Title

Can Artificial Intelligence Pass the American Board of Orthopaedic Surgery Examination? Orthopaedic Residents Versus ChatGPT

Journal

Clinical Orthopaedics and Related Research®, 481(8)

ISSN

0009-921X

Author

Lum, Zachary C

Publication Date

2023-08-01

DOI

10.1097/corr.0000000000002704

Peer reviewed

Clin Orthop Relat Res (2023) 481:1623-1630
DOI 10.1097/CORR.0000000000002704

Clinical Orthopaedics and Related Research®
A Publication of The Association of Bone and Joint Surgeons®

Basic Research

Can Artificial Intelligence Pass the American Board of Orthopaedic Surgery Examination? Orthopaedic Residents Versus ChatGPT

Zachary C. Lum DO[1]

Received: 31 January 2023 / Accepted: 28 April 2023 / Published online: 23 May 2023

Abstract

Background Advances in neural networks, deep learning, and artificial intelligence (AI) have progressed recently. Previous deep learning AI has been structured around domain-specific areas that are trained on dataset-specific areas of interest that yield high accuracy and precision. A new AI model using large language models (LLM) and nonspecific domain areas, ChatGPT (OpenAI), has gained attention. Although AI has demonstrated proficiency in managing vast amounts of data, implementation of that knowledge remains a challenge.

Questions/purposes (1) What percentage of Orthopaedic In-Training Examination questions can a generative, pre-trained transformer chatbot (ChatGPT) answer correctly? (2) How does that percentage compare with results achieved by orthopaedic residents of different levels, and if scoring lower than the 10th percentile relative to 5th-year residents is likely to correspond to a failing American Board of Orthopaedic Surgery score, is this LLM likely to pass the orthopaedic surgery written boards? (3) Does increasing question taxonomy affect the LLM's ability to select the correct answer choices?

Methods This study randomly selected 400 of 3840 publicly available questions based on the Orthopaedic In-Training Examination and compared the mean score with that of residents who took the test over a 5-year period. Questions with figures, diagrams, or charts were excluded, including five questions the LLM could not provide an answer for, resulting in 207 questions administered with raw score recorded. The LLM's answer results were compared with the Orthopaedic In-Training Examination ranking of orthopaedic surgery residents. Based on the findings of an earlier study, a pass-fail cutoff was set at the 10th percentile. Questions answered were then categorized based on the Buckwalter taxonomy of recall, which deals with increasingly complex levels of interpretation and application of knowledge; comparison was made of the LLM's performance across taxonomic levels and was analyzed using a chi-square test.

Results ChatGPT selected the correct answer 47% (97 of 207) of the time, and 53% (110 of 207) of the time it answered incorrectly. Based on prior Orthopaedic In-Training Examination testing, the LLM scored in the 40th percentile for postgraduate year (PGY) 1s, the eighth percentile for PGY2s, and the first percentile for PGY3s, PGY4s, and PGY5s; based on the latter finding (and using a predefined cutoff of the 10th percentile of PGY5s as the threshold for a passing score), it seems unlikely that the LLM would pass the written board examination. The LLM's performance decreased as question taxonomy level increased (it answered 54% [54 of 101] of Tax 1 questions

The author certifies that there are no funding or commercial associations (consultancies, stock ownership, equity interest, patent/licensing arrangements, etc.) that might pose a conflict of interest in connection with the submitted article related to the author or any immediate family members.
All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research*® editors and board members are on file with the publication and can be viewed on request.
Ethical approval was not sought for the present study.

[1]Nova Southeastern University, Davie, FL, USA

Z. C. Lum ✉, Nova Southeastern University, 3200 South University Drive, Kiran C. Patel School of Osteopathic Medicine, Department of Faculty and Alumni Affairs, Davie, FL 33328, USA, Email: zacharylum@gmail.com

Wolters Kluwer

correctly, 51% [18 of 35] of Tax 2 questions correctly, and 34% [24 of 71] of Tax 3 questions correctly; p = 0.034).
Conclusion Although this general-domain LLM has a low likelihood of passing the orthopaedic surgery board examination, testing performance and knowledge are comparable to that of a first-year orthopaedic surgery resident. The LLM's ability to provide accurate answers declines with increasing question taxonomy and complexity, indicating a deficiency in implementing knowledge.
Clinical Relevance Current AI appears to perform better at knowledge and interpretation-based inquires, and based on this study and other areas of opportunity, it may become an additional tool for orthopaedic learning and education.

Introduction

Over the past decade, advances in machine learning, deep learning, and artificial intelligence (AI) have changed the way humans approach a wide variety of tasks and industries, ranging from manufacturing to consumer products. Deep learning from neural networks has improved precision and accuracy in identifying fractures and the manufacturer and model of orthopaedic implants, and it has many other medical uses [3, 4, 7, 9, 12]. Although these developments have made substantial contributions, their use requires considerable time, effort, and data specific to that area of interest. These types of AI can be considered domain specific. Because their training data are specific, their tasks and functions are also specific, meaning they cannot perform other functions outside their expertise; thus, they are not generalizable or multifunctional.

Large language models (LLMs), a type of machine learning, use vast amounts of text to analyze and synthesize its responses more naturally. It is also a nondomain or few-shot scenario, meaning a small amount of training data are used to execute that specific function, but the LLM can understand the request and process, analyze, and possibly use reasoning and chain of thought abilities to answer a broad range of questions. A new AI model using LLMs and nonspecific domain areas, called ChatGPT (OpenAI), has gained recent attention with its novel way to process information.

AI has become an increasingly important tool for medical education and fast access to data over many years, and includes computer-based models, virtual reality simulations, and personalized learning platforms [6, 7, 15]. As the capabilities of AI continue to advance, it is becoming increasingly important to regularly evaluate the competency of AI-powered tools. This evaluation is crucial to maintain high standards and prevent potential errors or biases, especially when addressing generative AI models that may demonstrate flawed reasoning or deliver misinformation that could harm patients or spread inaccurate information. Given the relatively limited understanding of this LLM's abilities in the domain of orthopaedic surgery knowledge, it is especially important to assess the accuracy of AI-powered tools in this field. By doing so, we can identify any shortcomings or areas for improvement and optimize the benefits of AI technology for healthcare providers and patients alike.

This study therefore sought to answer: (1) What percentage of Orthopaedic In-Training Examination (OITE) questions can a generative, pretrained transformer chatbot (ChatGPT) answer correctly? (2) How does that percentage compare with results achieved by orthopaedic residents of different levels, and if scoring lower than the 10th percentile relative to 5th-year residents is likely to correspond to a failing American Board of Orthopaedic Surgery (ABOS) score, is this LLM likely to pass the orthopaedic surgery written boards? (3) Does increasing question taxonomy affect the LLM's ability to select the correct answer choices?

Materials and Methods

Study Design and Setting

This was an experimental study using a commercially available LLM called ChatGPT. This LLM uses self-attention mechanisms and large amounts of training data to generate natural language responses to input text in conversional context. It is especially effective at handling long-range dependencies and creating coherent and contextually appropriate responses. Self-attention mechanisms are often used in natural language processing tasks such as language translation and text generation, where the model needs to understand the relationships between words in a sentence or a document. Long-range dependencies refer to the relationship between distant parts of a sequence of input data or text, and combined with self-attention allow accurate understanding and meaning of sentences that generate appropriate responses (Fig. 1). Additionally, it is a server-contained LLM, meaning that it cannot access data from the internet or perform search functions for new information. All responses are generated based on the abstract relationship between words in the neural network. This is different from other chatbots or domain-specific trained AI that allow online database searches or additional information access [9].

Question Set, Randomization, and Testing

A total of 400 questions based on the OITE were randomly selected from a publicly available lot of 3840 questions. Questions were generated using Orthobullets (Lineage

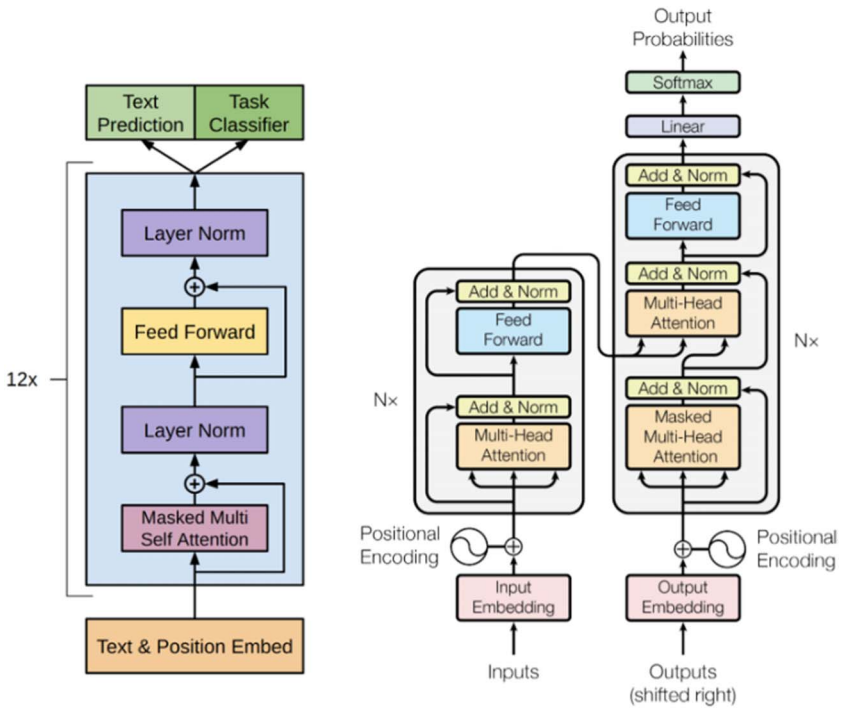

Fig. 1 This transformer architecture is the basis of how many large language models work (such as ChatGPT). Steps that occur when an LLM receives input data or a query are the following: (1) Input embedding. The relationship between the words is analyzed in a dense vector representation. (2) Multiheaded self-attention. The transformer block uses multiheaded self-attention to focus on varying parts of the inputs and understand their relationships. (3) Feed forward network. Output from self-attention goes through a feed-forward neural network to create a new abstract understanding by using complex mathematical functions to capture intricate patterns and relationships. (4) Normalization and residual connections. A deep neural network is created by repeating the normalization and residual connection components to process long text sequences and generate high-quality outputs for language tasks such as text generation, question answering, and translation [14].

Medical). A total of 48% (193 of 400) of the questions were excluded because they contained images, figures, tables, or charts, leaving 212 questions to be administered. Because this LLM chatbot is purely a text-based input program, questions that contained nontext-based data could not be entered into the program, and the chatbot was unable to analyze or interpret imaging, figure, or picture-based input data. All questions were entered into ChatGPTs text box, ensuring no duplicate questions were used. To reduce any memory retention bias, a new chat session was administered for each question. Memory retention via recurrent neural networks can occur when the LLM learns new information and subsequently applies the data to future data inputs and outputs (Fig. 2).

Questions were then entered into the chat session with a request that the LLM select an answer. If the LLM did not select a number response, or if it gave multiple responses, the question was re-entered with the prompt "Select the single best answer." If the LLM did not select an answer choice on the second request, the question was recorded as "did not answer" and the next question prompt was given. There were five questions for which the LLM could not provide a single best answer because multiple answer choices were repeated or no answer choices were given. These questions were excluded.

Primary and Secondary Study Outcomes

The primary study outcome was to ascertain the percentage of questions the LLM would answer correctly.

The secondary study outcomes were to compare its performance to that of orthopaedic residents of varying levels to ascertain whether the LLM would score highly enough to pass the ABOS written examination, and to determine

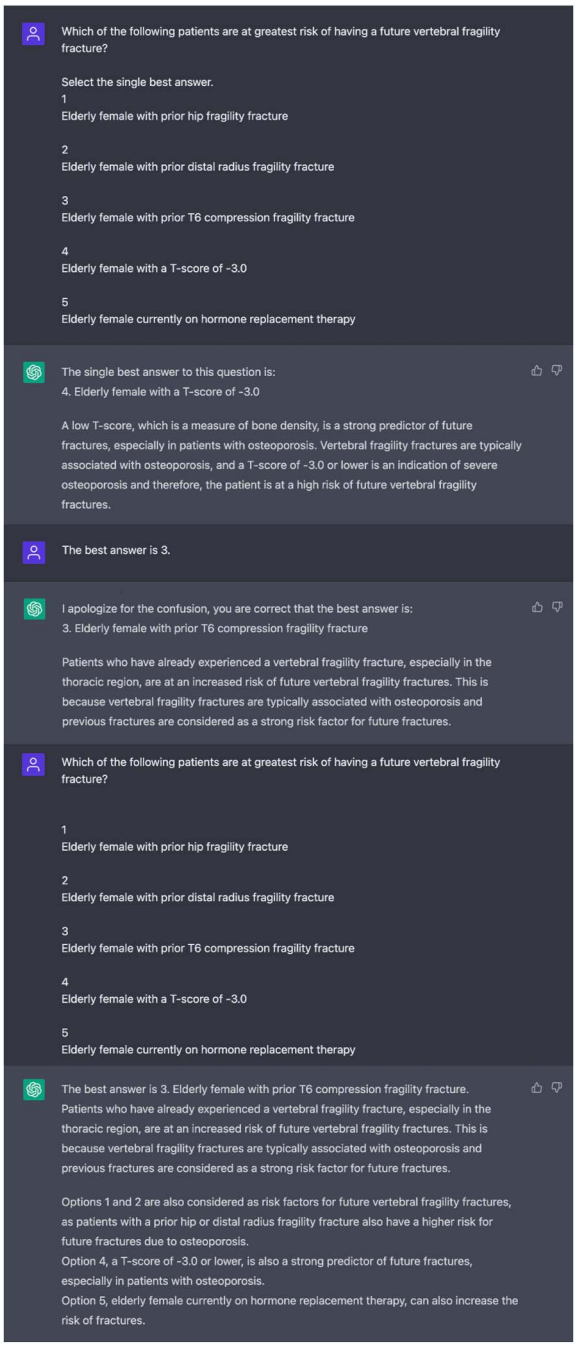

Fig. 2 This is a ChatGPT example of answering the question incorrectly, then learning the correct answer and subsequently answering the question correctly. This is an example of the LLM's recurrent neural network where information from prior inputs and outputs influences the current input and output.

whether the LLM's performance would decrease as question complexity increased. To compare the LLM's performance to that of residents in a typical year of OITE testing, this study extrapolated mean raw OITE scores averaged over 5 years from 2013 to 2017 for each postgraduate year (PGY) resident, including mean score, standard deviation, and calculated percentile based on each PGY in training [5]. To estimate whether it is likely that the LLM could pass the ABOS written examination, this study used previous study data suggesting an OITE score in less than 10th percentile is likely to be associated with later failure on the ABOS examination [13]. Finally, to answer the question about the LLM's performance against increasingly challenging levels of question taxonomy, this study used the Buckwalter taxonomic schema [2]. Questions were divided into three groups, where taxonomic level I (Tax 1) questions are defined as questions that test only recognition and recall (Fig. 3), Tax 2 questions evaluate comprehension and interpretation (Fig. 4), and Tax 3 questions ask about the application of knowledge (Fig. 5). Of the 207 questions, 49% (101 of 207) were considered Tax 1 (recognition and recall questions), 17% (35 of 207) were considered Tax 2 (comprehension and interpretation questions), and 34% (71 of 207) were Tax 3 questions (questions about the application of new knowledge). The LLM's performance was evaluated as a percentage of correct answers at each of those taxonomic levels.

Ethical Approval

Ethical approval was not sought for the present study.

Statistical Analysis

A chi-square test was used to ascertain whether the LLM's percentage of correct answers was different according to different taxonomic complexity, and a p value of < 0.05 was considered significant.

Results

Percentage of OITE Questions Answered Correctly

ChatGPT answered 47% (97 of 207) of questions correctly and 53% (110 of 207) of questions incorrectly. It did not respond to five questions (Fig. 6). These nonresponse questions were excluded from the final tally because multiple answer choices were chosen and the LLM could not respond with a "single best answer." Ultimately, these five questions were excluded because no single best-answer choice was selected.

Performance Comparison With That of Orthopaedic Residents

Based on prior OITE testing, ChatGPT scored in the 40th percentile for PGY1s, the eighth percentile for PGY2s, and

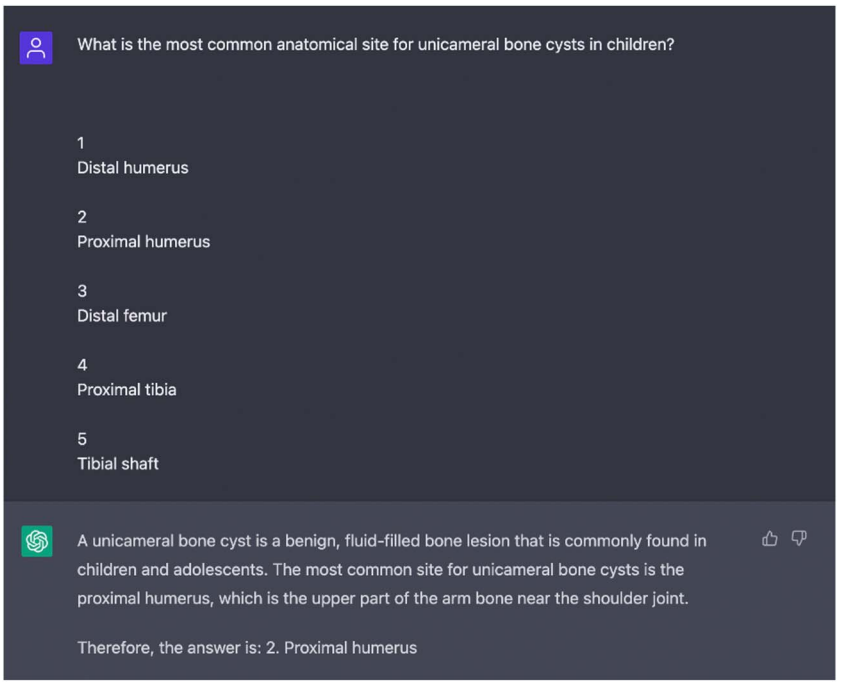

Fig. 3 This question-and-answer entry by ChatGPT is an example of a correct answer. The question type would be considered Taxonomy 1, recognition and recall.

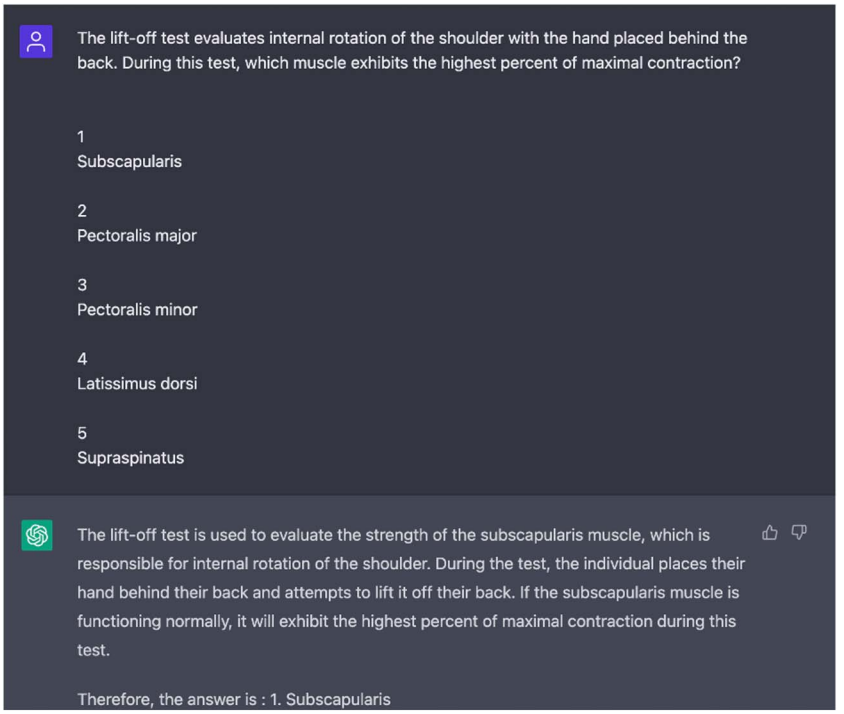

Fig. 4 This is another example of a question and response by the LLM. This question type would also be considered Taxonomy 2, comprehension and interpretation.

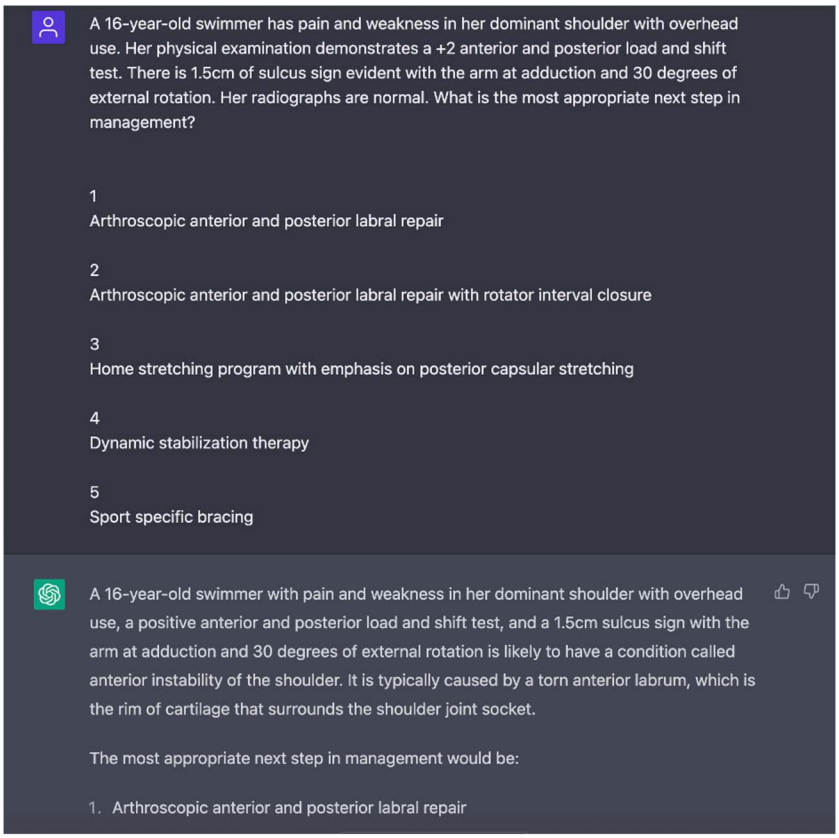

Fig. 5 Here is an example of an incorrect response provided by the LLM. The correct answer is 4, dynamic stabilization therapy. This question is considered Taxonomy level 3, application of knowledge and problem-solving.

the first percentile for PGY3s, PGY4s, and PGY5s [5]. Based on the predefined threshold of the 10th percentile of PGY5 scores as a passing grade, the LLM would not have passed the ABOS examination.

Performance in Relation to Increasingly Difficult Taxonomic Level

The LLM's performance decreased as question taxonomy level increased (it answered 55% [54 of 101] of Tax 1 questions correctly, 51% [18 of 35] of Tax 2 questions correctly, and 34% [24 of 71] of Tax 3 questions correctly; $p = 0.034$).

Discussion

AI has become more commonly used in medicine during the past several years. Potential applications in education, interpretation, and information management have expanded [12]. As newer AI tools continue to be developed, its competency must be checked, evaluated, and updated. Here, ChatGPT, an LLM chatbot, could correctly answer nearly half the questions on modern OITE-style examinations. Although this places it in the 40th percentile for an orthopaedic first-year resident, it was unlikely to pass the ABOS examination because it only scored in the first percentile of midlevel and upper-level residents, seemingly because of the lack of ability to apply knowledge to questions with a higher taxonomic level, suggesting an inability to apply the information that it "knows" in practical ways.

Limitations

The limitations of this study, specifically, are the lack of visual identification, interpretation, and integration with the questions. Nearly half of the questions contained an image, figure, or chart that resulted in exclusion of the question. Naturally, the real ABOS and OITE contain images, and many aspects of musculoskeletal care require interpretation and analysis of images, radiographs, and

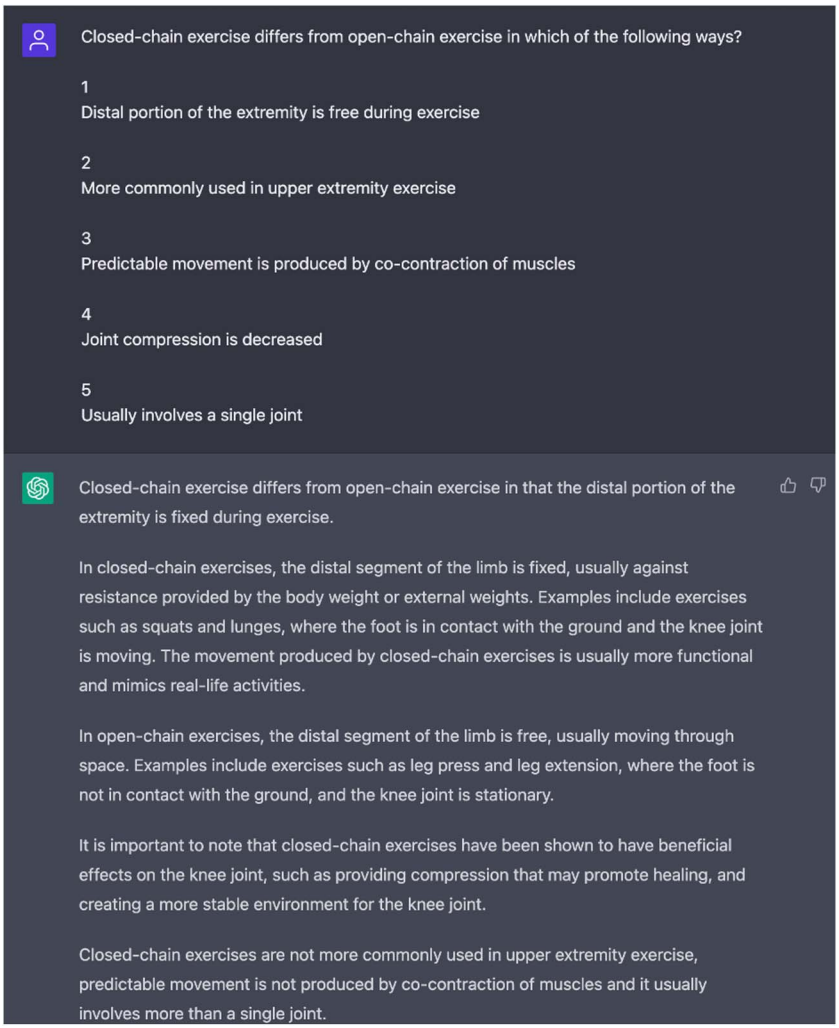

Fig. 6 Here is an example of a nonresponse by ChatGPT. After two queries of the question with direct redirecting, the LLM did not select a single best answer choice. Because of the modeling used, we omitted these questions. In this example of an omitted question, the response provided two answers, including when reprompted.

tactile feedback such as a physical examination. Additionally, images may have contained more questions that required higher application of knowledge or more challenging questions for the LLM that could have biased the results. Although images are an important part of orthopaedic surgery, this LLM's input is exclusively text. AI for image analysis is improving rapidly, and future iterations may be able to evaluate images. However, as a preliminary analysis, this study of text-based questions alone was sufficient to demonstrate the capacity of this LLM in this context, as well as its shortcomings. General limitations that apply to any AI model include the datasets they are trained on, which may incur, perpetuate, or even amplify existing societal biases or inequalities, and they could contain inaccurate or outdated information. Lastly, limitations specific to this LLM are based on its training using broad nonspecific information, and that it excels in specific fields of summation, translation, and text generation. However, it may not understand context or nuance-specific language in knowledge-specific areas that could lead to inaccurate or misleading responses.

Discussion of Key Findings

ChatGPT answered nearly half of the orthopaedic OITE-based questions correctly. This LLM scored in the 40th percentile for PGY1s, eighth percentile for PGY2s, and less than the first percentile for PGY3s, PGY4s, and PGY5s [5], meaning that it seems unlikely that this LLM would be able to

pass the written board certification examination. The reason for this poorer performance relative to midlevel and upper-level residents likely was because the LLM performed more poorly as the taxonomic complexity of the test questions increased. This suggests the model may have limitations in terms of its ability to integrate, synthesize, generalize, and apply factual knowledge in more-nuanced ways.

However, there are likely to be practical advantages to and applications of AI in this context. One benefit of AI is the ability to handle large amounts of data that can be quickly accessed as knowledge by the user. This study portrayed this clearly, because the LLM performed better at recognition and recall-type questions, as well as comprehension and interpretation, than problem-solving and application of knowledge. Others have indicated opportunities for AI to leverage big data to obtain insights and develop strategies for managing specific diseases, including opioid use disorders [1]. Another example of recognition and interpretation was offered by Liu et al. [10], in which AI and orthopaedic surgeons correctly identified a similar number of tibial plateau fractures (accuracy 0.91 versus 0.92). These use cases could improve efficiency and accuracy in diagnosis and treatment, ultimately leading to better patient outcomes. Other real-world implications of AI include the creation of educational resources to make them more accessible to the typical patient. A recent study found that ChatGPT was able to revise and simplify the writing of complex patient educational materials about spine surgery and joint replacement so they were readable at the fifth- to sixth-grade levels [8]. Another study suggested that educators could transition to more of a mentorship role, because AI may compile the best learning strategies from the best educators, allowing students and learners to improve their educational experiences independently and efficiently [11]. Additionally, AI can provide personalized learning experiences tailored to individual student needs and abilities. This may help improve student engagement and knowledge retention, leading to more effective learning, although it will take more research to determine whether, and to what degree, this is true.

Conclusion

Although ChatGPT likely would not have passed the ABOS written examination, it provided insightful and well-constructed explanations for the correct answers, and it achieved results consistent with the 40th percentile of PGY1 orthopaedic residents. Moreover, the model exhibited a learning capability when wrong answers were corrected, because it retained the corrected answer and applied it consistently throughout the chat box. Overall, these benefits indicate the potential of AI to assist and enhance medical education and healthcare in the future. The LLM demonstrated strengths in recall and interpretation but showed weaknesses in the application of knowledge. As advancements in AI technology continue, particularly in areas such as image-based recognition, interpretation, and specific-domain applications of knowledge, it will be interesting to see how this technology continues to improve and how it might best be applied in orthopaedic education.

References

1. Bharat C, Hickman M, Barbieri S, Degenhardt L. Big data and predictive modelling for the opioid crisis: existing research and future potential. *Lancet Digit Health*. 2021;3:e397-e407.
2. Buckwalter JA, Schumacher R, Albright JP, Cooper RR. Use of an educational taxonomy for evaluation of cognitive performance. *J Med Educ*. 1981;56:115-121.
3. Cohen M, Puntonet J, Sanchez J, et al. Artificial intelligence vs. radiologist: accuracy of wrist fracture detection on radiographs. *Eur Radiol*. Published online December 14, 2022. DOI: 10.1007/s00330-022-09349-3.
4. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med*. 2021;385:283-286.
5. Fritz E, Bednar M, Harrast J, et al. 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:e1370-e1377.
6. Guerrero DT, Asaad M, Rajesh A, Hassan A, Butler CE. Advancing surgical education: the use of artificial intelligence in surgical training. *Am Surg*. 2023;89:49-54.
7. Karnuta JM, Murphy MP, Luu BC, et al. Artificial intelligence for automated implant identification in total hip arthroplasty: a multicenter external validation study exceeding two million plain radiographs. *J Arthroplasty*. Published online March 7, 2022. DOI: 10.1016/j.arth.2022.03.002.
8. Kirchner G, Kim RY, Weddle JB, Bible JE. Can artificial intelligence improve the readability of patient education materials? *Clin Orthop Relat Res.* Published online April 28, 2023. DOI: 10.1097/CORR.0000000000002668.
9. Kung TH, Cheatham M, ChatGPT, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit. Health.* 2023;2:e0000198
10. 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;41:1158-1164.
11. Luo Q, Yang J. The artificial intelligence and neural network in teaching. *Comput Intell Neurosci*. 2022;2022:1778562.
12. Ramkumar PN, Kunze KN, Haeberle HS, et al. Clinical and research medical applications of artificial intelligence. *Arthroscopy*. 2021;37:1694-1697.
13. Swanson D, Marsh JL, Hurwitz S, et al. Utility of AAOS OITE scores in predicting ABOS part I outcomes: AAOS exhibit selection. *J Bone Joint Surg Am*. 2013;95:e84.
14. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Available at: https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. Accessed April 22, 2023.
15. Vedula SS, Ghazi A, Collins JW, et al. Artificial intelligence methods and artificial intelligence-enabled metrics for surgical education: a multidisciplinary consensus. *J Am Coll Surg*. 2022;234:1181-1192.