship to an AI language model such as ChatGPT (5).

Despite its many impressive capabilities, ChatGPT has certain limitations and undesirable features. According to information provided by OpenAI, the company that developed the program, ChatGPT is capable of citing non-existent articles and processing non-existent data. Given the risk of introducing not only erroneous information but also plagiarism into academic publications, this raises concerns over the reliability of scientific records. In addition, it should be noted that ChatGPT's responses to questions may be incorrect, yet presented in a coherent manner, potentially creating a dangerous situation for non-healthcare professionals reliant on the program's output. The provision of inaccurate data by ChatGPT could lead to negative outcomes in future research or healthcare decisions.

The use of artificial intelligence programs to search internet data and find answers to many questions is increasing daily. As evidenced by publications related to ChatGPT in 2023, studies across various scientific fields, including public health and orthopedic surgery, have been conducted (6). It is still a matter of debate whether passing grades can be obtained in some written exams using this program (7,8). This situation has led to restrictions on the use of the program in some countries and universities.

The aim of this study was to demonstrate the level of success of ChatGPT, which has recently become a popular topic and is gaining popularity in academic circles, in a multiple-choice orthopedic exam by comparing it with the answers of orthopedic residents.

MATERIALS AND METHODS

Study Design

This observational study is planned to be conducted at a tertiary hospital that is actively involved in resident training. The study participants comprised 31 orthopedic residents from the Department of Orthopedics and Traumatology. These residents

were selected based on voluntary participation and were stratified into five groups according to their level of experience: 6 months to 1 year, 1-2 years, 2-3 years, 3-4 years, and 4-5 years. This stratification ensured a diverse range of expertise and perspectives within the field of orthopedics.

Test Design

A comprehensive test consisting of 50 multiple-choice questions was designed to assess knowledge in various domains of orthopedics, including basic orthopedics, trauma, spine, orthopedic tumors, arthroplasty, and pediatric orthopedics. The questions, each with only one correct answer, were meticulously crafted by a working group of senior orthopedic professors, ensuring the validity and relevance of the content. Some examples of the questions asked to ChatGPT are presented in Figures 1-3.

Data Collection: Residents

The test was conducted by the residents under fair, controlled conditions to maintain the integrity of the responses. The time allocated, environment, and mode of answer submission were standardized for all participants. Responses were collected and anonymized for further analysis.

Data collection - ChatGPT

The same set of questions was presented to the ChatGPT 3.5 program, developed by OpenAI, at two different times to evaluate consistency in responses. For scenario-based questions, we used the same ChatGPT session to benefit from the AI's memory retention capabilities. For independent questions, a new session was initiated for each question to simulate a fresh interaction, mimicking a real-world clinical query scenario.

Ethics

The study received ethical approval from the Ethical Committee of Gazi University (approval number: E-77082166-604.01.02-643268, date: 27.04.2023). The research team ensured that all aspects of the study were conducted in accordance with the highest standards of academic integrity and ethical research practice.

Statistical Analysis

Data from both the residents' exams and ChatGPT responses were collated and coded for analysis. Responses were categorized as "correct", "incorrect", or "invalid/no answer". For ChatGPT, additional categorization was done for "consistent response" and "different explanations."

Statistical analysis was conducted using IBM SPSS (IBM Corp. Released 2020. IBM SPSS Statistics for Macintosh, Version 27.0. Armonk, NY: IBM Corp. Descriptive statistics were generated to summarize the basic features of the data. This included computation of means, standard deviations, and ranges for the number of correct answers. A comparison was then made between the answers provided by the AI and those given by the orthopedic residents. This comparison focused on not only the correctness of the answers but also the reasoning and explanation provided, especially for complex or scenario-based questions.

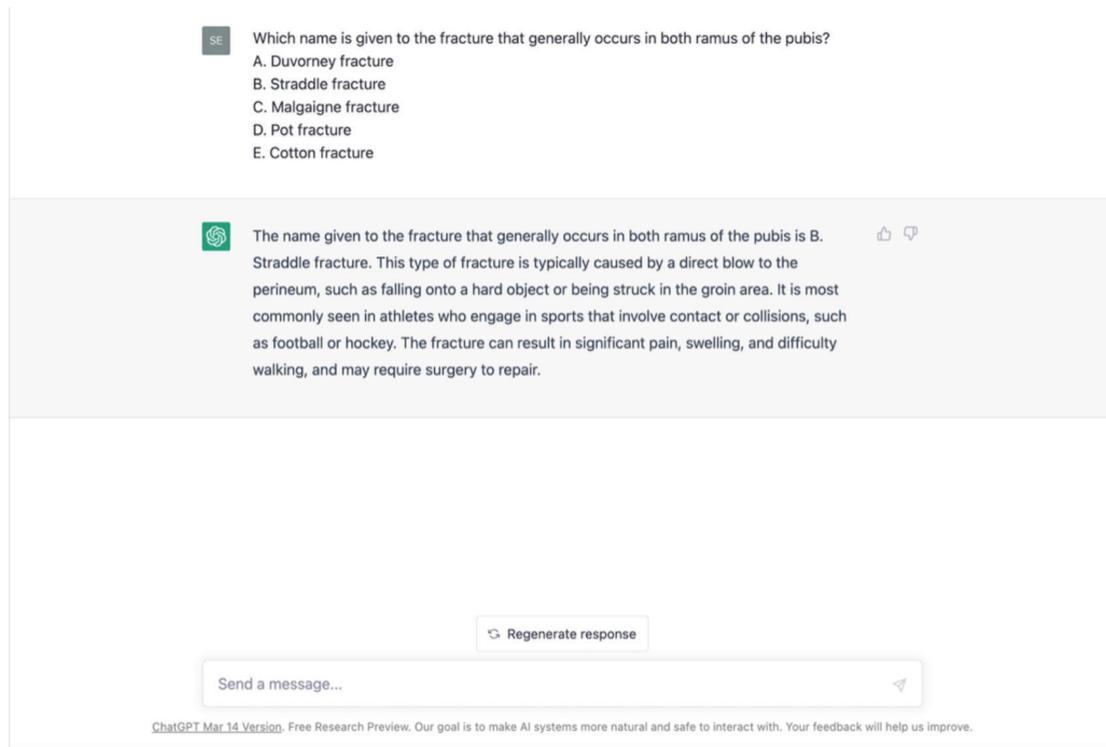


Figure 1. Example question that ChatGPT answered correctly.

ChatGPT: Chat Generative Pre-trained Transformer.

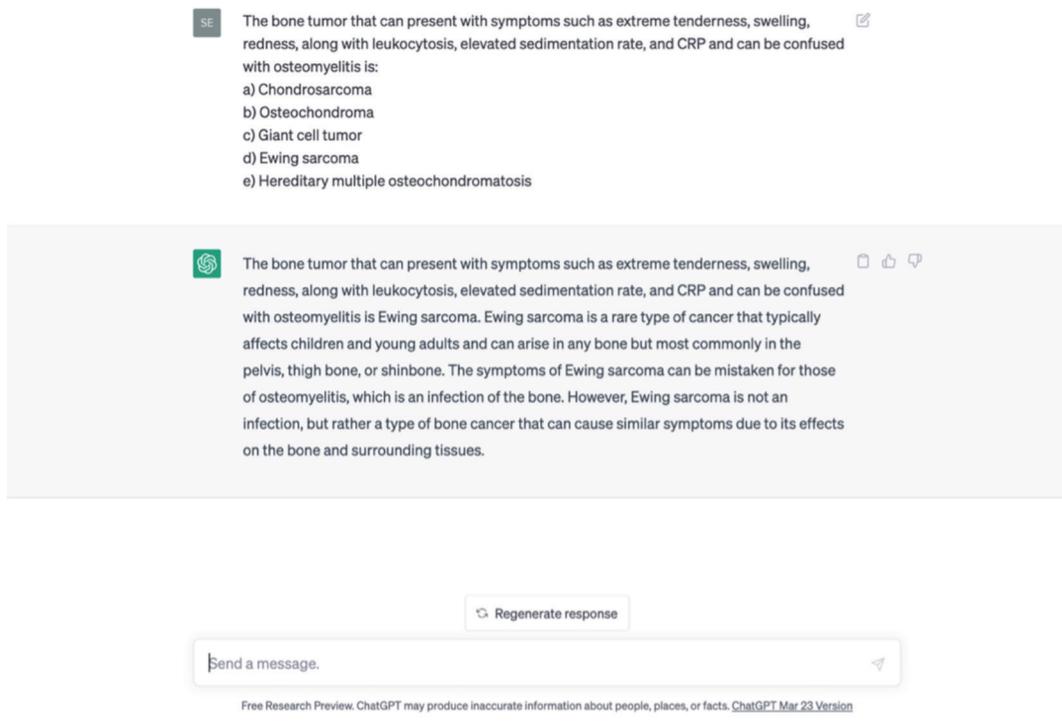


Figure 2. An English question example that ChatGPT answered correctly.

ChatGPT: Chat Generative Pre-trained Transformer.

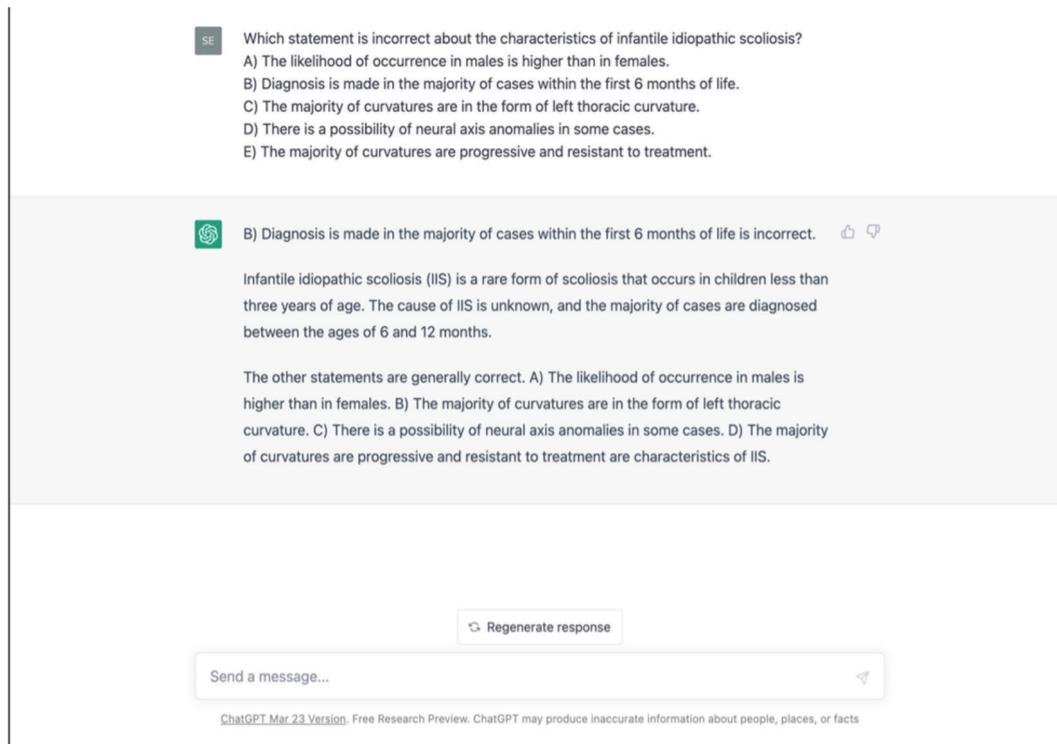


Figure 3. Example question that ChatGPT answered incorrectly.

ChatGPT: Chat Generative Pre-trained Transformer.

RESULTS

The exam results of 31 orthopedic resident doctors with a seniority ranging from 6 months to 5 years were included. Among the 31 orthopedic residents, 7 of them (22.6%) had seniority between 6 months and 1 year, 6 of them (19.35%) had seniority between 1 and 2 years, 6 of them (19.35%) had seniority between 2 and 3 years, 6 of them (19.35%) had seniority between 3 and 4 years, and the remaining 6 of them (19.35%) had seniority between 4 and 5 years (Figure 4).

The number of correct answers obtained by 31 orthopedic resident doctors who took the exam was calculated to have a minimum of 23 and a maximum of 40 out of 50, with a mean of 30.81. The mean score of orthopedic residents with seniority between 6 months and 1 year was calculated to be 25.86 (± 2.26) correct out of 50 multiple-choice questions. The mean score of residents with seniority between 1 and 2 years was also determined to be 25.33 (± 3.67). The mean of correct answers for residents with a seniority between 2 and 3 years was 29.89 (± 5.49). The mean of correct answers for residents with a seniority between 3 and 4 years was 35.5 (± 2.42). The mean number of correct responses for the most experienced orthopedic residents with a seniority of 4 to 5 years was computed as 38.33 (± 1.5).

The ChatGPT was asked 50 multiple-choice orthopedic questions via the chatbot link <https://chat.openai.com/chat> in both Turkish and English at different times. Consistent answers were provided

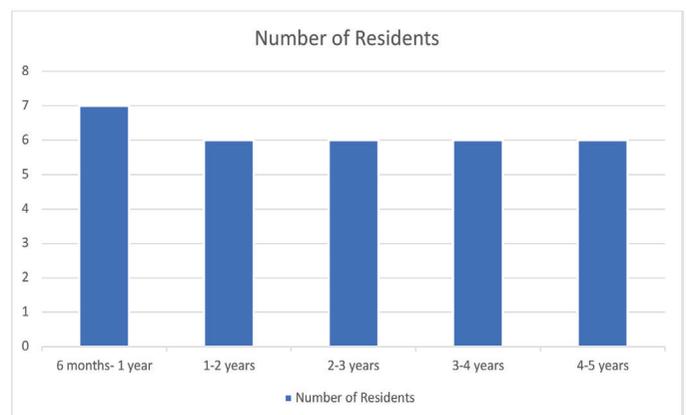


Figure 4. Number of residents.

by the program regardless of the language or time of questioning. However, the program provided different explanations for the same answer when the questions were asked at different times. The program's answers were internally consistent in different languages and at different times. ChatGPT provided the correct answer for 25 of the 50 multiple-choice questions. It indicated that two questions were incorrect, stating that the correct answer was not among the choices. It gave incorrect answers to 23 questions (Figure 5).

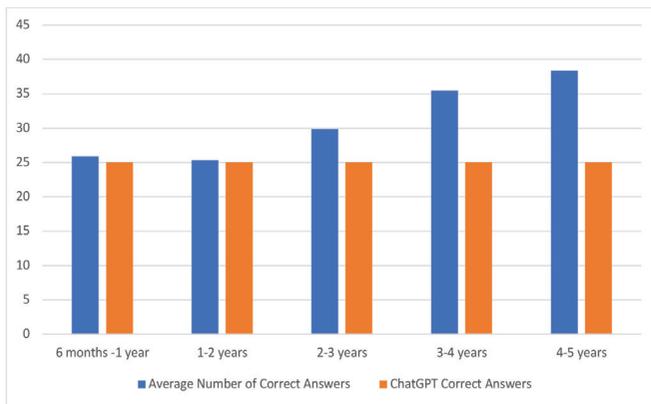


Figure 5. Comparison of residents' and ChatGPT's correct answers.

ChatGPT: Chat Generative Pre-trained Transformer.

DISCUSSION

Our study adds to the growing body of research evaluating the capabilities of AI, specifically ChatGPT, in the medical field. In our analysis, ChatGPT demonstrated a level of knowledge comparable to that of orthopedic residents with 6 months to 2 years of experience, correctly answering 50% of the questions. However, it showed limitations in questions requiring interpretation or inference, and there were concerns about the accuracy and reliability of its sources.

A study highlighted that ChatGPT 3.5, along with ChatGPT 4, was prone to generate fabricated bibliographic citations, a phenomenon categorized as a type of "hallucination" (9). This issue was obvious in our study as well, where ChatGPT provided false information with fabricated sources. This phenomenon poses significant concerns for the use of AI in academic and clinical settings where the accuracy of sources is paramount.

Upon examination of its responses, it can be considered a potential danger that ChatGPT presents false information in a fluent and well-formed manner, even when it is incorrect. In addition, ChatGPT's success rate in a multiple-choice orthopedic exam was found to be inadequate. Upon reviewing the literature, it is possible for the ChatGPT artificial intelligence program to achieve near-passing grades in certain exams.

In the study conducted by Fijačko et al. (7), the questions from two distinct exams developed by the American Heart Association were directed to ChatGPT for analysis. ChatGPT answered 68.4% and 76% of the questions correctly in these exams. In this study, ChatGPT could not answer a few questions correctly, exceeding the passing threshold of the exams. In our study, ChatGPT answered 50% of the questions correctly.

In another research study, the "United States Medical Licensing Exam" questions consisting of three stages were presented to ChatGPT, and ChatGPT approached the passing score in almost all stages (8). In a research conducted in a non-medical domain, ChatGPT was exposed to four distinct final exam questions from a law faculty, and it successfully achieved a passing grade for all of the exams (10). In our study, ChatGPT answered a similar number of questions correctly as the first-year resident. This may indicate that ChatGPT has more knowledge in certain areas.

Sahin et al. (11) reported that ChatGPT is a successful study assistant; however, the way the questions are asked is important in the success of ChatGPT. Yapar et al. (12) mentioned in their study that ChatGPT can provide strong support for patients in home care in the early period after orthopedic procedures.

In another study evaluating the success of ChatGPT-3.5, ChatGPT-4, and orthopedic residents, it was shown that orthopedic residents were more successful than ChatGPT and ChatGPT-4 was more successful than ChatGPT-3.5 (13). This was similar to the result in our study.

A study in orthopedics showed that the ChatGPT answered approximately 65% of the questions about anterior cruciate ligament surgery correctly (14). However, although ChatGPT provides guidance and effectively adapts to different target audiences, it cannot replace the expertise of orthopedic surgeons in diagnosis and treatment planning because of its limited knowledge in orthopedics and potential for inaccurate answers.

Analyzing these studies, it can be concluded that ChatGPT can produce more positive results in non-medical fields, but it may not provide sufficient results due to the large number of variables involved in medical subjects. Considering the results of our study, the performance of ChatGPT is limited, and although it seems to be helpful in solving some exam questions, it is not sufficient to provide accurate information. Despite its potential to produce different answers to the same questions at different times with different explanations, it should not be overlooked that ChatGPT can be used in academic settings and multiple-choice exams, albeit in a limited way. Although its current medical use appears to be limited, the accuracy of the information provided by the program may increase over time with further research and development. However, it should also be noted that there is a risk that both positive and negative practices may increase as the program improves, raising ethical concerns.

Study Limitations

Our study was limited by the sample size and scope of the questions. Future studies could utilize a larger pool of participants and questions from standardized exams such as the orthopedic board exam for a more comprehensive evaluation.

CONCLUSION

ChatGPT was found to have entry-level knowledge compared with orthopedic residents. It may provide accurate information in answering certain theoretical questions, but the information it provides for questions requiring interpretation and inference may not be at the desired level. However, the accuracy of the theoretical knowledge may increase with updates developed over time.

Ethics

Ethics Committee Approval: The study received ethical approval from the Ethical Committee of Gazi University (approval number: E-77082166-604.01.02-643268, date: 27.04.2023).

Informed Consent: It was obtained.

Authorship Contributions

Surgical and Medical Practices: S.Y., Concept: S.Y., Design: S.Y., Supervision: T.T., Data Collection or Processing: M.O., Analysis or Interpretation: S.C.Y., T.T., Literature Search: A.A., A.C.B., M.A.T., S.C.Y., Writing: S.Y., A.A., S.C.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

- Rudolph J, Tan S, Tan S. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching*, 2023.
- Rivas P, Zhao L. Marketing with ChatGPT: Navigating the Ethical Terrain of GPT-Based Chatbot Technology. *AI* 2023; 4: 375-84.
- Polonsky MJ, Rotman JD. Should Artificial Intelligent Agents be Your Co-author? Arguments in Favour, Informed by ChatGPT. *Australasian Marketing Journal* 2023; 31: 91-6.
- Biswas S. ChatGPT and the Future of Medical Writing. *Radiology* 2023; 307: e223312.
- Thorp HH. ChatGPT is fun, but not an author. *Science* 2023; 379: 313.
- Biswas SS. Role of Chat GPT in Public Health. *Ann Biomed Eng* 2023; 51: 868-9.
- Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American heart association course? *Resuscitation* 2023; 185: 109732.
- 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.
- Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep* 2023; 13: 14045.
- Choi JH, Hickman KE, Monahan A, Schwarcz D. ChatGPT goes to law school. Available at SSRN, 2023.
- Sahin MC, Sozer A, Kuzucu P, Turkmen T, Sahin MB, Sozer E, et al. Beyond human in neurosurgical exams: ChatGPT's success in the Turkish neurosurgical society proficiency board exams. *Comput Biol Med* 2024; 169: 107807.
- Yapar D, Demir Avci Y, Tokur Sonuvar E, Egerci ÖF, Yapar A. ChatGPT's potential to support home care for patients in the early period after orthopedic interventions and enhance public health. *Jt Dis Relat Surg* 2024; 35: 169-76.
- 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.
- 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.