JB&JS
OPEN ACCESS

AMERICAN ORTHOPAEDIC ASSOCIATION

AOA Critical Issues in Education

Evaluating ChatGPT Performance on the Orthopaedic In-Training Examination

Justin E. Kung, MD, Christopher Marshall, BS, Chase Gauthier, BS, Tyler A. Gonzalez, MD, MBA, and J. Benjamin Jackson III, MD, MBA

Investigation performed at Prisma Health-Midlands University of South Carolina School of Medicine, Columbia, South Carolina

Background: Artificial intelligence (AI) holds potential in improving medical education and healthcare delivery. ChatGPT is a state-of-the-art natural language processing AI model which has shown impressive capabilities, scoring in the top percentiles on numerous standardized examinations, including the Uniform Bar Exam and Scholastic Aptitude Test. The goal of this study was to evaluate ChatGPT performance on the Orthopaedic In-Training Examination (OITE), an assessment of medical knowledge for orthopedic residents.

Methods: OITE 2020, 2021, and 2022 questions without images were inputted into ChatGPT version 3.5 and version 4 (GPT-4) with zero prompting. The performance of ChatGPT was evaluated as a percentage of correct responses and compared with the national average of orthopedic surgery residents at each postgraduate year (PGY) level. ChatGPT was asked to provide a source for its answer, which was categorized as being a journal article, book, or website, and if the source could be verified. Impact factor for the journal cited was also recorded.

Results: ChatGPT answered 196 of 360 answers correctly (54.3%), corresponding to a PGY-1 level. ChatGPT cited a verifiable source in 47.2% of questions, with an average median journal impact factor of 5.4. GPT-4 answered 265 of 360 questions correctly (73.6%), corresponding to the average performance of a PGY-5 and exceeding the corresponding passing score for the American Board of Orthopaedic Surgery Part I Examination of 67%. GPT-4 cited a verifiable source in 87.9% of questions, with an average median journal impact factor of 5.2.

Conclusions: ChatGPT performed above the average PGY-1 level and GPT-4 performed better than the average PGY-5 level, showing major improvement. Further investigation is needed to determine how successive versions of ChatGPT would perform and how to optimize this technology to improve medical education.

Clinical Relevance: AI has the potential to aid in medical education and healthcare delivery.

Ethical approval was not sought for this study because it consists of nonidentifying and anonymous clinical data that exposed minimal risk to the patients.

Disclosure: The **Disclosure of Potential Conflicts of Interest** forms are provided with the online version of the article (http://links.lww.com/JBJSOA/A553).

Introduction

Artificial intelligence (AI) holds great potential in improving medical education and healthcare delivery[1]. ChatGPT, a state-of-the-art language model chatbot publicly released by OpenAI in November of 2022, has demonstrated impressive performance for various examinations, such as the Scholastic Aptitude Test (SAT), Uniform Bar Exam, and MBA degree examination, often scoring in the 90th percentile[2,3]. When recently evaluated on the United States Medical Licensing Examination (USMLE) Step examinations, a comprehensive test for medical students and graduates, ChatGPT answered approximately 60% of questions correctly and was only marginally better than the passing threshold[4]. However, ChatGPT was able to show significant improvement in a period of months, suggesting that it may not be far off from potentially serving as an educational or clinical decision-making tool[5].

In recent studies, ChatGPT has been evaluated on multiple subspecialty training examinations. In the Plastic Surgery In-Service Examination, ChatGPT was noted to perform at the level of a first-year resident[6]. In the Dermatology Specialty Certificate Examination (SCE), ChatGPT had an overall score of 63.1% but GPT-4 showed a substantial improvement to 90.5%[7]. To the best of our knowledge, ChatGPT has not yet been evaluated specifically on the Orthopaedic In-Training Examination (OITE) and whether ChatGPT can provide verifiable sources to support its answers. In this study, we explore how well ChatGPT performs on a standardized examination in the orthopedic surgical specialty, to provide a benchmark and perhaps determine if it has the competency necessary to serve as an educational tool in this field.

The objectives of this study are (1) to determine how well ChatGPT performs on the OITE, a key evaluation tool for orthopedic surgery residents; (2) to compare the performance of version 3.5 to the more advanced GPT-4 model; and (3) to evaluate how often ChatGPT cites a verifiable source for its information. This investigation aims to evaluate ChatGPT and GPT-4 as potential educational aid by determining if it has the necessary knowledge base to provide accurate information to residents and students in orthopedics.

Methods

The OITE 2020, 2021, and 2022 were used to evaluate the performance of ChatGPT. Questions were accessed through ResStudy, an educational tool released by the American Academy of Orthopaedic Surgeons (AAOS)[8]. ChatGPT only accepts textual input in its current form; thus, questions with images were excluded from the study. Because prompt engineering has been shown to have a significant impact on large language models (LLMs) output, a zero-prompting approach was used. In our zero-prompting approach, ChatGPT was only given the information that would be available to a person taking the OITE, that is, the question and answer choices, without any additional information. Questions were formatted as they would appear on the OITE; with each question block being followed by 4 multiple choice answers on different lines.

Two different AI models were used in this study: ChatGPT 3.5 and GPT-4. ChatGPT 3.5 and GPT-4 are similar in which both are LLMs developed by OpenAI and but they differ in key areas, including model size and computational ability[3,9]. The 2 models were evaluated in July 2023. Each model was presented with the same set of OITE questions, and their responses were recorded for analysis. For each question, the multiple choice answers selected by ChatGPT were manually reviewed and compared with the corresponding correct answer. To compare the performances of ChatGPT 3.5 and GPT-4, we calculated the percentage of correct answers for each model. To understand its relative performance, the percentage of correct answers for ChatGPT 3.5 and GPT-4 were compared with the percentage of correct answers for orthopedic surgery residents at different years of training at Accreditation Council for Graduate Medical Education-accredited programs, as described by the OITE 2020, 2021, and 2022 technical reports provided by the AAOS[10,11]. Finally, ChatGPT 3.5 and GPT-4 were asked to cite a source for the answer they provided. These answers were categorized as journal articles, books, or websites. The sources were verified as being real through PubMed and Google searches of the source information. If ChatGPT cited a journal article, the journal's impact factor was recorded as well. Percentages of real sources cited, type of source provided, and average median impact factor were calculated.

Results

There were 215 questions on the OITE 2020 examination, 213 on the OITE 2021 examination, and 207 questions on the OITE 2022 examination (Table I). Of these, 129 (60%) OITE 2020, 107 (50.2%) OITE 2021, and 124 (59.9%) OITE 2022 questions were text-only questions and thus were included in the study. ChatGPT 3.5 answered 76 (58.9%) of 129, 56 (52.3%) of 107, and 64 (51.6%) of 124 questions correctly, with a total of 196 (54.3%) of 360 questions correct. ChatGPT 3.5's performance was similar to the average percentage correct for a postgraduate year (PGY-1) resident on the same examinations, which is 51%, 55%, and 51% for the OITE 2020, 2021, and 2022, respectively.

GPT-4 performed better than ChatGPT 3.5, answering 96 (74.4%) of 129, 78 (72.9%) of 107, and 91 (73.4%) of 124 questions correctly, with a total of 265 (73.6%) of 360 questions correct. GPT-4's performance is comparable with a PGY-5 performance on the examinations, with their percentage correct being and 68%, 71%, and 73% for OITE 2020, 2021, and 2022, respectively. Of note, GPT-4's percent correct on each OITE was higher than the corresponding passing standard for the American Board of Orthopaedic Surgery part I examination, which was determined to be 63%, 69, and 69% for OITE 2020, 2021, and 2022, respectively[10-12].

For the OITE 2020, 2021, and 2022 examinations, ChatGPT 3.5 provided a verifiable source for 51.2%, 47.7%, and 42.7% of the questions (Table II). An average of 66.0%, 16.8%, and 17.2% of sources cited by ChatGPT 3.5 were journal articles, books, and websites, respectively. Of the journal articles cited by ChatGPT 3.5, the average median impact factor of the journal cited was 5.4.

GPT-4 performed better than ChatGPT 3.5 in source information as well, providing a verifiable source for 85.3%,

TABLE I ChatGPT 3.5 and GPT-4 Results for OITE 2020, 2021, and 2022*

	OITE 2020	OITE 2021	OITE 2022	Overall
Total questions	215	213	209	637
Questions without images	129	107	124	360
Number correct (ChatGPT 3.5)	76	56	64	196
Percentage correct (ChatGPT 3.5)	58.9%	52.3%	51.6%	54.3%
Number correct (ChatGPT 4)	96	78	91	265
Percentage correct (ChatGPT 4)	74.4	72.9%	73.4%	73.6%
Corresponding ABOS passing level	63%	69%	69%	67%

*ABOS = American Board of Orthopaedic Surgery, and OITE = Orthopaedic In-Training Examination.

89.7%, and 88.7% of the OITE 2020, 2021, and 2022 questions, respectively. An average of 80.6%, 10.6%, and 8.8% of sources cited by GPT-4 were journal articles, books, and websites, respectively. The average median impact factor of the journal articles cited by GPT-4 was 5.2.

Discussion

ChatGPT, a state-of-the-art language model developed by OpenAI, has demonstrated remarkable achievements in various domains. One notable achievement is its passing score on the USMLE[4,5]. Although a higher standard will need to be set if it is ever to gain credibility as an educational or clinical decision-making tool, its current performance and rapid improvement suggest this standard may be feasible in due time. We sought to determine if ChatGPT could be used in a similar fashion for orthopedic residents by determining its competency of orthopedic knowledge through the OITE. Both ChatGPT 3.5 and GPT-4 performed well in the examination, with each performing at approximately the PGY-1 and PGY-5 level, respectively, on the 2020, 2021, and 2022 examinations. Overall, this study establishes a new benchmark for ChatGPT performance on the OITE.

Multiple studies have evaluated ChatGPT on subspecialty in-training examinations. ChatGPT performed at the level of a first-year resident on the Plastic Surgery In-Service Examination, answering approximately 55% of questions correctly[6,13]. On the Dermatology SCE, ChatGPT 3.5 scored 63.1% and GPT-4 scored 90.5%[7]. In the field of orthopedics, Lum et al. evaluated ChatGPT on Orthobullets (Lineage Medical) practice questions and reported a correct answer 47% of the time. In our study, we evaluated recent OITE questions, included GPT-4 performance, and performed an analysis on whether a verifiable source could be provided[14].

It is clear ChatGPT 3.5 and GPT-4 did not perform at the same level as they had on the Uniform Bar Exam or SAT but did perform at a similar level as they had on the USMLE Step examinations. One reason for this could be the way ChatGPT obtains its information. In the information ChatGPT draws its answers from, there may be conflicting sources of information about different topics. This conflicting information may be detrimental to ChatGPT's ability to answer a question correctly. In addition, ChatGPT is only trained on information up to September 2021, so new information that the medical examinations test on may not be available information that

TABLE II ChatGPT 3.5 and GPT-4 Verifiable Sources for OITE 2020, 2021, and 2022*

	OITE 2020	OITE 2021	OITE 2022	Overall
ChatGPT 3.5				
Verifiable Source	66 (51.2%)	51 (47.7%)	53 (42.7%)	170 (47.2%)
Journal article	85 (57.0%)	73 (68.9%)	77 (62.6%)	235 (65.8%)
Book	19 (12.8%)	18 (17.0%)	23 (18.7%)	60 (16.8%)
Website	24 (16.1%)	15 (14.2%)	23 (18.7%)	62 (17.2%)
GPT-4				
Verifiable Source	110 (85.3%)	96 (89.7%)	110 (88.7%)	316 (87.8%)
Journal article	109 (81.3%)	89 (76.1%)	108 (84.4%)	306 (80.7%)
Book	14 (10.4%)	13 (11.1%)	13 (10.2%)	40 (10.6%)
Website	11 (8.2%)	15 (12.8%)	7 (5.5%)	33 (8.7%)

*OITE = Orthopaedic In-Training Examination.

ChatGPT can draw from. Finally, in medicine specifically, there can be multiple potentially correct answers to a given question with only one best answer, which may cause difficulty for the AI when there is correct information supporting each answer. A potential solution to these issues would be to train an AI model on only peer-reviewed medical literature, such as that obtained from PubMed. In December 2022, a model known as BioMedLM 2.7B was developed using the GPT-2 architecture trained on PubMed abstracts[15]. It answered 50.3% of USMLE-type practice questions correctly, which surpassed other state-of-the-art models at the time. Perhaps applying this approach to the newer GPT-4 architecture as well as training the model to account for the level of evidence would strengthen its performance. This would potentially allow models to select the most correct answer, even when there are multiple sources with evidence supporting a particular answer choice.

In a previous study, 2 versions of ChatGPT, released only months apart, showed substantial improvement from 36.7% to 60% on the USMLE Step 1[5]. Similarly in our study, GPT-4 exhibited a considerable performance boost from version 3.5's average performance of 54.3% to GPT-4's average performance of 73.6% on the OITE, highlighting the rapid pace of development in AI models. Although the results are encouraging, it remains to be seen whether ChatGPT may be used to help physicians make medical decisions. Multiple studies have suggested that deep learning models for automated image analysis can be synergistic with clinicians resulting in superior predictions compared with those of clinicians alone. Krogue et al. developed a deep learning model for automatic identification and classification of hip fractures and determined that model usage could boost a resident's performance to that of an unaided fellowship-trained attending physician[16]. However, whether LLMs can help improve clinician performance has not been clearly shown. Furthermore, the demonstration of clinical knowledge is not merely measured by performance on multiple choice questions, and further study is needed to assess whether ChatGPT is able to answer clinical and management questions correctly in an open format. Overall ChatGPT and generative AI models have demonstrated significant improvements in their knowledge base, but further study is needed to fully elucidate if these models can be trusted as clinical decision-making tools.

An educational tool is effective if it can facilitate learning and if the information provided by the tool is correct and verifiable. Our study looked to answer this question to determine if ChatGPT can provide correct information and, therefore, be useful as an educational tool. We found GPT-4 performed well, scoring at a PGY-5 level on the 2020, 2021, and 2022 OITEs. In addition, GPT-4 cited a real source for 87.9% of questions and drew from influential journals, given the high average median impact factor for its cited journal article sources. Our results indicate GPT-4 may be useful as an educational tool for residents. Specifically, users may ask GPT-4 orthopedic-related questions and have a dialog with the AI surrounding the question. These conversations can help further enforce concepts by allowing users to explore topics in a conversational manner at their own pace. Users may ask ChatGPT follow-up questions regarding a particular topic, allowing them to gain more information beyond the static explanations provided in traditional learning material. Further development and investigation of AI models are needed to establish their reliability, but once credited, they may be able to provide orthopedic knowledge, clarify information, provide case-based learning, and promote evidence-based medicine[14].

Despite these results, there are several limitations to our study. First, the current version of ChatGPT is unable to analyze images, making it difficult to evaluate an essential skill for orthopedic surgeons. However, given the rapid progress in deep learning research, we anticipate that future AI models will incorporate image analysis capabilities. Second, we observed that ChatGPT occasionally provided different answers to the exact same question, raising concerns about the consistency of its performance. Finally, when examining instances where ChatGPT provided a verifiable source of information but still gave an incorrect answer, we found ChatGPT either cited outdated articles, articles with low levels of evidence, or drew incorrect conclusions based on certain article phrases, which were not representative of the article's overall conclusions. Therefore, even when ChatGPT cites a verifiable source, the information drawn from said source may be outdated or wrong entirely. Gilson et al. noted that when evaluated on USMLE-style questions, ChatGPT provided logical justifications for all of its answer choices but was still subject to logical errors[4]. Lum et al. noted that ChatGPT performance decreased as Buckwalter taxonomy level of recall increased to involve complex levels of interpretation and application of knowledge[14]. These logical errors and assertions of false facts are quite concerning and have even been termed as the *hallucination effect* that LLMs are susceptible to[17]. Although previous studies have demonstrated high concordance and insight within the model's explanations, errors in explanations are a valuable aspect to explore in future research[5].

AI is a growing technology that has been affecting different medical specialties at various speeds. Many fields in medicine have been investigating the role of AI in a clinical decision-making capacity. Owing to the development of advanced AI models capable of image-recognition tasks, radiologists are perhaps closest to deploying AI into their clinical work, and some have even posited that radiologists should lead AI initiatives[18]. In the realm of internal medicine, with continuous glucose monitoring, AI algorithms are used to determine the necessity for insulin dose adjustments remotely[19]. Although orthopedics has yet to largely integrate AI models for clinical decision-making, this is the direction the field of medicine seems to be headed. As for where we are now, our results show that AI tools are improving and demonstrate a solid knowledge base. Although this is only one aspect of delivering appropriate orthopedic care, it is feasible that AI, such as ChatGPT, will continue to evolve and may soon be able to contribute to medical decisions in surgical patients. Further research is needed to fully elucidate the capacity of ChatGPT to aid orthopedic surgeons in making clinical decisions. Owing to the rapidly changing standard set by AI, it is important for orthopedic surgeons to be involved in the integration of AI into this field and to guide it to a position where it can be used in providing excellent patient care. In its current form, ChatGPT demonstrated comparable knowledge with that of orthopedic residents, and with further advancement, may possibly be used in orthopedic medical education, patient education, and clinical decision-making. ■

Justin E. Kung, MD[1]
Christopher Marshall, BS[2]
Chase Gauthier, BS[1]
Tyler A. Gonzalez, MD, MBA[1]
J. Benjamin Jackson III, MD, MBA[1]

[1]Department of Orthopedic Surgery, Prisma Health-Midlands University of South Carolina, Columbia, South Carolina

[2]University of South Carolina School of Medicine, Columbia, South Carolina

E-mail address for J.E. Kung: Justin.Kung@prismahealth.org

References

1. St Mart JP, Goh EL, Liew I, Shah Z, Sinha J. Artificial intelligence in orthopaedics surgery: transforming technological innovation in patient care and surgical training. Postgrad Med J. 2022;99(1173):687-94.

2. Katz DM, Bommarito MJ, Gao S, Arredondo P. GPT-4 Passes the bar exam. Social Science Research Network. 2023. doi: 10.2139/ssrn.4389233.

3. GPT-4 technical report. Available at: https://cdn.openai.com/papers/gpt-4.pdf. Accessed April 20, 2023.

4. 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.

5. 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.

6. 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:sjad130. doi: 10.1093/asj/sjad130

7. Passby L, Jenko N, Wernham A. Performance of ChatGPT on dermatology specialty certificate examination multiple choice questions. Clin Exp Dermatol. 2023: llad197.doi: 10.1093/ced/llad197

8. ResStudy–Orthopaedic Exam Question Bank|American Academy of Orthopaedic Surgeons. Available at: https://www.aaos.org/education/examinations/ResStudy/. Accessed April 20, 2023.

9. OpenAI platform. Available at: https://platform.openai.com. Accessed July 14, 2023.

10. 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 April 20, 2023.

11. Orthopaedic in-training examination (OITE) technical report 2022. Available at: https://www.aaos.org/globalassets/education/product-pages/oite/oite-2022-technical-report-20230125.pdf. Accessed April 20, 2023.

12. Incrocci M. Orthopaedic In-Training Examination (OITE) Technical Report 2020. Available at: https://www.aaos.org/globalassets/education/product-pages/oite/oite-2020-technical-report_website.pdf. Accessed April 20, 2023.

13. Gupta R, Herzog I, Park JB, Weisberger J, Firouzbakht P, Ocon V, Chao J, Lee ES, Mailey BA. Performance of ChatGPT on the plastic surgery inservice training examination. Aesthet Surg J. 2023:sjad128. doi: 10.1093/asj/sjad128

14. Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery examination? Orthopaedic residents versus ChatGPT. Clin Orthop Relat Res. 2023;481(8):1623-30.

15. Stanford CRFM. Available at: https://crfm.stanford.edu/2022/12/15/pubmedgpt.html. Accessed April 20, 2023.

16. Krogue JD, Cheng KV, Hwang KM, Toogood P, Meinberg EG, Geiger EJ, Zaid M, McGill KC, Patel R, Sohn JH, Wright A, Darger BF, Padrez KA, Ozhinsky E, Majumdar S, Pedoia V. Automatic hip fracture identification and functional subclassification with deep learning. Radiol Artif Intell. 2020;2(2):e190023.

17. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, Moy L. ChatGPT and other Large Language Models are double-edged swords. Radiology. 2023;307(2): e230163.

18. Santomartino SM, Siegel E, Yi PH. Academic radiology departments should lead artificial intelligence initiatives. Acad Radiol. 2023;30(5):971-4.

19. Nomura A, Noguchi M, Kometani M, Furukawa K, Yoneda T. Artificial intelligence in current diabetes management and prediction. Curr Diab Rep. 2021; 21(12):61.