Journal of Hand Surgery Global Online 6 (2024) 164–168

Contents lists available at ScienceDirect

Journal of Hand Surgery Global Online

journal homepage: www.JHSGO.org

Original Research

Comparison of Artificial Intelligence to Resident Performance on Upper-Extremity Orthopaedic In-Training Examination Questions

Yagiz Ozdag, MD, * Daniel S. Hayes, BS, * Gabriel S. Makar, MD, * Shahid Manzar, MEng, * Brian K. Foster, MD, * Mason J. Shultz, BS, * Joel C. Klena, MD, * Louis C. Grandizio, DO *

* Department of Orthopaedic Surgery, Geisinger Musculoskeletal Institute, Geisinger Commonwealth School of Medicine, Danville, PA

ARTICLE INFO

Article history:
Received for publication July 10, 2023
Accepted in revised form October 28, 2023
Available online December 11, 2023

Key words:
Artificial intelligence
ChatGPT
Orthopaedic In-Training Examination
Resident education
Upper extremity

Purpose: Currently, there is a paucity of prior investigations and studies examining applications for artificial intelligence (AI) in upper-extremity (UE) surgical education. The purpose of this investigation was to assess the performance of a novel AI tool (ChatGPT) on UE questions on the Orthopaedic In-Training Examination (OITE). We aimed to compare the performance of ChatGPT to the examination performance of hand surgery residents.

Methods: We selected questions from the 2020–2022 OITEs that focused on both the hand and UE as well as the shoulder and elbow content domains. These questions were divided into two categories: those with text-only prompts (text-only questions) and those that included supplementary images or videos (media questions). Two authors (B.K.F. and G.S.M.) converted the accompanying media into text-based descriptions. Included questions were inputted into ChatGPT (version 3.5) to generate responses. Each OITE question was entered into ChatGPT three times: (1) open-ended response, which requested a free-text response; (2) multiple-choice responses without asking for justification; and (3) multiple-choice response with justification. We referred to the OITE scoring guide for each year in order to compare the percentage of correct AI responses to correct resident responses.

Results: A total of 102 UE OITE questions were included; 59 were text-only questions, and 43 were media-based. ChatGPT correctly answered 46 (45%) of 102 questions using the Multiple Choice No Justification prompt requirement (42% for text-based and 44% for media questions). Compared to ChatGPT, postgraduate year 1 orthopaedic residents achieved an average score of 51% correct. Post-graduate year 5 residents answered 76% of the same questions correctly.

Conclusions: ChatGPT answered fewer UE OITE questions correctly compared to hand surgery residents of all training levels.

Clinical relevance: Further development of novel AI tools may be necessary if this technology is going to have a role in UE education.

The Orthopaedic In-Training Examination (OITE) was initially developed in 1960 as the first in-training examination provided to resident trainees.[1,2] The purpose of the OITE was to offer programs an objective measure to assess the effectiveness of their educational systems and evaluate the clinical knowledge of residents relative to their peers.[3] Over time, the OITE has evolved into a yearly administered examination consisting of 275 multiple-choice (MC) questions for all hand surgery residents in training. The examination covers 11 domains, encompassing various orthopedic subspecialties, practice management, and fundamental scientific concepts.[4–8] Each year, approximately 16% of the questions pertain to upper-extremity (UE) topics, such as the hand, wrist, shoulder, and elbow.[6,9–11] Residents receive a gross score for a total of 275 questions and a percentile rank compared to their peers. Prior investigations have demonstrated a correlation between increasing OITE scores and passing the American Board of Orthopaedic Surgery Part I examination, with the correlation strengthening across postgraduate years (PGYs) 2–5.[12,13]

Declaration of interests: No benefits in any form have been received or will be received related directly to this article.

Corresponding author: Louis C. Grandizio, DO, Department of Orthopaedic Surgery, Geisinger Musculoskeletal Institute, Geisinger Commonwealth School of Medicine, 16 Woodbine Lane, Danville, PA 17821.

E-mail address: chris.grandizio@gmail.com (L.C. Grandizio).

https://doi.org/10.1016/j.jhsg.2023.10.013

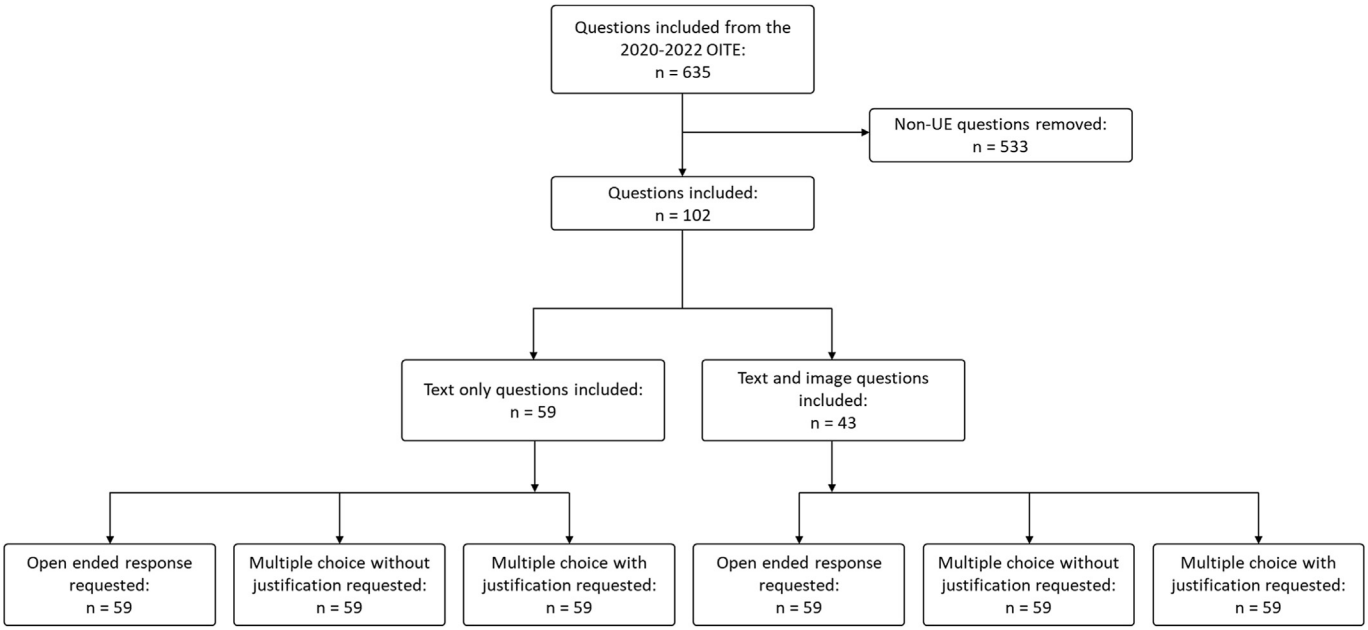

Figure 1. Flowchart demonstrating the selection for the UE OITE questions included in the study as well as breakdown of text-only and media questions.

Over the past decade, advancements in technology have expanded, particularly with the development of deep learning models and artificial intelligence (AI). OpenAI's ChatGPT (Chat Generative Pre-Trained Transformer) is an AI model that operates on an internet-based, continuously reinforced large language model framework. It possesses the capability to generate new sequences or responses based on the given input.[14] ChatGPT has been used to assist in scientific writing, literature reviews, and formulating research questions.[15] Although ethical concerns regarding this emerging technology continue to be discussed, OpenAI's technology may play an increasing role in clinical practice and education. It can streamline clinical workflows, improve documentation, and generate diverse clinical vignettes within medical education.[16] Notably, recent demonstrations have showcased ChatGPT's ability to write articles, serve as a virtual teaching assistant, and provide reasonably accurate responses to standardized medical questions.[17] Furthermore, ChatGPT was able to pass the United States Medical Licensing Examination step 1.[18] Additionally, it has achieved scores exceeding 80% on various assessments, including the Law School Admission Test, multiple Advanced Placement examinations, and the Uniform Bar Examination.[14,18]

Currently, there is a paucity of investigations examining ChatGPT's proficiency in answering clinical questions within the field of UE surgery. Evaluating these technologies' ability to address OITE questions can provide insight into its capacity to interpret orthopedic knowledge and may help define its potential role in postgraduate medical education. The purpose of this investigation was to assess the performance of ChatGPT on the OITE UE questions and compare it to the performance of resident trainees. We hypothesized that the performance of ChatGPT on the OITE would be inferior to that of surgical trainees.

Materials and Methods

Approval from the institutional review board was not sought for this study because it did not involve any direct or indirect participation of human subjects. We selected questions from the OITEs from the years 2020, 2021, and 2022 that specifically focused on both the hand and UE as well as the shoulder and elbow content domains (Fig. 1). These questions were divided into two categories: those with only text prompts (text-only questions) and those that included supplementary images or videos (media questions).

Currently, the ChatGPT model lacks the ability to independently analyze and interpret media files. Therefore, to incorporate the information present in the images or videos, two authors (B.K.F. and G.S.M.) carefully examined, interpreted, and converted the accompanying media into text-based descriptions (Fig. 2). Initially, if the image had a description or if the question itself provided a description of the image, we used that information. If such information was not available, we referred to the question's explanation, where the question writer at times provided a clear definition of the picture. Lastly, if neither of these sources provided the necessary details, the images were described by one of the two authors, ensuring that essential information was conveyed within the explanation (Fig. 2).

After converting supplemental media items into text descriptions, the OITE questions were inputted into ChatGPT (version 3.5) to generate responses. Each OITE question was entered into ChatGPT three times, each with a different answer requirement. For the first answer requirement, an open-ended response (OE) was requested, where only the question or prompt was provided without any MC options. The second answer requirement involved providing MC responses alongside the question or prompt, allowing the AI tool to select the most appropriate response. This answer requirement, which reflects the way in which residents are required to answer OITE questions, was defined as Multiple Choice No Justification (MCNJ). The third answer requirement also included MC responses, but this time, the AI tool was asked to select the most suitable response and provide a justification (Multiple Choice with Justification [MCWJ]).

Descriptive statistics were employed for this study. We conducted separate comparisons between the performance of ChatGPT on text-only questions and media questions from the included OITE questions. The responses generated by ChatGPT were categorized as either correct or incorrect for each of the three answer requirements (OE, MCNJ, and MCWJ). In order to evaluate ChatGPT's performance relative to hand surgery residents, we referred to the scoring guide for each year, which provides percentiles indicating

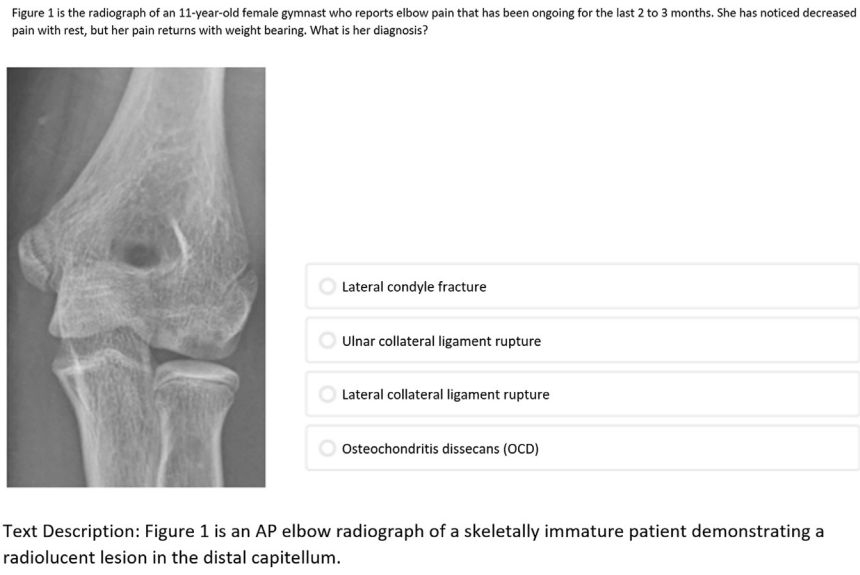

Figure 2. Example of a media-based OITE question that was converted to text format. AP, anteroposterior; OCD; osteochondritis dissecans.

the overall performance of residents based on their PGY. Because residents taking the OITE are required to answer using the MCNJ answer requirement, this was considered the primary outcome measure for comparing ChatGPT's answers with the responses provided by residents who answered the same examination questions. The percentage of correct responses by residents was compared to the percentage of correct responses by ChatGPT based on their respective PGY levels to assess for comparable competency in a very general sense.

Results

Figure 1 presents a flowchart illustrating the selection of OITE questions and the assessment method used. A total of 635 OITE questions from the 2020–2022 OITEs were identified, out of which 533 were excluded because they were outside of the hand/wrist and shoulder/elbow content domains. A total of 102 UE questions were included for this study, with 59 being text-only questions and 43 being media-based questions (Fig. 1). Table 1 provides information on the total number of questions within the UE domains as well as the distribution of question types. The questions were inputted three times, each with different prompt criteria, and Table 2 displays the percentage of correct responses categorized by text-only, image-only, and combined questions.

In total, ChatGPT 3.5 correctly answered 46 (45%) of 102 questions using the MCNJ prompt requirement. Within the MCNJ answer requirement, ChatGPT answered 25 (42%) of 59 text-only questions correctly, compared to 26 (44%) of 43 media questions. For the MCWJ prompt, ChatGPT answered 46 (45%) of 102 questions correctly. For the OE responses, ChatGPT provided correct answers for 38 (37%) of 102 questions. Additional information on correct responses can be found in Table 2. To assess the reliability of ChatGPT in generating consistent responses to the same OITE question, the responses generated under MCNJ and MCWJ were compared. Overall, ChatGPT produced identical responses for 79 (77%) of 102 questions.

Table 3 includes the 2020–2022 UE OITE scores and percentiles for residents in our institution and nationwide, stratified by PGY in training. With an overall percentage of correct responses of 45% using the MCNJ prompt, the ChatGPT score was 6% lower than that of PGY-1 interns nationwide, who achieved an average score of 51% correct responses. As PGY in training increased, the performance gap between ChatGPT and the residents also increased, with PGY-5 residents averaging 76% correct responses.

Discussion

ChatGPT answered 45% of UE OITE questions correctly when given a MC prompt without requiring justification (MCNJ). It is important to note that a score of 69% is generally considered a good indicator of competency in this context.[19] In comparison, first-year hand surgery residents (PGY-1) typically achieve an average score of approximately 51% to 55% on the OITE when considering all content domains, which is considerably higher than the performance of ChatGPT in this scenario.[19] For UE OITE questions, PGY-1 residents nationwide correctly answered 44% to 53% of questions, depending on the year. As expected, both nationwide and institutional performances (range of questions answered corrected for UE OITE questions) also improved as the resident year in training increased. These findings highlight that further development is necessary before ChatGPT can be used in educational scenarios for hand surgery trainees because it presently performs below the level of a PGY-1 trainee. The OITE questions are presented in a MC format with 4–5 options in most cases. If an individual were to guess randomly, they would typically score approximately 20% to 25% of the questions correctly.[19] To put these results in context, ChatGPT performed twice as well as one would expect from random guessing.

Our testing encompassed 59 text-based and 42 media-based OITE questions focused on the UE. The questions involving images introduce an added layer of complexity that current language models such as ChatGPT are not yet capable of evaluating independently. Previous studies examining ChatGPT's ability to answer OITE questions excluded questions with figures, diagrams, or charts.[20] Similar studies performed in radiology and ophthalmology also excluded images and videos from their analysis and included text-only questions in their studies.[21,22] Regardless of question format, ChatGPT achieved similar results in both question formats, with an accuracy of 42% in text-only scenarios and 43% in scenarios involving media questions. When examining specific prompt requirements, we observed comparable performance between text-only and media questions when using the OE, MCWJ, and MCNJ answer requirements. However, ChatGPT only produced identical responses for 79 (77%) of 102 questions when comparing

Table 1
Distribution of UE OITE Questions by Year

OITE Year	Total Questions	Text-Based Questions	Image-Based Questions
2020 OITE	27	14	13
2021 OITE	41	20	21
2022 OITE	34	25	9
Total included	102	59	43

Table 2
AI Performance on Hand- and UE-Related OITE Questions

OITE Question Type	Number of Correct Responses, n (%)
Text only	
OE	22 (37)
MCNJ	25 (42)
MCWJ	26 (46)
Answer without justification matches answer with justification	42 (71)
Average	25 (42)
Text and image	
OE	17 (40)
MCNJ	19 (44)
MCWJ	20 (47)
Answer without justification matches answer with justification	33 (77)
Average	19 (43)
Combined	
OE	38 (37)
MCNJ	46 (45)
MCWJ	46 (45)
Answer without justification matches answer with justification	79 (77)
Average	43 (42)

Table 3
Performance on OITE UE Questions of National and Institution Residents Stratified by PGY

Performer	Number Correct by OITE Year, n (%)			
	2020	2021	2022	Average
Institution PGY-1 residents	18 (44)	32 (62)	26 (59)	25 (56)
Institution PGY-2 residents	25 (63)	34 (65)	32 (73)	30 (67)
Institution PGY-3 residents	29 (73)	37 (71)	31 (70)	32 (71)
Institution PGY-4 residents	28 (70)	40 (77)	33 (75)	34 (74)
Institution PGY-5 residents	30 (75)	38 (73)	35 (80)	34 (76)
Nationwide PGY-1 residents	18 (44)	27 (52)	23 (53)	23 (50)
Nationwide PGY-2 residents	20 (51)	31 (60)	26 (60)	26 (57)
Nationwide PGY-3 residents	23 (57)	34 (66)	30 (67)	29 (64)
Nationwide PGY-4 residents	24 (61)	36 (69)	31 (72)	31 (68)
Nationwide PGY-5 residents	25 (63)	37 (71)	33 (74)	32 (70)

the MCWJ and MCNJ answer requirements. These results highlight the need for continued assessment of AI technology with respect to medical education, particularly when models that can incorporate imaging become more widely available.

One potential application of an AI tool is its use in medical education. Previous studies have demonstrated the successful implementation of AI in optimizing educational systems at the university level.[23] These studies have also indicated that proposed curriculum changes suggested by AI systems were well received by both faculty and students.[23] Furthermore, a separate study focusing on the utility of ChatGPT in healthcare education highlights its ability to reduce costs, improve communication skills, and provide personalized educational feedback and mentoring.[16] Previous research has indicated that ChatGPT exhibits enhanced proficiency across various fields and educational levels. Notably, ChatGPT 3.5 achieved success in passing the United States Medical Licensing Examination step 1, obtained a passing grade on law school examinations conducted at the University of Minnesota, achieved positive results on multiple Advanced Placement examinations, and came close to passing the Multistate Bar Examination.[14,18,24–26] Specifically, ChatGPT 3.5 achieved a score of 50% compared to the 25% that would be expected through random guessing on the Multistate Bar Examination.[26] In the context of these prior investigations, ChatGPT appears better suited at present to answer undergraduate- and graduate-level examination questions as opposed to postgraduate-level medical examination questions. Although not directly assessed with our present methodology, there may be a role for AI in vetting and validating questions written for standardized surgical examinations, particularly if this technology can achieve performance similar to that of trainees. At present, it appears that these AI tools may require further advancement and refinement prior to having a more substantial role in UE surgical education.

Our investigation possesses several limitations that warrant consideration. In our methodology, we converted media provided in the OITE into text when the original prompt lacked a reasonable explanation. This transcoding process was carried out independently by two authors. However, it is important to acknowledge that potential reliability biases may exist because we did not explore this aspect within the scope of our study. This methodology has not been previously used. Notably, there have been advancements in the development of context-sensitive, image-based AI algorithms that could potentially complement language models such as ChatGPT in the future. However, we did additionally categorize and assess text-only questions. At the time of data collection, ChatGPT version 4.0 was not publicly available, and it is uncertain if updated and improved versions would have influenced our results, particularly with respect to image analysis. Future updated versions of ChatGPT may improve accuracy and reliability in test-taking scenarios.[14] The rapid progress in AI technology suggests that the application of AI as an educational tool may become more feasible in the future; however, additional investigations will be required. Additionally, although the OITE is recognized as a comprehensive examination that all hand surgery residents must take, it may not entirely replicate actual clinical encounters with patients.[27]

In conclusion, ChatGPT exhibited relatively lower performance in answering questions within the hand/wrist and shoulder/elbow content domains of the OITE compared to PGY-1 orthopaedic surgery residents. At present, this technology may have limited applications in UE surgical education. As AI technology continues to rapidly develop, future investigations will be required to determine its role in UE surgical education.

Acknowledgments

The authors would like to thank Jessica Temple for her assistance.

References

1. Lackey WG, Jeray KJ, Tanner S. Analysis of the musculoskeletal trauma section of the Orthopaedic In-Training Examination (OITE). *J Orthop Trauma*. 2011;25(4):238–242.
2. Mankin HJ. The Orthopaedic In-Training Examination (OITE). *Clin Orthop Relat Res*. 1971;75:108–116.
3. Evaniew N, Holt G, Kreuger S, et al. The orthopaedic in-training examination: perspectives of program directors and residents from the United States and Canada. *J Surg Educ*. 2013;70(4):528–536.
4. Le HV, Wick JB, Haus BM, Dyer GSM. Orthopaedic In-Training Examination: history, perspective, and tips for residents. *J Am Acad Orthop Surg*. 2021;29(9):e427–e437.
5. American Academy of Orthopaedic Surgeons. Orthopaedic In-Training Examination (OITE) technical report 2020. Published November 2020. Accessed June 8, 2023. https://www.aaos.org/globalassets/education/product-pages/oite/oite-2020-technical-report_website.pdf
6. Grandizio LC, Huston JC, Shim SS, Graham J, Klena JC. Levels of evidence for hand questions on the Orthopaedic In-Training Examination. *Hand (N Y)*. 2016;11(4):484–488.
7. Grandizio LC, Huston JC, Shim SS, Parenti JM, Graham J, Klena JC. Levels of evidence have increased for musculoskeletal trauma questions on the Orthopaedic In-Training Examination. *J Surg Educ*. 2015;72(2):258–263.
8. Walsh CT, Grandizio LC, Klena JC, Parenti JM, Cush GJ. Levels of evidence for foot and ankle questions on the Orthopaedic In-Training Examination: 15-year trends. *J Surg Educ*. 2016;73(6):999–1003.
9. LeBrun DG, Premkumar A, Ellsworth B, Shen TS, Cross MB, Fufa DT. Analysis of hand surgery questions on Orthopedic In-Training Examination from 2014 to 2019. *Hand (N Y)*. 2022;17(5):975–982.
10. Bartlett LE, Klein B, White PB, et al. An updated analysis of shoulder and elbow questions on the Orthopedic In-Training Examination. *J Shoulder Elbow Surg*. 2022;31(11):e562–e568.
11. Martin AS, McMains MC, Shacklett AG, Awan HM. Hand surgery questions on the Orthopaedic In-Training Examination: analysis of content and reference. *J Hand Surg Am*. 2018;43(6):568.e1–568.e6.
12. 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(24):e1370–e1377.
13. Dougherty PJ, Walter N, Schilling P, Najibi S, Herkowitz H. Do scores of the USMLE Step 1 and OITE correlate with the ABOS Part I certifying examination?: a multicenter study. *Clin Orthop Relat Res*. 2010;468(10):2797–2802.
14. OpenAI. ChatGPT. Accessed June 7, 2023. https://openai.com/chatgpt
15. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst*. 2023;47(1):33.
16. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. 2023;11(6):887.
17. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Preprint repository is Research Square. Posted online February 28, 2023. rs.3.rs–2566942. https://doi.org/10.21203/rs.3.rs-2566942/v1
18. Gilson A, Safranek CW, Huang T, et al. 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.
19. American Academy of Orthopaedic Surgeons. Orthopaedic In-Training Examination (OITE) technical report 2022. Published January 25, 2023. Accessed June 7, 2023. https://www.aaos.org/globalassets/education/product-pages/oite/oite-2022-technical-report-20230125.pdf
20. Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery Examination? Orthopaedic residents versus ChatGPT. *Clin Orthop Relat Res*. 2023;481(8):1623–1630.
21. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmol*. 2023;141(6):589–597.
22. 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.
23. Zhang F. Design and application of artificial intelligence technology-driven education and teaching system in universities. *Comput Math Methods Med*. 2022;2022:8503239.
24. Choi JH, Hickman KE, Monahan A, Schwarcz DB. ChatGPT Goes to Law School (January 23, 2023). 71 Journal of Legal Education 387 (2022). Accessed December 5, 2023. http://dx.doi.org/10.2139/ssrn.4335905
25. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health*. 2023;2(2):e0000198.
26. Bommarito MJ, Katz DM. GPT takes the bar exam. *SSRN Electronic Journal*. Published online December 31, 2022. Accessed December 5, 2023. http://dx.doi.org/10.2139/ssrn.4314839
27. Van Heest AE, Armstrong AD, Bednar MS, et al. American Board of Orthopaedic Surgery's initiatives toward competency-based education. *JB JS Open Access*. 2022;7(2):e21.00150.