



AOA Critical Issues in Education

ChatGPT Performs at the Level of a Third-Year Orthopaedic Surgery Resident on the Orthopaedic In-Training Examination

Diane Ghanem, MD, Oscar Covarrubias, BS, Micheal Raad, MD, Dawn LaPorte, MD, FAOA, and Babar Shafiq, MD, FAOA

Investigation performed at The Johns Hopkins Hospital, Baltimore, Maryland

Introduction: Publicly available AI language models such as ChatGPT have demonstrated utility in text generation and even problem-solving when provided with clear instructions. Amidst this transformative shift, the aim of this study is to assess ChatGPT's performance on the orthopaedic surgery in-training examination (OITE).

Methods: All 213 OITE 2021 web-based questions were retrieved from the AAOS-ResStudy website (<https://www.aaos.org/education/examinations/ResStudy>). Two independent reviewers copied and pasted the questions and response options into ChatGPT Plus (version 4.0) and recorded the generated answers. All media-containing questions were flagged and carefully examined. Twelve OITE media-containing questions that relied purely on images (clinical pictures, radiographs, MRIs, CT scans) and could not be rationalized from the clinical presentation were excluded. Cohen's Kappa coefficient was used to examine the agreement of ChatGPT-generated responses between reviewers. Descriptive statistics were used to summarize the performance (% correct) of ChatGPT Plus. The 2021 norm table was used to compare ChatGPT Plus' performance on the OITE to national orthopaedic surgery residents in that same year.

Results: A total of 201 questions were evaluated by ChatGPT Plus. Excellent agreement was observed between raters for the 201 ChatGPT-generated responses, with a Cohen's Kappa coefficient of 0.947. 45.8% (92/201) were media-containing questions. ChatGPT had an average overall score of 61.2% (123/201). Its score was 64.2% (70/109) on non-media questions. When compared to the performance of all national orthopaedic surgery residents in 2021, ChatGPT Plus performed at the level of an average PGY3.

Discussion: ChatGPT Plus is able to pass the OITE with an overall score of 61.2%, ranking at the level of a third-year orthopaedic surgery resident. It provided logical reasoning and justifications that may help residents improve their understanding of OITE cases and general orthopaedic principles. Further studies are still needed to examine their efficacy and impact on long-term learning and OITE/ABOS performance.

The authors have no funding, financial relationships, or conflict of interest to disclose.

Disclosure: The **Disclosure of Potential Conflicts of Interest** forms are provided with the online version of the article (<http://links.lww.com/JBJSOA/A576>).

Copyright © 2023 The Authors. Published by The Journal of Bone and Joint Surgery, Incorporated. All rights reserved. This is an open access article distributed under the terms of the [Creative Commons Attribution-Non Commercial-No Derivatives License 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0/) (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

Introduction

The rise of artificial intelligence (AI) in the last decade has revolutionized numerous domains, including health care, bringing a paradigm shift in diagnosis, treatment, and patient care across various specialties¹⁻⁴. Recent strides have been made in using AI algorithms, specifically OpenAI's language model "ChatGPT," as promising educational tools for medical training, yet their proficiency remains under continuous scrutiny^{1,5-7}. ChatGPT has demonstrated an unprecedented ability to understand, generate, and contextualize human language⁸⁻¹⁰, making it a focal point of interest for this study.

Standardized examinations have long been considered a cornerstone in measuring cognitive competency and academic achievement¹¹⁻¹⁴. Their fixed nature and predetermined scoring methods offer a consistent yardstick for gauging intellectual acumen across diverse demographics. Consequently, the performance of AI in this context presents a rich yet unexplored terrain for quantifying AI's understanding of complex cognitive tasks and simulating human-like problem-solving skills. The concept of AI superseding human ability in specialized areas, such as medical education, remains contentious¹⁵. In recent studies assessing the performance of ChatGPT on the United States Medical Licensing Examination, Gilson et al. found that ChatGPT achieved passing grades on Step 1 and Step 2 examinations⁶, and Kung et al. found similar results when assessing Step 1, Step 2CK, and Step 3⁷. Humar et al. found that ChatGPT performed at the level of a first-year plastic surgery resident on the Plastic Surgery In-service Training Examination⁵.

Nevertheless, the orthopaedic realm presents unique challenges owing to its procedural, radiographical, and clinical intricacies³. Every year, US orthopaedic surgery residents take the Orthopaedic Surgery In-Training Examination (OITE), which is overseen by the American Academy

of Orthopaedic Surgeons (AAOS)^{11,16,17}. Introduced in 1963, the OITE was the first standardized examination implemented in the medical field^{11,16}. The OITE assesses residents' knowledge across 10 levels of competency, ranging from basic science to surgical concepts¹⁷. It requires an increasing familiarity with orthopaedic principles and critical thinking skills to apply this knowledge in scenarios that require medical decision-making.

To our knowledge, no study has evaluated the performance of AI on these complex context-dependent OITE questions. The aim of this study is to evaluate the ability of ChatGPT to apply its comprehensive knowledge and human-like problem solving skills to pass the highly specialized OITE.

Materials and Methods

OITE 2021 Practice Test

ChatGPT is an AI language model that was developed by OpenAI (2022). Its latest version has neither real-time internet access nor the ability to update its training data beyond its informational cut-off date of September 2021^{9,18}. Owing to its inability to provide the most current data for some questions, the 2021 OITE practice test was used to avoid underperformance. It was accessed using the AAOS-ResStudy website (<https://www.aaos.org/education/examinations/ResStudy>) in May 2023. All 213 OITE web-based questions were retrieved. Similar to the official OITE, content domains and distributions included were basic science (11%), foot and ankle (10%), hand and wrist (10%), hip and knee (14%), oncology (9%), pediatrics (13%), shoulder and elbow (10%), spine (8%), sports medicine (5%), and trauma (12%)¹⁹.

ChatGPT Testing

For each OITE question, 2 independent reviewers—one orthopaedic surgery postdoctoral research fellow and a final year

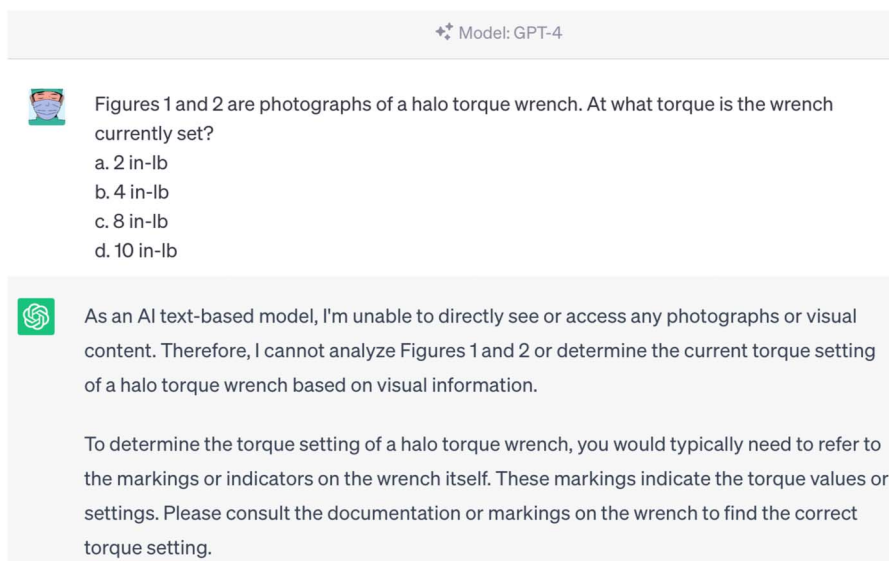


Fig. 1
An example of an excluded media-containing question that relies purely on the photographs and cannot be rationalized by ChatGPT Plus (version 4.0).

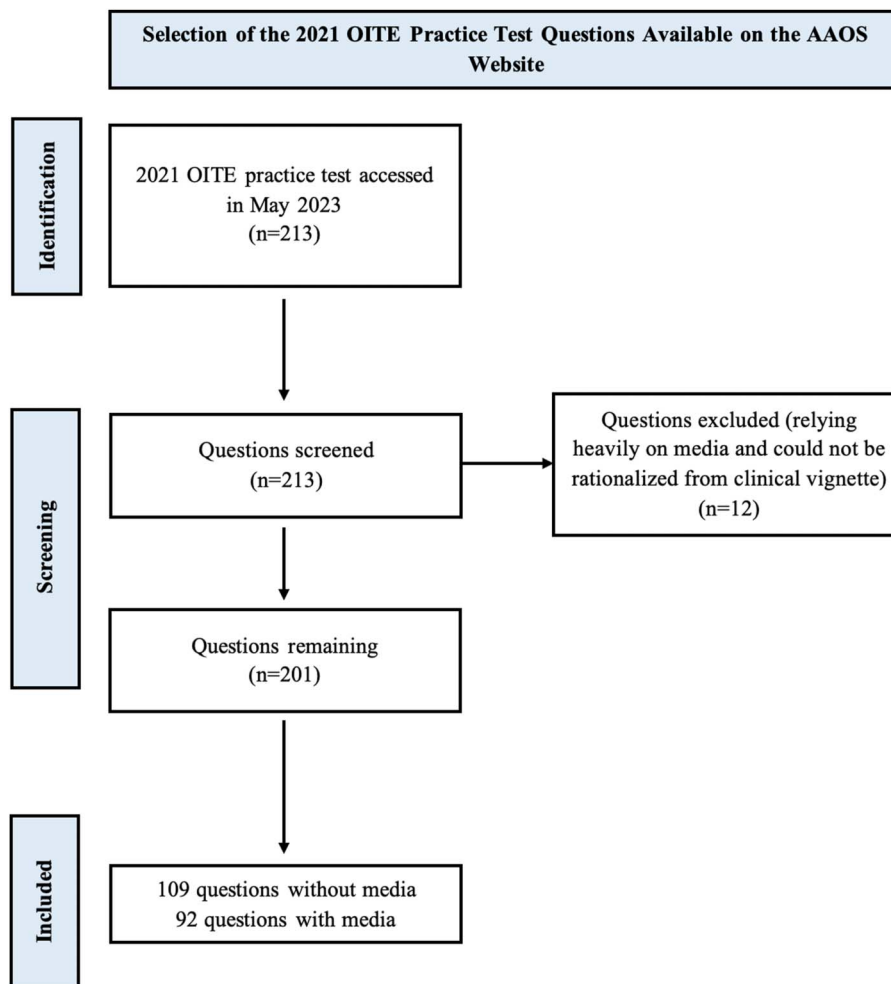


Fig. 2

A flow diagram outlining the selection of the 2021 OITE practice test questions available on the AAOS website. AAOS = American Academy of Orthopaedic Surgeons and OITE = Orthopaedic Surgery In-Training Examination.

medical student—copied and pasted the question stem and its associated response options into ChatGPT Plus (version 4.0, May 24 version) (<https://chat.openai.com>) and then recorded the ChatGPT-generated answer into a separate Microsoft Excel document 2023 (Version 16.73; Microsoft). This was done to ensure that the data was consistent and unlikely to be skewed by a single observer's biases to accurately gauge ChatGPT's performance.

Because ChatGPT is a language model that only accepts text input, all media-containing questions were flagged and carefully examined to avoid underestimating ChatGPT's performance. Twelve OITE media-containing questions that relied exclusively on images (clinical pictures, radiographs, magnetic resonance imaging, and computed tomography scans) and could not be rationalized from the clinical presentation were excluded (Fig. 1). Media-containing questions that could be rationalized from the information provided in the stem were included. To mitigate any potential for confounding variables, all inputted question stems and responses were formatted exactly as they seemed on the AAOS website. The stem included a clinical

presentation and question, followed by the answer choices in a multiple-choice format. Given ChatGPT's ability to learn from previous prompts, a new chat was created for each question to avoid altering raw performance. A new chat initiates a novel conversation with ChatGPT, in a separate tab, eliminating any learning that may have occurred because of the previous inquiry.

Statistical Analysis

Cohen Kappa coefficient was used to examine the agreement of ChatGPT-generated responses between the 2 independent observers. Scores from both independent observers were used to calculate the mean ChatGPT Plus performance for each question, and the means were used for all statistical analyses. Descriptive statistics were used to summarize the performance (% correct) of ChatGPT Plus. Pearson χ^2 was used to compare the number of correct media-containing and nonmedia-containing questions. The 2021 OITE technical report's norm table was used to compare ChatGPT Plus' performance on the OITE to the national performance of orthopaedic surgery residents for that same year. A p-value of 0.05 was set to

TABLE 1 Summary Table of the 201 Included Orthopaedic Surgery In-Training Examination Questions, With Performance Stratified by Media Presence in the Question Stem

	Number (n)	Correct (n)	Incorrect (n)	Performance (%)	p-value*
OITE questions	201	123	78	61.2%	—
OITE questions with media	92	53	39	57.6%	0.416
OITE questions without media	109	70	39	64.2%	

*p-value calculated using Pearson χ^2 . OITE = Orthopaedic Surgery In-Training Examination.

determine statistical significance. Statistical tests and analyses were performed using R software (Version 4.3.0.; R Foundation for Statistical Computing).

Results

OITE Questions


Two hundred one questions were evaluated by ChatGPT Plus (Fig. 2). Ninety-two media-containing questions, which correspond to 45.8% of the 201 questions, were included. All

ChatGPT-generated answers were marked as either correct or incorrect. Only one question remained inconclusive because ChatGPT thought there were 2 possible answers from the choices given. Both suggested answers were wrong, and therefore, this ChatGPT-generated answer was marked as incorrect.


ChatGPT Performance

ChatGPT had an overall score of 61.2% (123/201). Excellent agreement was observed between raters for the 201 ChatGPT-

Model: GPT-4

 A 9-year-old female presents with a 1-month history of pain in the right knee. Plain radiographs show a well-circumscribed, 1-cm lytic lesion in the tibial epiphysis. What is the most likely diagnosis?

- Giant cell tumor of bone
- Clear cell chondrosarcoma
- Giant cell tumor of tendon sheath
- Chondroblastoma

 This clinical scenario is consistent with a chondroblastoma.

Chondroblastomas are rare, benign bone tumors that account for less than 1% of all primary bone tumors. They occur most commonly in children and young adults, typically at the ends of long bones. The knee (the distal femur and proximal tibia) is a frequent location. The classic radiographic appearance of a chondroblastoma is a lytic, well-circumscribed lesion in the epiphysis or apophysis of a long bone.

So the answer is (d) Chondroblastoma.

Let's quickly go through the other options:

- Giant cell tumors of the bone are typically found in adults aged 20-40 and are most often found around the knee, but they are typically located in the metaphysis or epiphysis, extending to the subchondral bone. They are rarely found in children.
- Clear cell chondrosarcoma is a very rare variant of chondrosarcoma and is generally seen in older adults.
- Giant cell tumor of tendon sheath is a type of soft tissue tumor and does not form in bone, so this option does not fit with the radiographic findings.

Fig. 3

An example of a question prompt entry and correct response on ChatGPT Plus (version 4.0).

generated responses, with a Cohen Kappa coefficient of 0.947 (p-value = 0.947). ChatGPT's score was 64.2% (70/109) on nonmedia questions (Table I, Fig. 3). Among the media-related answered questions, 57.6% (53/92) of questions were rationalized from the stem and answered correctly (Table I, Fig. 4). There was no statistically significant difference between ChatGPT's performance on questions with and without media (p = 0.416) (Table I). When compared with the performance of all national orthopaedic surgery residents in 2021, ChatGPT Plus performed at the level of the average PGY3 (Table II).

Discussion

This is the first study to assess the performance of the AI language model ChatGPT on the OITE. It included all 2021 OITE questions available on the AAOS ResStudy website. ChatGPT Plus performed well at the level of the average PGY-3, with a satisfactory score of 61.2% on the OITE. When compared with the current literature regarding ChatGPT's performance on standardized examinations⁵⁻⁷, this overall score on the 2021 OITE is an impressive accomplishment for such a complex

medical examination. This exceptional performance may be attributable to this study using ChatGPT Plus (version 4.0) instead of previous versions including ChatGPT 3.5 and earlier. This version was chosen because of its superior performance and consistency despite its monthly subscription cost and its longer processing time to generate a response.

When stratifying questions based on media inclusion, there was no statistically significant difference between the rates of correct responses. ChatGPT demonstrated the potential to rationalize using the clinical vignette and make a reasonable choice despite missing key information. This is in line with current teachings, which suggest that clinical diagnoses and medical decisions, just like multiple-choice answers, rely heavily on the provided history.

The orthopaedic postdoctoral research fellow and final-year medical student used the same method and question format when inputting their queries to ChatGPT. Yet, comparing the generated texts between the 2 reviewers, 2 different responses were provided by ChatGPT in answering the same OITE questions. Although this may ensure originality, this lack of reproducibility may impede reliability if used for the

🔄 Model: GPT-4





Figure 1 shows failure of lumbopelvic fixation surgery to treat an adult spinal deformity. Which of the following most likely contributed to this failure?

- Large mismatch between pelvic incidence (PI) and lumbar lordosis (LL)
- Use of a synthetic allograft
- Small mismatch between PI and LL
- Inadequate correction of the coronal balance



Given the choices, the most likely contributing factor for the failure of lumbopelvic fixation surgery in the case of an adult spinal deformity would be:

- Large mismatch between pelvic incidence (PI) and lumbar lordosis (LL)

The PI-LL mismatch is an important parameter to consider in spinal deformity surgery. Pelvic incidence is a constant anatomical parameter that should ideally be matched by lumbar lordosis to maintain a balanced spine. If there is a large mismatch, it could lead to an inappropriate surgical correction and subsequently a failure of the lumbopelvic fixation.

The choice b, "Use of a synthetic allograft," would not necessarily lead to failure unless there were other complications such as infection, non-union, or rejection of the synthetic material. Similarly, the choice c, "Small mismatch between PI and LL," would actually be ideal as it implies that the spine is more likely to be in balance.

The choice d, "Inadequate correction of the coronal balance," could also contribute to surgical failure. However, the question and the given image specifically mentioned lumbopelvic fixation, which usually focuses on sagittal alignment (anteroposterior view). Therefore, without additional information, choice a is more likely.

But remember, in real clinical cases, a combination of factors can often contribute to surgical failure, and multiple variables should be taken into account.

Fig. 4

An example of a radiograph-containing question prompt entry and correctly rationalized response on ChatGPT Plus (version 4.0).

TABLE II Summary Table of All National Orthopaedic Surgery Residents' Average Performance on the 2021 Orthopaedic Surgery In-Training Examination Per Postgraduate-Year Level and ChatGPT Plus Performance, Ranked by Increasing Performance

Level	Number (n)	Mean Score (/263)	Performance (%)
PGY1	965	132.12	50%
PGY2	1,032	145.44	55.5%
PGY3	1,011	159.29	61%
ChatGPT Plus*	2	123*	61.2%
PGY4	1,015	167.6	64%
PGY5	963	175.02	67%

*ChatGPT Plus (version 4.0) mean score and performance was calculated by 2 independent reviewers over a total of 201 questions. PGY = Postgraduate Year.

purposes of medical education. Upon examination of the data, we found that ChatGPT selected different answer choices for 4 of its incorrect responses. For that reason, the Cohen Kappa coefficient was calculated. It demonstrated excellent agreement of ChatGPT-generated choices between reviewers despite the variability, suggesting that although ChatGPT Plus generates different justifications, it often results in the same final answer because of its reliance on key orthopaedic principles.

We acknowledge a few potential limitations of this study. First, ChatGPT is a text-based AI and is unable to view or analyze images^{18,20}. This is a key limitation because most medical specialties, such as orthopaedic surgery, necessitate physical examination and imaging to obtain a complete clinical picture and subsequently reach a diagnosis. However, as shown in this study, ChatGPT Plus always provides the reader with general information about a given topic or a justification to support its claims (Figs. 3 and 4) and occasionally describes the method that would allow the reader to make an informed decision (Fig. 1). Second, as an AI model with a knowledge cut-off of September 2021, ChatGPT does not have access to post-2021 or real-time data^{9,18}. It is thus unable to provide us with the latest updates or use newer references in its response. This is further restricted by ChatGPT's use of publicly available external information which, more often than not, excludes high-impact journals because of subscription fees. Thus, ChatGPT may give out information that is inaccurate or out of date^{9,18,21}. In addition, it is often misleadingly quick in generating "plausible-sounding but incorrect or nonsensical answers," as warned the OpenAI ChatGPT founder^{9,18}. It is just a matter of time before OpenAI developers render their newly incorporated real-time internet access feature available to the public¹⁸, an addition that will potentially enhance the efficacy and reliability of ChatGPT-4. Because ChatGPT and other language models can at times fabricate information to support its claims when providing responses, users must have some prior level of knowledge and

understanding of the subject matter to avoid misinformation. Despite this risk, the interesting and detailed responses provided using ChatGPT can enhance understanding among informed users who can cross reference ChatGPT's responses with reputable, published sources.

Multiple study tools simply provide residents with a "correct/incorrect" feedback, omitting the rationale behind the answer. ChatGPT is able to bridge that gap because it instantly generates a well-supported response with detailed explanations. In addition, given its ability to encode and save information for future use, residents can, if need be, engage in a discussion about the prompt or even ask ChatGPT for clarification or topic simplification. Hence, with the increasing digitization of medical education, integrating AI language models, such as ChatGPT, into training curriculums may offer an efficient and sustainable way to enhance orthopaedic training and augment learning outcomes, although further studies are still needed to examine their efficacy and impact on long-term learning and OITE/ABOS performance.

It is important to note that, despite showing its ability to answer a wide range of questions, ChatGPT might not always perform as well as human test-takers on standardized examinations; in fact, this study shows that PGY-4 and PGY-5 residents performed better than the AI model. This might be because of the fact that although ChatGPT can process the information it has been trained on, it does not truly "understand" its context. Standardized tests often rely on implicit knowledge, worldviews, and nuanced understandings that may not be explicitly stated in the question stem. Pattern recognition embedded in ChatGPT cannot replace the residents' gestalt understanding of orthopaedic concepts or their intuitive leaps based on previous clinical and surgical experiences. These qualities are inherent to human cognition and make standardized testing a complex arena for AI.

Our study is the first to assess the performance of OpenAI's now widely available language model "ChatGPT" on the OITE. With an overall score of 61.2%, ChatGPT has demonstrated an ability to assimilate, reason through clinical problems, and reproduce knowledge in the field of orthopaedic surgery equivalent to a third-year orthopaedic surgery resident. More importantly, its ability to provide logical reasoning and thorough explanations of answer choices may help residents grasp evidence-based information and improve their understanding of OITE cases and general orthopaedic principles. ■

Diane Ghanem, MD¹
 Oscar Covarrubias, BS²
 Micheal Raad, MD¹
 Dawn LaPorte, MD, FAOA¹
 Babar Shafiq, MD, FAOA¹

¹Department of Orthopaedic Surgery, The Johns Hopkins Hospital, Baltimore, Maryland

²School of Medicine, The Johns Hopkins University, Baltimore, Maryland

E-mail address for D. Ghanem: dghanem1@jh.edu

References

1. Bi AS. What's important: the next academic—ChatGPT AI? *J Bone Joint Surg.* 2023;105(11), 893-5.
2. Dergaa I, Chamari K, Zmijewski P, Ben Saad H. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. *Biol Sport.* 2023;40(2):615-22.
3. Bernstein J. Not the last word: ChatGPT can't perform orthopaedic surgery. *Clin Orthop Relat Res.* 2023;481(4):651-5.
4. Mogali SR. Initial impressions of ChatGPT for anatomy education. *Anat Sci Educ.* 2023;10.1002/ase.2261.
5. Humar P, Asaad M, Bengur FB, Nguyen V. ChatGPT is equivalent to first year plastic surgery residents: evaluation of ChatGPT on the plastic surgery in-service exam. *Aesthet Surg J.* 2023;43(12):NP1085-NP1089.
6. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ.* 2023;9:e45312.
7. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health.* 2023;2(2):e0000198.
8. Ollivier M, Pareek A, Dahmen J, Kayaalp ME, Winkler PW, Hirschmann MT, Karlsson J. A deeper dive into ChatGPT: history, use and future perspectives for orthopaedic research. *Knee Surg Sports Traumatol Arthrosc.* 2023;31(4):1190-2.
9. OpenAI. GPT-4 technical report. 2023 Available at: <https://cdn.openai.com/papers/gpt-4.pdf>. Accessed May 17, 2023.
10. The Lancet Digital Health. ChatGPT: friend or foe? *Lancet Digit Health.* 2023; 5(3):e102.
11. Marsh JL, Hruska L, Mevis H. An electronic orthopaedic in-training examination. *J Am Acad Orthop Surg.* 2010;18(10):589-96.
12. Silvestre J, Zhang A, Lin SJ. Analysis of references on the plastic surgery in-service training exam. *Plast Reconstr Surg.* 2016;137(6):1951-7.
13. DePasse JM, Haglin J, Eitorai AEM, Mulcahey MK, Ebersson CP, Daniels AH. Orthopedic in-training examination question metrics and resident test performance. *Orthop Rev (Pavia).* 2017;9(2):7006.
14. Fritz E, Bednar M, Harrast J, Marsh JL, Martin D, Swanson D, Tornetta P, Van Heest A. Do orthopaedic in-training examination scores predict the likelihood of passing the American board of orthopaedic surgery part I examination? An update with 2014 to 2018 data. *J Am Acad Orthop Surg.* 2021;29(24):e1370-e1377.
15. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT—reshaping medical education and clinical management. *Pak J Med Sci.* 2023;39(2):605-7.
16. Mankin HJ. The orthopaedic in-training examination (OITE). *Clin Orthop Relat Res.* 1971;75:108-16.
17. LaPorte DM, Marker DR, Seyler TM, Mont MA, Frassica FJ. Educational resources for the orthopedic in-training examination. *J Surg Educ.* 2010;67(3): 135-8.
18. OpenAI. Introducing ChatGPT. Available at: <https://openai.com/blog/chatgpt>. Accessed May 27, 2023.
19. American Academy of Orthopaedic Surgeons. Orthopaedic in-training examination (OITE) technical report 2021. Available at: <https://www.aaos.org/globalassets/education/product-pages/oite/oite-2021-technical-report.pdf>. Accessed October 6, 2023.
20. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, Rizzo C. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health.* 2023;11:1166120.
21. Flanagan A, Bibbins-Domingo K, Berkwits M, Christiansen SL. Nonhuman “authors” and implications for the integrity of scientific publication and medical knowledge. *JAMA.* 2023;329(8):637-9.