JB&JS

OPEN ACCESS

AMERICAN ORTHOPAEDIC ASSOCIATION

AOA Critical Issues in Education

Exploring the Performance of ChatGPT in an Orthopaedic Setting and Its Potential Use as an Educational Tool

Arthur Drouaud, BS, Carolina Stocchi, BS, Justin Tang, BS, Grant Gonsalves, BA, Zoe Cheung, MD,
Jan Szatkowski, MD, and David Forsh, MD

Investigation performed at George Washington University School of Medicine, Washington, DC

Introduction: We assessed ChatGPT-4 vision (GPT-4V)'s performance for image interpretation, diagnosis formulation, and patient management capabilities. We aim to shed light on its potential as an educational tool addressing real-life cases for medical students.

Methods: Ten of the most popular orthopaedic trauma cases from OrthoBullets were selected. GPT-4V interpreted medical imaging and patient information, providing diagnoses, and guiding responses to OrthoBullets questions. Four fellowship-trained orthopaedic trauma surgeons rated GPT-4V responses using a 5-point Likert scale (strongly disagree to strongly agree). Each of GPT-4V's answers was assessed for alignment with current medical knowledge (accuracy), rationale and whether it is logical (rationale), relevancy to the specific case (relevance), and whether surgeons would trust the answers (trustworthiness). Mean scores from surgeon ratings were calculated.

Results: In total, 10 clinical cases, comprising 97 questions, were analyzed (10 imaging, 35 management, and 52 treatment). The surgeons assigned a mean overall rating of 3.46/5.00 to GPT-4V's imaging response (accuracy 3.28, rationale 3.68, relevance 3.75, and trustworthiness 3.15). Management questions received an overall score of 3.76 (accuracy 3.61, rationale 3.84, relevance 4.01, and trustworthiness 3.58), while treatment questions had an average overall score of 4.04 (accuracy 3.99, rationale 4.08, relevance 4.15, and trustworthiness 3.93).

Conclusion: This is the first study evaluating GPT-4V's imaging interpretation, personalized management, and treatment approaches as a medical educational tool. Surgeon ratings indicate overall fair agreement in GPT-4V reasoning behind decision-making. GPT-4V performed less favorably in imaging interpretation compared with its management and treatment approach performance. The performance of GPT-4V falls below our fellowship-trained orthopaedic trauma surgeon's standards as a standalone tool for medical education.

Jan Szatkowski: Stock Holder of Bullet Health.

Disclosure: The **Disclosure of Potential Conflicts of Interest** forms are provided with the online version of the article (http://links.lww.com/JBJSOA/A700).

Copyright © 2024 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 (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

Artificial intelligence (AI) development has rapidly increased with potential applications in clinical decision support, medical documentation, and patient education[1-3]. However, its use also raises concerns about accuracy, privacy, and ethical considerations[2-4]. Despite these limitations, AI tools such as ChatGPT have the potential to revolutionize healthcare delivery by providing personalized, data-driven insights and supporting the work of healthcare professionals[5].

ChatGPT is a large language model (LLM) trained on immense amounts of textual data, allowing it to generate human-like responses to users' questions. Since its release, it has gained attention, with a significant increase in publications exploring its use in the medical field across various specialties[6-10]. Since the study conducted by Kung et al. demonstrated that ChatGPT can complete and pass the United States Medical Licensing Examination (USMLE) without prior program-specific training, widespread conversation about its potential applications in medical education began[11]. The study found that ChatGPT's responses had a high concordance and low self-contradiction, which authors deemed sound clinical reasoning. These findings sparked inquiries about ChatGPT's potential in clinical decision-making and educating future clinicians. ChatGPT provides instant information on medical topics such as diseases, treatments, and procedures. It allows students to engage in simulated dialogues, discuss medical case management, and receive personalized, adaptive tutoring that aligns with precision medical education principles[12,13]. Moreover, ChatGPT can be used as a testing tool, explaining why certain answers to multiple choice questions (MCQ) are correct or incorrect[14]. The American Medical Association (AMA) has supported AI's role in generating USMLE-style questions and simulating patient interactions and acknowledges its integration into medical curricula to enhance clinical skills, including diagnosis of medical conditions[15]. A survey of international students entering residency programs showed that 57% would use ChatGPT for examination preparation and 53.2% believed it improved their learning[14]. As of today, ChatGPT-4 excelled in orthopaedic MCQ performance, outperforming average Post Graduate Year 5 (PGY-5) orthopaedic residents in Orthopaedic In-Training Examination questions[16]. It also provided evidence-based answers to common patient queries about total hip arthroplasty and surpassed ChatGPT-3.5 and Google Bard in formulating orthopaedic decisions based on OrthoBullets cases[17,18]. Despite these promising aspects, limitations persisted in the tool's inability to answer image-based questions[16-19].

The major challenge previously explored is benchmarking AI tools against human clinicians for clinical education[18]. As of September 25, 2023, ChatGPT has added new voice and image capabilities, allowing users to input one or more images in the ChatGPT chatbox[20]. The introduction of ChatGPT-4 vision (GPT-4V), which can understand text and interpret images, makes it potentially useful for interpreting medical images, thereby enhancing its appeal as an educational tool. As the popularity of ChatGPT continues to surge, medical students are increasingly leveraging its capabilities to address questions and access information on various topics[21]. Evaluating ChatGPT's clinical reasoning behind decision-making for orthopaedic cases with multifaceted answers has yet to be explored. Therefore, we utilized real-life orthopaedic trauma cases sourced from OrthoBullets, the world's largest orthopaedic educational platform[22]. The platform presents diverse orthopaedic cases, incorporating radiographs, patient history, and physical examination findings, along with OrthoBullets uploaded poll questions addressing management and treatment strategies. Trauma cases were deliberately selected for their unspecific nature in terms of fracture locations, encompassing scenarios involving various parts of the body.

We aim to assess GPT-4V's performance in interpreting medical images and its ability to rationalize its decision-making capabilities for different MCQ related to these orthopaedic trauma cases. With the rise in its use, our study analyzes whether GPT-4V can be adequately used as a reliable educational tool for medical professionals.

Methods

GPT-4V Prompting

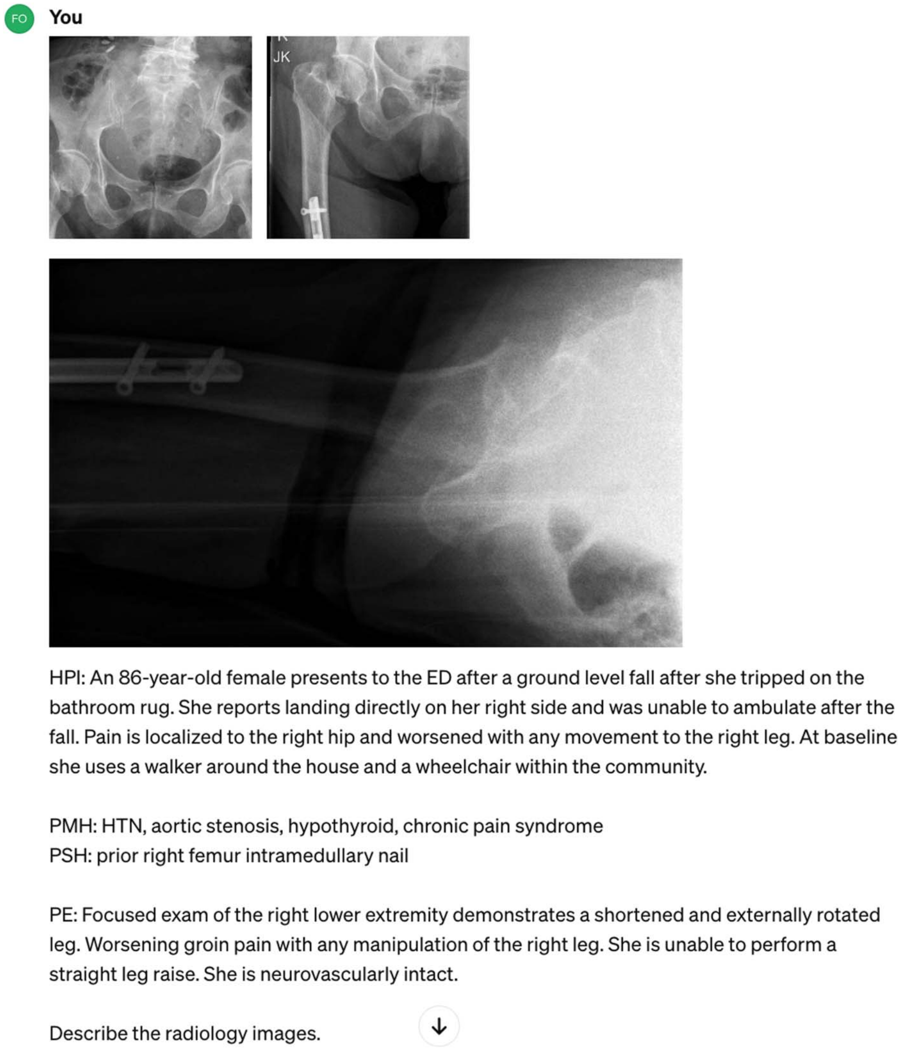
Ten orthopaedic trauma cases were selected from the OrthoBullets website in September 2023. The “Cases” section of OrthoBullets provides detailed clinical scenarios for orthopaedic education, featuring patient histories, radiographic images, diagnostic information, and treatment options. It includes management and treatment questions for members to enhance learning of orthopaedic principles. We were granted written approval by the founder of OrthoBullets to use their data, in adherence with their terms and conditions. All 10 cases were selected from the “Trauma” category and “Popular” subcategory. Two cases of each fracture type, including femoral neck, femoral shaft, distal femur, proximal humeral, and distal radial fractures, were included in the study. As the femur cases constituted the majority (3 of 5), we intentionally included cases from the radius and humerus to analyze different fracture types. The selected cases were chosen based on the clarity of images, sufficient patient history, and popularity determined by at least 8,000 poll responses from OrthoBullets members. Although this is not an exhaustive list, we included cases with and without a clear consensus on management and treatment to fully evaluate ChatGPT-4V rationalizations behind its decision-making. For our investigation, we specifically utilized the ChatGPT-4 vision, November 21, 2023, version[20]. We inputted medical images and patient details that were uploaded by surgeons on the OrthoBullets website into GPT-4V's chat box from September to October 2023. Our first question was to “Describe the radiology images” for each case (Fig. 1). We then input the uploaded OrthoBullet's management and treatment MCQs on each case into the GPT-4V chat box (Figs. 2 and 3). We defined a management question as one that involves determining the appropriate strategy for a patient case. A treatment question focuses on identifying suitable interventions or surgical options for fractures and assessing the user's understanding of therapeutic choices. Each case had

1 question for imaging and a variable number of questions pertaining to management or treatment categories based on the specific case on the OrthoBullets website. For example, our femoral neck fracture was taken from case #C101423 which consists of 3 management questions (questions 1-3) and 9 treatment questions (questions 4-11). Across 10 cases, 97 questions were input into ChatGPT, composed of 10 imaging, 35 management, and 52 treatment questions. The heterogeneity of the number of questions is due to the nature of questions one can ask for each subcategory, seeing as only 1 imaging question was asked per case by OrthoBullets.

Qualitative Analysis of GPT-4V

Four fellowship-trained orthopaedic trauma surgeons assessed the GPT-4V explanations in medical image interpretation, patient management, and treatment MCQs. Each of GPT-4V's answers was assessed for alignment with current medical knowledge and guidelines (accuracy), whether they were logical and understandable (rationale), their relevancy to the specific case or scenario (relevance), and whether the surgeons would trust the medical information provided in the healthcare setting (trustworthiness) (Fig. 4). Surgeons rated answers on a 5-point Likert scale from "strongly disagree" to "strongly agree," providing qualitative assessments of GPT-4V explanations.

Means, standard deviations, medians, and interquartile ranges were calculated to summarize surgeon ratings across all 97 questions posed to GPT-4V and categorized separately by imaging and MCQ. We used the intraclass correlation coefficient (ICC) employing a 2-way mixed-effects model with single measures as described by Shrout and Fleiss to assess the level of agreement between the surgeon ratings of GPT-4V's performance[23]. We interpreted the ICC values as described by Cichetti (excellent 0.75-1.00; good 0.60-0.74; fair 0.40-0.59; poor <0.40)[24]. This study was conducted from September 2023 to January 2024.



HPI: An 86-year-old female presents to the ED after a ground level fall after she tripped on the bathroom rug. She reports landing directly on her right side and was unable to ambulate after the fall. Pain is localized to the right hip and worsened with any movement to the right leg. At baseline she uses a walker around the house and a wheelchair within the community.

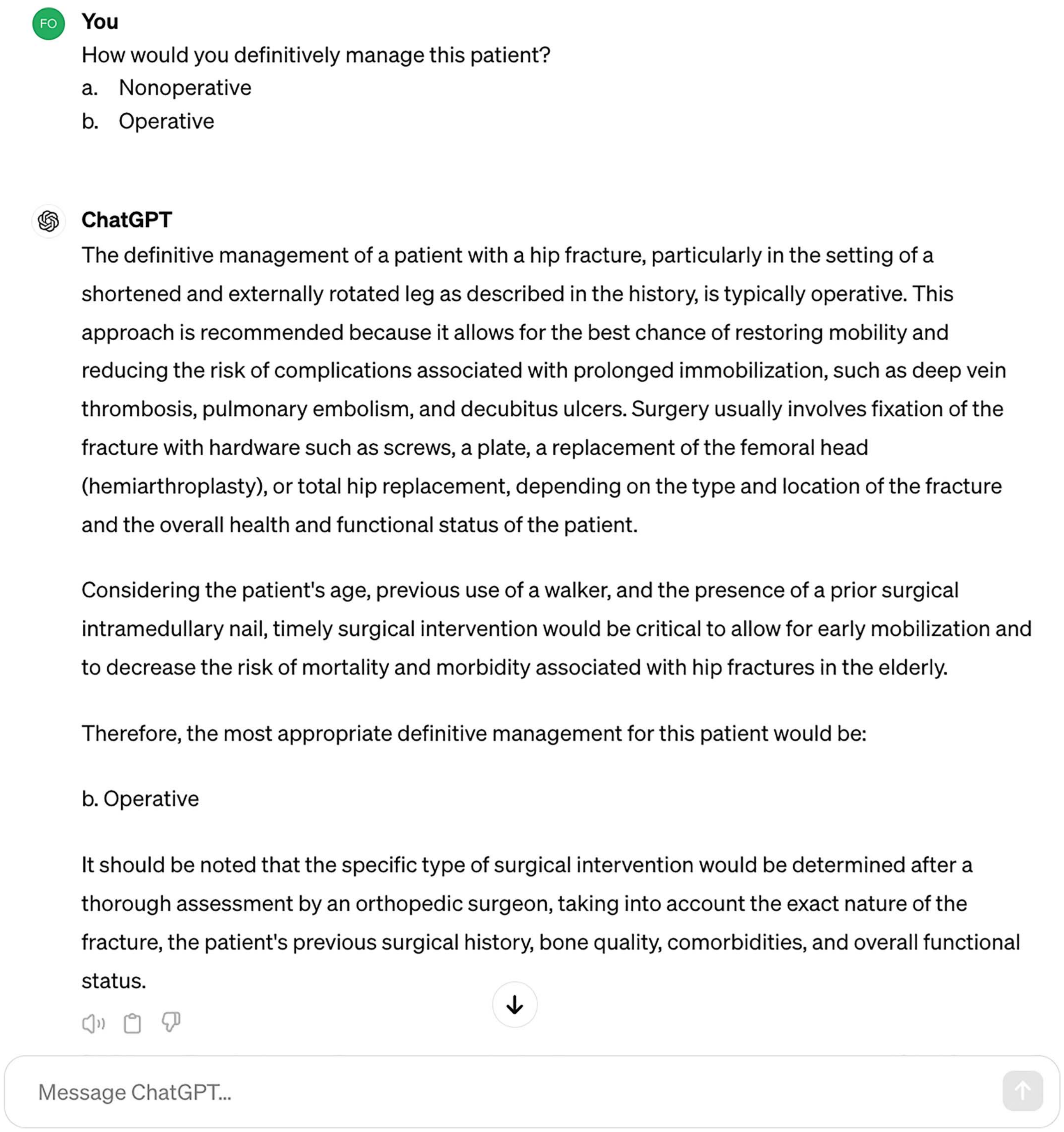
PMH: HTN, aortic stenosis, hypothyroid, chronic pain syndrome
PSH: prior right femur intramedullary nail

PE: Focused exam of the right lower extremity demonstrates a shortened and externally rotated leg. Worsening groin pain with any manipulation of the right leg. She is unable to perform a straight leg raise. She is neurovascularly intact.

Describe the radiology images.

Fig. 1

Imaging interpretation question input into ChatGPT-4 vision screenshot (femoral neck case 2).



You

How would you definitively manage this patient?

a. Nonoperative

b. Operative

ChatGPT

The definitive management of a patient with a hip fracture, particularly in the setting of a shortened and externally rotated leg as described in the history, is typically operative. This approach is recommended because it allows for the best chance of restoring mobility and reducing the risk of complications associated with prolonged immobilization, such as deep vein thrombosis, pulmonary embolism, and decubitus ulcers. Surgery usually involves fixation of the fracture with hardware such as screws, a plate, a replacement of the femoral head (hemiarthroplasty), or total hip replacement, depending on the type and location of the fracture and the overall health and functional status of the patient.

Considering the patient's age, previous use of a walker, and the presence of a prior surgical intramedullary nail, timely surgical intervention would be critical to allow for early mobilization and to decrease the risk of mortality and morbidity associated with hip fractures in the elderly.

Therefore, the most appropriate definitive management for this patient would be:

b. Operative

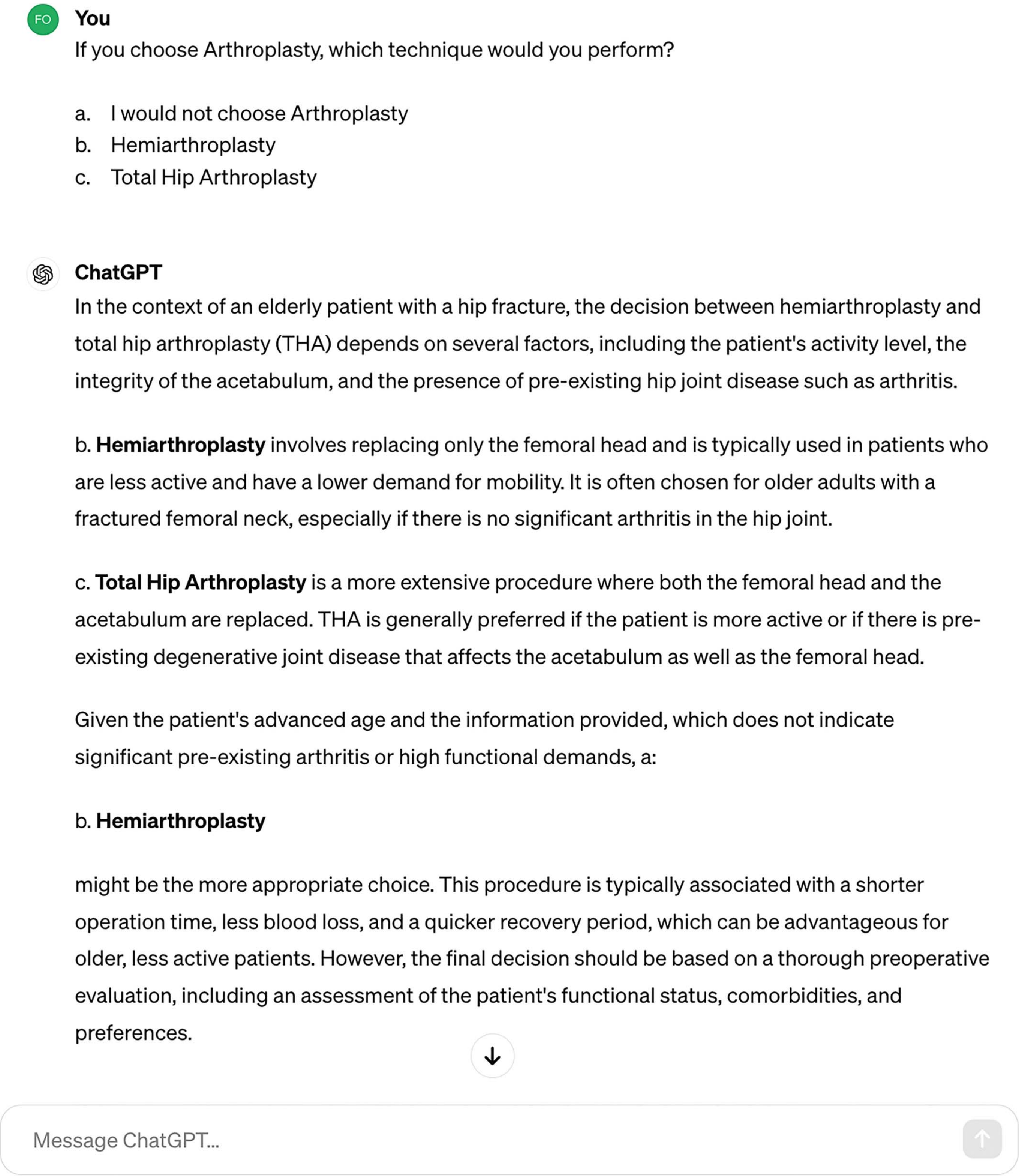
It should be noted that the specific type of surgical intervention would be determined after a thorough assessment by an orthopedic surgeon, taking into account the exact nature of the fracture, the patient's previous surgical history, bone quality, comorbidities, and overall functional status.

Fig. 2

Example of management question input into ChatGPT-4 vision screenshot (femoral neck case 2).

Results

When aggregated across all 97 questions, the mean surgeon rating was 3.78 for accuracy, 3.95 for rationale, 4.06 for relevance, and 3.72 for trustworthiness (Table I). Examples of both a highly rated and poorly rated GPT-4V interpretation from the surgeons are provided in Supplemental Figure 1 and Supplemental Figure 2. When stratified by imaging, MCQ management, and MCQ treatment, the mean rating for each of the 4 characteristics was qualitatively highest for MCQ treatment, followed by MCQ management and then imaging questions. The largest differences were a mean difference of 0.78 points for trustworthiness and 0.71 points for accuracy for MCQ treatment



You

If you choose Arthroplasty, which technique would you perform?

a. I would not choose Arthroplasty
b. Hemiarthroplasty
c. Total Hip Arthroplasty

ChatGPT

In the context of an elderly patient with a hip fracture, the decision between hemiarthroplasty and total hip arthroplasty (THA) depends on several factors, including the patient's activity level, the integrity of the acetabulum, and the presence of pre-existing hip joint disease such as arthritis.

b. **Hemiarthroplasty** involves replacing only the femoral head and is typically used in patients who are less active and have a lower demand for mobility. It is often chosen for older adults with a fractured femoral neck, especially if there is no significant arthritis in the hip joint.

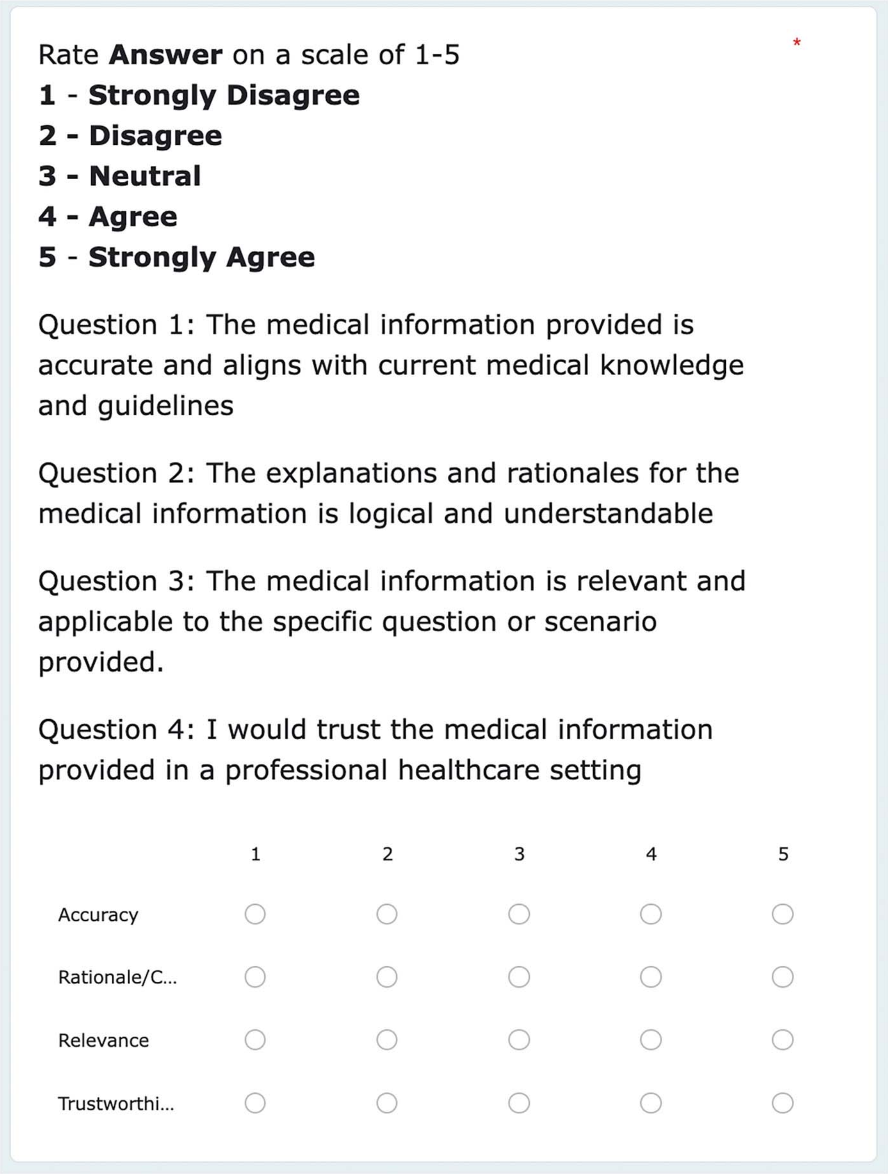
c. **Total Hip Arthroplasty** is a more extensive procedure where both the femoral head and the acetabulum are replaced. THA is generally preferred if the patient is more active or if there is pre-existing degenerative joint disease that affects the acetabulum as well as the femoral head.

Given the patient's advanced age and the information provided, which does not indicate significant pre-existing arthritis or high functional demands, a:

b. **Hemiarthroplasty**

might be the more appropriate choice. This procedure is typically associated with a shorter operation time, less blood loss, and a quicker recovery period, which can be advantageous for older, less active patients. However, the final decision should be based on a thorough preoperative evaluation, including an assessment of the patient's functional status, comorbidities, and preferences.

Fig. 3
Example of treatment question input into ChatGPT-4 vision screenshot (femoral neck case 2).



Rate **Answer** on a scale of 1-5 *

1 - Strongly Disagree
2 - Disagree
3 - Neutral
4 - Agree
5 - Strongly Agree

Question 1: The medical information provided is accurate and aligns with current medical knowledge and guidelines

Question 2: The explanations and rationales for the medical information is logical and understandable

Question 3: The medical information is relevant and applicable to the specific question or scenario provided.

Question 4: I would trust the medical information provided in a professional healthcare setting

	1	2	3	4	5
Accuracy	○	○	○	○	○
Rationale/C...	○	○	○	○	○
Relevance	○	○	○	○	○
Trustworthi...	○	○	○	○	○

Fig. 4
Likert Scale provided to surgeons to rate ChatGPT-4 vision performance of OrthoBullets cases.

vs. imaging. The median surgeon rating for accuracy, rationale, relevance, and trustworthiness was 4 across all questions, imaging, and MCQs except for trustworthiness for imaging questions (median 3.5, interquartile range [IQR] IQR 2-4). Surgeon agreement was fair to poor when ratings were examined across all questions, and it was stronger overall for imaging vs. MCQ (Table I). Specifically, interrater agreement of ratings was strongest for accuracy and trustworthiness for imaging questions.

Discussion

In the orthopaedic domain, where nuanced expertise is crucial, ChatGPT's potential as a clinical decision-support tool is compelling[25]. Our study focuses on the latest version, GPT-4V, which integrates text and image interpretation capabilities[20].

Our study is the first to evaluate the quality of GPT-4V's image interpretations and clinical decision-making, as assessed by fellowship-trained orthopaedic trauma surgeons. Unlike other studies, we focused on questions without clear-cut answers to reflect real-life scenarios, where clinical judgment is needed. While AI can handle MCQs, it struggles with questions requiring reasoning. This highlights why we would not recommend using AI as a learning tool for medical students in clinical settings. We included only fellowship-trained orthopaedic trauma surgeons in the grading process because their specialized knowledge and experience enable them to accurately identify errors in GPT-4V's answers, ensuring a high-quality peer review and reliable assessment of inaccuracies.

TABLE I Surgeon Ratings and Intraclass Correlation Coefficients of ChatGPT-4 Vision Ability to Answer OrthoBullets Questions*

	Mean (Standard Deviation)	Median (IQR)	ICC (Agreement)
All questions			
Accuracy	3.78 ± 1.06	4 (3-5)	0.40 (fair)
Rationale	3.95 ± 0.94	4 (4-5)	0.28 (poor)
Relevance	4.06 ± 0.82	4 (4-5)	0.20 (poor)
Trustworthiness	3.72 ± 1.03	4 (3-4)	0.40 (fair)
Imaging questions			
Accuracy	3.28 ± 1.22	4 (2-4)	0.75 (excellent)
Rationale	3.68 ± 1.07	4 (3-4)	0.51 (fair)
Relevance	3.75 ± 1.04	4 (3.5-4)	0.29 (poor)
Trustworthiness	3.15 ± 1.23	3.5 (2-4)	0.76 (excellent)
Multiple choice questions			
Accuracy			
Management	3.61 ± 1.12	4 (3-4)	0.32 (poor)
Treatment	3.99 ± 0.93	4 (4-5)	
Rationale			
Management	3.84 ± 1.05	4 (3-5)	0.23 (poor)
Treatment	4.08 ± 0.81	4 (4-5)	
Relevance			
Management	4.01 ± 0.86	4 (4-5)	0.16 (poor)
Treatment	4.15 ± 0.72	4 (4-5)	
Trustworthiness			
Management	3.58 ± 1.09	4 (3-4)	0.30 (poor)
Treatment	3.93 ± 0.89	4 (3.5-5)	

*ICC = intraclass correlation coefficient: excellent: 0.75 to 1.00; good: 0.60 to 0.74; fair: 0.40 to 0.59; poor: <0.40, and IQR = interquartile range.

Imaging Analysis

The capability of GPT-4V to interpret medical images represents a significant milestone in orthopaedic education, a domain where previous studies have not explored image interpretation capabilities[16-19]. Our surgeons rated the accuracy of GPT-4V image interpretation as 3.28/5.00 with strong agreement between raters. This suggests that our surgeons rated GPT-4V's performance as neutral in imaging analysis. The average ratings for the rationale of the imaging interpretation were between "neutral" and "agree," with a moderate agreement between our surgeons. Although our surgeons rated the relevance of GPT-4V's responses to each case the highest out of our 4 metrics, we interpret this finding with caution with poor agreement between our raters. However, surgeons fairly agree on a cautious approach to utilizing GPT-4V in professional settings, with trust scores being the lowest of other metric or question types.

Examination of ratings for relevance and rationale for each image interpretation identified areas for improvement. Our raters scored the relevance of GPT-4V's answers the highest for image interpretation, showcasing its ability to tailor information to the presented scenario. Improvements can be made such as enhancing clarity and avoiding generalized language when explaining medical images.

We recognize AI's transformative capabilities, especially its capacity for learning through repetition, make it a valuable asset in medical education[2]. While improvements in radiographic reading accuracy are needed, the positive 'neutral to agree' responses suggest potential for future integration into educational frameworks. However, GPT-4V's image interpretation is not currently ready for integration into medical education. This consensus aligns with broader research advocating for rigorous validation and testing of AI algorithms on representative data sets before considering their implementation[25].

Management and Treatment

MCQ were used to evaluate how well GPT-4V could differentiate between answer choices, as OrthoBullets presents their polls to surgeons around the world in this way. It serves to simulate possible questions medical students may ask GPT-4V, therefore analyzing its potential to educate them on each answer choice. In our analysis for accuracy, rationale, relevance, and trustworthiness regarding MCQs, we found consistent mean ratings above 3.5 on a 5-point Likert scale. Notably, the relevance of management and treatment questions received the highest ratings on average, but with poor agreement among surgeons regarding ChatGPT's performance in this area. Examples of poorly rated and well-rated MCQ from GPT-4V are provided in Supplemental Figures 1 and 2.

All answers for treatment were rated even higher than for management, which leads us to believe ChatGPT is more adapted to provide next-step answers than answers for questions regarding the beginning of a case (Table I). Previous research examined ChatGPT-4's performance in orthopaedic questions, finding it agreed with the OrthoBullets majority in 66 of 97 cases (68.0%). However, this study had only evaluated the frequency that ChatGPT picked the answer choice that the majority of members in OrthoBullets picked for each question on their polls, not the quality of reasoning why it picked each answer[18]. ChatGPT-3.5 exhibits poorer performance on questions requiring higher-order thinking compared with lower-order recall and understanding[26]. Similarly, in assessments using sequential "select all that apply" questions from clinical cases, ChatGPT shows lower accuracy on initial diagnostic steps and management compared with final diagnosis, indicating better performance with more information[27]. Moreover, GPT-4V struggles to properly navigate clinical scenarios with more ambiguous courses of action (e.g., ruling out extra or unnecessary diagnostic testing or clinical intervention)[27].

Educational Implications

ChatGPT's growing accessibility highlights its emerging role as a powerful tool in medical education. Many medical students already view AI as beneficial to physicians, helping with quick access to information and reducing errors[28]. As ChatGPT becomes more integrated into daily tools, such as through Apple's latest updates, where users can access ChatGPT directly from their phones, its accessibility in both educational and clinical settings is growing[29,30]. As stated, it has been used to assist with test preparation by explaining correct and incorrect answers to MCQ[14]. The AMA has endorsed AI's use in generating questions and simulating patient interactions, recognizing its potential to enhance clinical skills, including diagnostic reasoning. Students turn to ChatGPT for answering practice questions, interpreting radiographs, or clarifying complex cases encountered in lectures or rotations. Given this increasing reliance, it is crucial to evaluate ChatGPT's accuracy and reliability, particularly in specialized fields such as orthopaedics, to ensure students are well-informed of its limitations. As research comparing ChatGPT with traditional search engines such as Google in medical contexts grows, the need for a thorough understanding of its strengths and weaknesses becomes all the more important.

The increased need for personalized learning supports ChatGPT's integration into education and highlights its strong accessibility [31]. Our fellowship-trained orthopaedic trauma surgeons, collectively express high agreement that imaging interpretation is "neutral," and generally agree with GPT-4V on the MCQ explanations. However, there is a lack of consensus among them regarding the specific ratings. This suggests that at its current state, GPT-4V may not perform at the level needed to be used as a medical educational tool. Further development of the LLM is needed to consistently reach "agree" and "strongly agree" evaluations with high ICC values between raters in all cases, to be considered a medical education tool for medical trainees. Based on its previous performance on simple MCQs, ChatGPT has been proposed as a "decision support" tool in clinical management, assisting with treatment suggestions based on patient symptoms and medical history[12]. However, the excitement surrounding its integration into clinical space should be tempered, acknowledging that explanations behind each answer fall short of acceptable clinical standards. While the potential is promising, it should not be used as a current tool for clinical reasoning in medical education.

This study is not without limitations. Specifically, only 4 board-certified orthopaedic trauma surgeons were used as graders, given the large number of answers that needed to be rated. These raters may not be representative of all surgeons. There was also overall poor interrater agreement for surgeon ratings, which may reflect differences in practice and training experiences as well as preference-sensitive decisions. Differences in ratings between individuals, especially for trustworthiness, may also reflect differences in biases in perceptions of AI. Future research should consider evaluating ChatGPT's stability of performance by requesting a response to each prompt several times. In addition, we are only able to examine the current state of ChatGPT and are unable to ascertain how close ChatGPT is to being ready for use as an educational tool for the medical field.

This study lays a foundation for integrating AI tools to enhance the orthopaedic educational landscape such as OrthoBullets and other online platforms. However, we recommend exercising caution with these tools, as we have highlighted areas of further development and refinement needed to meet the needs for accurate orthopaedic education. Future studies are needed to assess specific changes that can be made to ChatGPT's answers and rationale to satisfy the requirements as an educational tool. In addition, a comprehensive study evaluating the current usage of ChatGPT by medical students would also help further the literature's understanding of this topic.

Conclusion

GPT-4V is rapidly reshaping medical education, showing potential as a support tool in orthopaedics trauma cases. Our analysis indicates a neutral consensus in GPT-4V's imaging interpretation performance; it demonstrates better capabilities in management and treatment queries, offering patient-specific advice effectively. The performance of GPT-4V falls below our fellowship-trained orthopaedic surgeons' standards as a tool for medical education. We stress ChatGPT's necessity for continued refinement to become widely accepted as a reliable medical education tool. Integrating AI-driven tools into educational programs presents exciting opportunities but must be meticulously evaluated before its integration.

Appendix

eA Supporting material provided by the author is posted with the online version of this article as a data supplement at jbjs.org (http://links.lww.com/JBJSOA/A701). This content has not been copyedited or verified. ■

Note: The authors thank Brocha Stern, PhD, for assistance with the analyses.

Arthur Drouaud, BS[1]
Carolina Stocchi, BS[2]
Justin Tang, BS[2]
Grant Gonsalves, BA[2]
Zoe Cheung, MD[3]
Jan Szatkowski, MD[4]
David Forsh, MD[2]

[1]George Washington University School of Medicine, Washington, District of Columbia

[2]Department of Orthopaedic Surgery, Mount Sinai, New York, New York

[3]Department of Orthopaedic Surgery, Staten Island University Hospital, Staten Island, New York

[4]Department of Orthopaedic Surgery, Indiana University Health Methodist Hospital, Indianapolis, Indiana

E-mail address for A. Drouaud: arthurdrouaud@gwu.edu

References

1. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res. 2023;25:e48568.

2. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595.

3. Garg RK, Urs VL, Agarwal AA, Chaudhary SK, Paliwal V, Kar SK. Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: a systematic review. Health Promot Perspect. 2023;13(3):183-91.

4. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, Floridi L. The ethics of AI in health care: a mapping review. Soc Sci Med. 2020;260:113172.

5. Tustumi F, Andreollo NA, Aguilar-Nascimento JE. Future of the language models in healthcare: the role of ChatGPT. Arq Bras Cir Dig. 2023;36:e1727.

6. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res. 2023;104(5):269-73.

7. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. J Am Coll Radiol. 2023;20:990-7.

8. Baker MN, Burruss CP, Wilson CL. ChatGPT: a supplemental tool for efficiency and improved communication in rural dermatology. Cureus. 2023;15(8):e43812.

9. Nedbal C, Naik N, Castellani D, Gauhar V, Geraghty R, Somani BK. ChatGPT in urology practice: revolutionizing efficiency and patient care with generative artificial intelligence. Curr Opin Urol. 2024;34(2):98-104.

10. Sharma SC, Ramchandani JP, Thakker A, Lahiri A. ChatGPT in plastic and reconstructive surgery. Indian J Plast Surg. 2023;56(4):320-5.

11. 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.

12. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT–reshaping medical education and clinical management. Pak J Med Sci. 2023;39(2):605-7.

13. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing ChatGPT for medical education: exploring its impact on doctors and medical students. JMIR Med Educ. 2024;10:e52483.

14. Alkhaaldi SMI, Kassab CH, Dimassi Z, Oyoun Alsoud L, Al Fahim M, Al Hageh C, Ibrahim H. Medical student experiences and perceptions of ChatGPT and artificial intelligence: cross-sectional study. JMIR Med Educ. 2023;9:e51302.

15. ChatGPT in Medical Education. Generative AI and the Future of Artificial Intelligence in Health Care. American Medical Association; 2024. Available at: https://www.ama-assn.org/practice-management/digital/chatgpt-medical-education-generative-ai-and-future-artificial. Accessed September 13, 2024.

16. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JBI. Evaluating ChatGPT performance on the orthopaedic in-training examination. JBJS Open Access. 2023;8(3):e23.00056.

17. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. J Bone Joint Surg Am. 2023;105(19):1519-26.

18. Agharia S, Szatkowski J, Fraval A, Stevens J, Zhou Y. The ability of artificial intelligence tools to formulate orthopaedic clinical decisions in comparison to human clinicians: an analysis of ChatGPT 3.5, ChatGPT 4, and Bard. J Orthop. 2024;50:1-7.

19. Zhou Y, Moon C, Szatkowski J, Moore D, Stevens J. Evaluating ChatGPT responses in the context of a 53-year-old male with a femoral neck fracture: a qualitative analysis. Eur J Orthop Surg Traumatol. 2024;34(2):927-55.

20. ChatGPT—release notes|OpenAI help center. Available at: https://help.openai.com/en/articles/6825453-chatgpt-release-notes. Accessed April 22, 2024.

21. Duarte F. Number of ChatGPT users (Jan 2024). Exploding Topics; 2024. Available at: https://explodingtopics.com/blog/chatgpt-users. Accessed January 24, 2024.

22. Cohn RM, Klein BJ, Bitterman AD, Nellans KW. Update on educational resources and evaluation tools for orthopaedic surgery residents. J Am Acad Orthop Surg. 2023;31(13):660-8.

23. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420-8.

24. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 1994;6(4):284-90.

25. Giorgino R, Alessandri-Bonetti M, Luca A, Migliorini F, Rossi N, Peretti GM, Mangiavini L. ChatGPT in orthopedics: a narrative review exploring the potential of artificial intelligence in orthopedic practice. Front Surg. 2023;10:1284015.

26. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology. 2023;307(5):e230582.

27. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, Landman A, Dreyer K, Succi MD. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res. 2023;25:e48659.

28. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ. 2022;22(1):772.

29. OpenAI and Apple announce partnership. Available at: https://openai.com/index/openai-and-apple-announce-partnership/. Accessed September 22, 2024.

30. Introducing Apple Intelligence for iPhone, iPad, and Mac. Apple Newsroom. Available at: https://www.apple.com/newsroom/2024/06/introducing-apple-intelligence-for-iphone-ipad-and-mac/. Accessed October 2, 2024.

31. Zhu C, Sun M, Luo J, Li T, Wang M. How to harness the potential of ChatGPT in education? Knowl Manag E Learn. 2023;15(2):133-52.