Comparison of ChatGPT–3.5, ChatGPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations

Patrick A. Massey, MD, MBA

Carver Montgomery, MD

Andrew S Zhang, MD

From the Department of Orthopaedic Surgery, Louisiana State University Health Sciences Center Shreveport, Shreveport, LA.

Correspondence to Dr. Massey patrick.massey@lsuhs.edu

J Am Acad Orthop Surg 2023;31:1173-1179

DOI: 10.5435/JAAOS-D-23-00396

ABSTRACT

Introduction: Artificial intelligence (AI) programs have the ability to answer complex queries including medical profession examination questions. The purpose of this study was to compare the performance of orthopaedic residents (ortho residents) against Chat Generative Pretrained Transformer (ChatGPT)-3.5 and GPT-4 on orthopaedic assessment examinations. A secondary objective was to perform a subgroup analysis comparing the performance of each group on questions that included image interpretation versus text-only questions.

Methods: The ResStudy orthopaedic examination question bank was used as the primary source of questions. One hundred eighty questions and answer choices from nine different orthopaedic subspecialties were directly input into ChatGPT-3.5 and then GPT-4. ChatGPT did not have consistently available image interpretation, so no images were directly provided to either AI format. Answers were recorded as correct versus incorrect by the chatbot, and resident performance was recorded based on user data provided by ResStudy.

Results: Overall, ChatGPT-3.5, GPT-4, and ortho residents scored 29.4%, 47.2%, and 74.2%, respectively. There was a difference among the three groups in testing success, with ortho residents scoring higher than ChatGPT-3.5 and GPT-4 ($P < 0.001$ and $P < 0.001$). GPT-4 scored higher than ChatGPT-3.5 ($P = 0.002$). A subgroup analysis was performed by dividing questions into question stems without images and question stems with images. ChatGPT-3.5 was more correct (37.8% vs. 22.4%, respectively, OR = 2.1, $P = 0.033$) and ChatGPT-4 was also more correct (61.0% vs. 35.7%, OR = 2.8, $P < 0.001$), when comparing text-only questions versus questions with images. Residents were 72.6% versus 75.5% correct with text-only questions versus questions with images, with no significant difference ($P = 0.302$).

Conclusion: Orthopaedic residents were able to answer more questions accurately than ChatGPT-3.5 and GPT-4 on orthopaedic assessment examinations. GPT-4 is superior to ChatGPT-3.5 for answering orthopaedic resident assessment examination questions. Both ChatGPT-3.5 and GPT-4 performed better on text-only questions than questions with images. It is unlikely that GPT-4 or ChatGPT-3.5 would pass the American Board of Orthopaedic Surgery written examination.

Background

Artificial intelligence (AI) chatbots are computer programs that have the ability to understand human language and maintain a conversation with users with detailed responses. Although the first chatbot was created in the 1960s, this technology has advanced markedly into the chatbots we recognize currently.[1] Chatbots on the market today including Socratic, Jasper, ChatGPT (Chat Generative Pretrained Transformer), and YouChat have advanced capabilities to answer questions, schedule appointments, carry a conversation, and even take and pass tests intended for professionals with years of study and training.[2,3]

ChatGPT is an AI chatbot created by OpenAI and launched in November 2022. Since its inception, millions have used the program to perform tasks and answer questions. The chatbot runs off of GPT-3.5 or GPT-4 (if you have a subscription), which are large language models that can use AI to respond to, and understand, the human language. GPT-4 is the latest model, released on March 14th, 2023, provided by OpenAI that is more advanced than the previous model (ChatGPT-3.5), which can reportedly solve more difficult problems with better accuracy because of improved reasoning logic and a much larger base of learned information.[4]

ChatGPT models have demonstrated the ability to successfully complete rigorous professional examinations in a number of medical and nonmedical specialties based primarily on the interpretation of textual questions and information.[5-7] Orthopaedic surgery in practice and on examinations is distinguished by the frequent need to synthesize imaging data in formulating treatment plans. Thus, we sought to examine the performance of these algorithms on a contemporary set of orthopaedic questions, which closely parallel the American Board of Orthopaedic Surgery (ABOS) examination, and to specifically analyze their performance on text-only questions versus questions requiring image interpretation. The purpose of this study was to compare the performance of orthopaedic residents (ortho residents) against ChatGPT-3.5 and GPT-4 on orthopaedic assessment examinations. A secondary objective was to perform a subgroup analysis comparing the performance of each group on questions that included image interpretation versus text-only questions.

Methods

The ResStudy orthopaedic examination question bank, endorsed by the American Academy of Orthopaedic Surgeons (AAOS), was used as the primary source of questions. This question database includes Orthopaedic In-Training Examination (OITE) questions, Orthopaedic Knowledge Update questions, and self-assessment examinations. Twenty questions were tested from each of nine sections, each a different orthopaedic subspecialty (Adult Reconstruction (Hip and Knee), Foot and Ankle, Trauma, Shoulder and Elbow, Hand and Wrist, Pediatrics, Adult Spine, Musculoskeletal Tumors and Diseases, and Sports Medicine) in both ChatGPT-3.5 and ChatGPT-4. An a priori power analysis was performed based on a pilot of the first 10 comparisons of correct answers for ChatGPT-3.5 and GPT-4. This determined that the number needed for a power of 0.8 was 155. Based on a predetermination to test evenly across all nine subspecialties, 20 questions were used per section, yielding 180 total questions or samples.

This study was conducted with a random selection process. The question bank software provides the option to randomly select questions from the total potential question pool. Questions were selected randomly to be answered by building a quiz in "Study Mode" with the options "unanswered," "correct," and "incorrect" selected to randomly choose from the total pool of questions, regardless of whether or not questions had been previously answered by the user. Question stems and answer choices were copied from the ResStudy question bank and pasted into the ChatGPT input bar. A new chat was opened for each question for the AI to try and prevent it from learning by reinforcement. All questions contained four answer choices and were labeled with A, B, C, and D to allow ChatGPT to differentiate between them.

ChatGPT did not have consistently available image interpretation, so no images were directly provided to

Table 1. **Orthopaedic Examination Scores of ChatGPT-3.5, GPT-4, and Orthopaedic Residents**

	Average	95% CI
ChatGPT-3.5	29.4%[a,b]	23.2%-36.4%
GPT-4	47.2%[a,b]	40.0%-54.5%
Orthopaedic residents	74.2%[a]	71.5%-76.9%

CI = confidence interval

[a] Orthopaedic residents scored higher than ChatGPT-3.5 and GPT-4 ($P < 0.001$ and $P < 0.001$, respectively).

[b] GPT-4 scored higher than ChatGPT-3.5 ($P = 0.002$).

either AI format. As of now, ChatGPT-4 only accepts images for research purposes, and this feature is not yet available to the public. However, questions including images were input in both Chatbox versions in the same manner as questions without images but with only the text in the prompt. The time for each answer response was recorded by starting a timer exactly when the question was submitted to ChatGPT and stopping the timer when the chatbot had completed its answer.

An answer was considered "correct" only if ChatGPT definitively provided a correct answer choice within its full answer. If ChatGPT refused to answer the question or provided the incorrect answer choice, it was scored as incorrect. After each question is answered in ResStudy, a percentage for each answer choice representing the percentage of ortho residents who also chose that answer is displayed. To compare ortho resident averages with ChatGPT-3.5 and GPT-4, the resident average correct for each question was averaged for each section. The averages for each section were then compared with the score for each section for ChatGPT-3.5 and GPT-4.

Statistical Analysis

Categorical data such as the answer choice being correct or incorrect were compared between ChatGPT-3.5 and GPT-4 using chi-square calculation with SPSS version 29 (IBM, Armonk, NY). Numerical data among the three groups were compared using ANOVA with post hoc testing using the Tukey test. Chi-square analysis was also used to compare the accuracy among the nine different subspecialties. The relationship between word count and correctness was evaluated with a regression analysis and the Pearson coefficient.

Results

Overall, ortho residents scored an average of 74.2% ± 5.6 with a 95% CI [71.5, 76.9] (Table 1). ChatGPT-3.5 and GPT-4 correctly answered 29.4% with a 95% CI [23.2%, 36.4%] and 47.2% with a 95% CI [40.0%, 54.5%], respectively. There was a difference among the three groups in testing success. Orthopaedic residents scored higher than ChatGPT-3.5 and GPT-4 ($P < 0.001$ and $P < 0.001$, respectively). GPT-4 scored higher than ChatGPT-3.5 ($P = 0.002$).

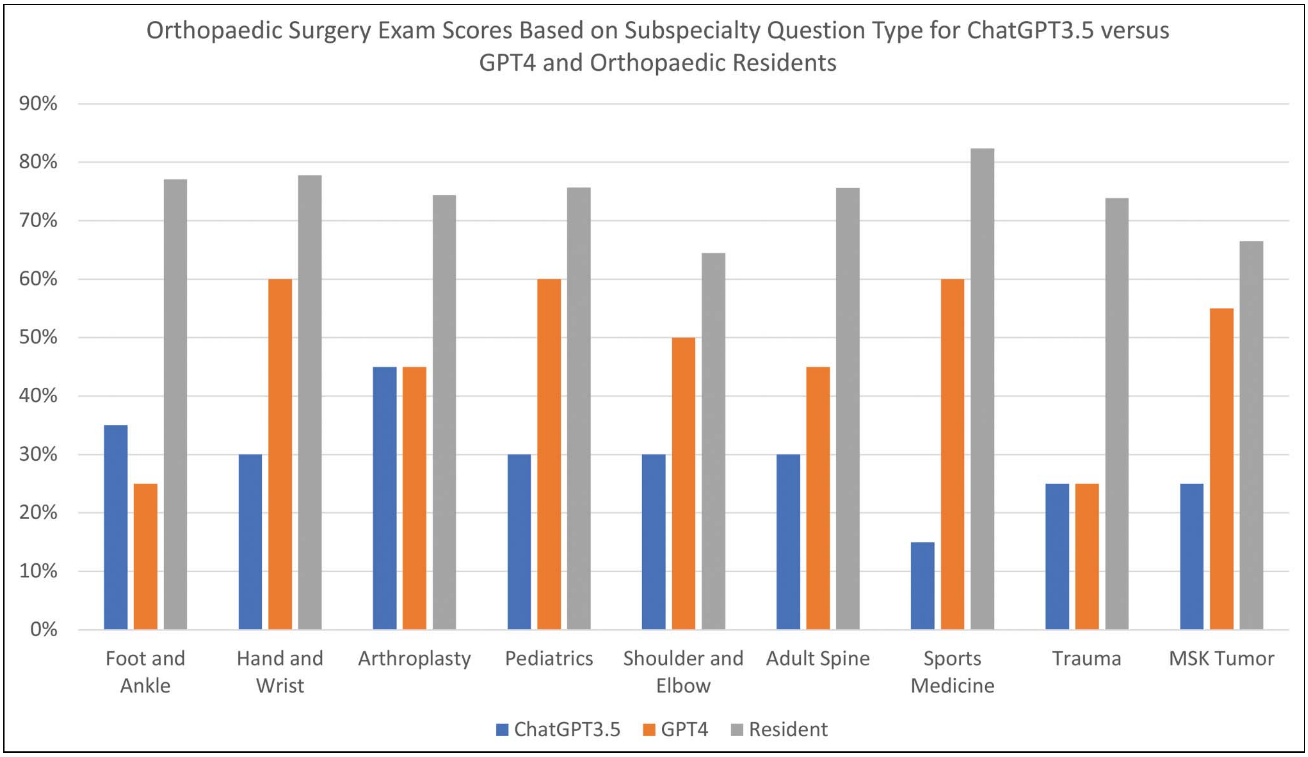
The average response time for ChatGPT-3.5 versus GPT-4 was 12.1 ± 4.7 seconds versus 31.8 ± 17.8 seconds, respectively. ChatGPT-3.5 was faster at responding to questions than GPT-4 ($P = 0.008$). There was no association between word count and correctness for each test question for ChatGPT-3.5 and GPT-4 ($P = 0.703$ and 0.718, respectively). There was a positive correlation between word count and correctness for ortho residents (R = 0.16, $P = 0.031$). Among the nine sections, there was no difference in the scores of GPT-4, ChatGPT-3.5, or ortho residents ($P = 0.131$, 0.755, and 0.067, respectively) (Figure 1). A post hoc power analysis using the discordant cell proportions of ChatGPT-3.5 and GPT-4 showed that with a sample size of 180, the power was 0.995.

Subgroup Analysis

A subgroup analysis was performed by dividing questions into question stems without images and question stems with images (Figure 2). There were 82 of the 180 questions (45.6%) that had no images in the question stem, whereas 98 (54.4%) did have images. ChatGPT-3.5 was more correct when comparing text-only questions versus questions with images (37.8% vs. 22.4%, respectively, OR = 2.1, $P = 0.033$). ChatGPT-4 was also more correct when comparing text-only versus image questions (61.0% vs. 35.7%, OR = 2.8, $P < 0.001$). These results are summarized in Table 2. Residents were 72.6% ± 19.9% correct on questions with no images and 75.5% ± 17.3% correct on questions with images, with no difference based on the presence of images ($P = 0.302$).

When evaluating only questions with images, ChatGPT-3.5, GPT-4, and ortho residents had no differences in accuracy among the nine sections ($P = 0.975$,

Figure 1



Bar graph showing orthopaedic surgery examination scores based on the subspecialty question type for ChatGPT-3.5 versus GPT-4 and residents. Arthroplasty = adult reconstruction (hip and knee), GPT = Generative Pretrained Transformer, MSK Tumor = musculoskeletal tumors and disease

Table 2. **Orthopaedic Assessment Examination Scores of Residents, ChatGPT-3.5, and GPT-4 for Questions Requiring Image Interpretation Versus Text-Only Questions**

	Questions With Images			Text-Only Questions			
	Mean	SD	95% CI	Mean	SD	95% CI	*P*
ChatGPT-3.5	22.40%	—	15.1% - 31.4%	37.80%	—	27.9% - 48.6%	*P* = 0.033
GPT-4	35.70%	—	26.8% - 45.5%	61.00%	—	50.2% - 71.0%	*P* < 0.001
Orthopaedic residents	75.50%	17.30%	73.0% - 78.0%	72.60%	19.90%	69.7% - 75.5%	*P* = 0.302

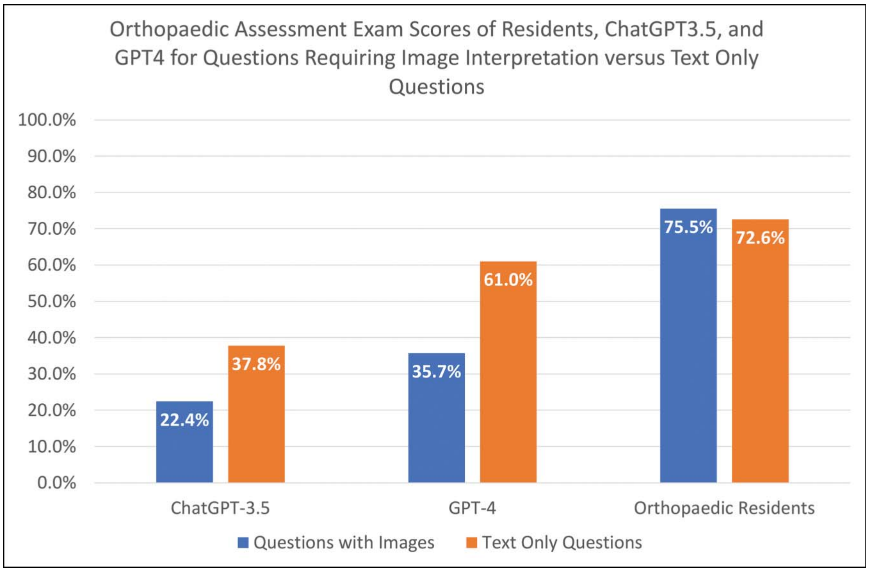
CI = confidence interval

P = 0.195, and *P* = 0.342, respectively). When evaluating only questions that had no images, ChatGPT-3.5 and GPT-4 had no differences in accuracy among the nine sections (*P* = 0.495 and *P* = 0.600, respectively). There was a difference among the nine sections with correctness for ortho residents when evaluating only questions with no images (*P* = 0.029). Post hoc analysis showed that only shoulder/elbow and adult reconstruction were significantly different, with ortho residents answering more adult reconstruction questions correct (*P* = 0.041).

Discussion

AI chatbot technology has developed at a rapid rate and has been increasingly used across various platforms in society.[8,9] ChatGPT, in particular, has become one of the fastest growing computer applications in history, gaining 100 million active users in just 2 months after becoming available to the public.[10] Like other forms of AI, ChatGPT is trained on an abundance of information including peer-reviewed journal articles, texts, news articles, and online resources.[11] Although this technology

Figure 2



Bar graph showing orthopaedic assessment examination scores of residents, ChatGPT-3.5, and GPT-4 for questions requiring image interpretation versus text-only questions. GPT = Generative Pretrained Transformer

remains in relative infancy, its ability to assimilate information and its accuracy continue to improve.

In the United States, orthopaedic surgery remains one of the most competitive subspecialty residencies to obtain with an average US Medical Licensing Examination (USMLE) Step 2 score of 259 and a match rate of approximately 60% of those who apply.[12,13] Despite the high-performing aptitude of these orthopaedic residents, there was still a 14% failure rate for the 2021 ABOS Part I Examination.[14] These written examinations are challenging, requiring a strong foundation of orthopaedic knowledge in tandem with critical thinking skills and advanced image interpretation. Given the proven test-taking abilities of ChatGPT across other disciplines, we sought to evaluate its performance on a contemporary set of orthopaedic questions, which closely parallel the ABOS examination, and to specifically analyze their performance on text-only questions versus questions requiring image interpretation.

In the current study, ortho residents outperformed ChatGPT-3.5 and GPT-4 at answering orthopaedic examination questions from a standardized popular question bank in this study. Overall, ortho residents were able to answer more questions correctly, whether an image was included in the question stem or not. ChatGPT-3.5 was 2.1 times more likely and GPT-4 was 2.8 times more likely to answer a question correctly when it was a text-only question compared with a question requiring image interpretation. This notable difference alludes to the comprehensive skill set required for answering orthopaedic assessment questions and can perhaps be translated to clinical practice. Unlike assessment examinations in other disciplines, orthopaedic examinations require particular scrutiny of radiographic images in conjunction with interpretation of clinical vignettes, which reflects the critical thinking needed every day by orthopaedic surgeons that may still remain outside the ability of these chatbots at this time.

It has already been proven that both ChatGPT-3.5 and GPT-4 models can successfully pass multiple standardized tests. Studies have shown that ChatGPT can achieve near-perfect scores on the Scholastic Aptitude Test,[4] pass a Master of Business Administration Final Examination on Operations Management at the University of Pennsylvania's Wharton School of Business,[5] and score 4s and 5s on multiple Advanced Placement (AP) examinations.[4] Not only can these AI programs pass nonmedical examinations, but they also pass medical licensing examinations. They have performed well on the MBBS and passed the USMLE Step 1 and Step 2 examinations.[6,15,16] GPT-4 has also placed in the 90th percentile on the Uniform Bar Examination.[4] It has demonstrated aptitude on medical specialty examinations, passing a Dutch Family Medicine Examination[17] and successfully solving higher-order pathology questions.[18]

Recently, a study compared both ChatGPT-3.5 and GPT-4 on the neurological surgery (ABNS) Self-Assessment

Examination 1.[19] The authors found that ChatGPT-3.5 and GPT-4 scored 73.4% and 83.4%, respectively. This is higher than the performance in the current study on orthopaedic assessment questions. In the study by Ali et al., 22% of the neurosurgery test questions had images, whereas at least 50% of orthopaedic examination questions have images. This may explain why ChatGPT-3.5 and GPT-4 had more difficulties in the current study, having difficulty answering orthopaedic examination questions requiring interpretation of images. Ali et al. also demonstrated that both ChatGPT-3.5 and GPT-4 had improved performance when no images were included in the questions. The authors demonstrated examination scores of 80.2% with ChatGPT-3.5 and 91% with GPT-4 with text-only questions. Although both AI models performed better with text-only questions on a neurosurgery examination, the results on orthopaedic examination questions were much lower with text-only questions.

In a collaboration between the ABOS and the AAOS, a linking study was conducted to correlate the content of the ABOS Part I Certifying Examination, the test required to obtain board certification, with the content of the AAOS OITE, taken during residency training. The linking study found that a minimum score of 69.2% correct on the OITE would approximate a passing designation on the ABOS Part I Certifying Examination.[20] As ResStudy comprised actual OITE questions, based on their overall test averages, residents would likely pass the examination created in our study, whereas ChatGPT-3.5 and GPT-4 would not. Even with text-only questions, GPT-4 obtained a 61% score, still below the minimum threshold needed to pass the ABOS Part I, whereas the average orthopaedic surgery resident scored well enough to pass.

However, the implications and potential of ChatGPT are still quite promising. With just one newer iteration of programming, GPT-4 demonstrated that although it took slightly longer to answer questions, it was still able to answer many more questions correctly compared with ChatGPT-3.5. GPT-4 was able to answer greater than or equal to the number of questions correctly by ChatGPT-3.5 overall and across all nine subspecialties. GPT-4 showed particular improvement when answering questions requiring no image interpretation. Although GPT-4 was still unable to reach a minimum passing score when no images were present, it did better answering 61% of questions correctly. It is notable that in our study, GPT-4 was not as successful as a multitude of other studies such as the neurosurgery examination, bar examination, and USMLE. This could be due to inadequate modeling by the engineers and lack of the orthopaedic literature used to train it. It could also be the complexity of questions in orthopaedic examinations. These AI models may also play a role in the future for medical question writing.[21] Although models may be developed to write standardized test questions, there are a variety of ethical and legal concerns. Finally, it should be noted that orthopaedic residents may now have access to using AI to assist them in answering test questions. Currently, standardized protocols for taking the ABOS Part I are already in place to have in-person proctors and computer systems that lock out access to third-party resources, such as GPT-4. However, when orthopaedic programs require residents take orthopaedic examinations for program-specific evaluation purposes, protocols such as mandated proctoring should be considered as AI usage during examination is undoubtedly a rising threat to testing integrity.

Limitations

This study is not without limitations. Although ChatGPT-3.5 and GPT-4 were evaluated using categorical data in the form of a correct or incorrect answer, ortho resident data were based on the national averages of correct answers for orthopaedic residents. To compare categorical data versus numerical data, the sections were averaged to return an average number or percentage for each section. This led to a smaller sample size, only when comparing the averages of the sections, but still showed notable differences, so this potential type II error was less of a concern. In addition, although comparisons of all 180 total questions were sufficiently powered, it should be noted that comparisons between each section with 20 questions in each were likely underpowered. In addition, there could have been the presence of human error for the timing data because each response was timed manually with a timer. However, this was all done with the same person performing the timing measurements in a consistent manner.

Conclusion

Orthopaedic residents were able to answer more questions accurately than ChatGPT-3.5 and GPT-4 on orthopaedic assessment examinations. GPT-4 is superior to ChatGPT-3.5 for answering orthopaedic resident assessment examination questions. Both ChatGPT-3.5 and GPT-4 performed better on text-only questions than questions with images. It is unlikely that GPT-4 or ChatGPT-3.5 would pass the American Board of Orthopaedic Surgery written examination.

References

1. Weizenbaum J: ELIZA—a computer program for the study of natural language communication between man and machine. *Commun ACM* 1966;9:36-45

2. Parviainen J, Rantala J: Chatbot breakthrough in the 2020s? An ethical reflection on the trend of automated consultations in health care. *Med Health Care Philos* 2022;25:61

3. Gkinko L, Elbanna A: The appropriation of conversational AI in the workplace: A taxonomy of AI chatbot users. *Int J Inf Manag* 2023;69:102568

4. GPT-4. Accessed April 13, 2023. https://openai.com/product/gpt-4.

5. Terwiesch C. Would chat GPT get a Wharton MBA? New white paper by Christian Terwiesch.

6. 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:e0000198

7. Katz DM, Bommarito MJ, Gao S, Arredondo P. GPT-4 Passes the bar exam. doi:10.2139/ssrn.4389233

8. Grudin J, Chatbots JacquesR: Humbots, and the quest for artificial general intelligence, in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Association for Computing Machinery, 2019, pp 1-11

9. Adamopoulou E, Moussiades L: Chatbots: History, technology, and applications. *Machine Learn Appl* 2020;2:100006

10. Hu K: *ChatGPT sets record for fastest-growing user base - analyst note*. Reuters.Available at: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/. Accessed April 12, 2023.

11. Where does ChatGPT get its information from? Scribbr. Available at: https://www.scribbr.com/frequently-asked-questions/chatgpt-information/ Accessed April 12, 2023.

12. Lubowitz JH, Brand JC, Rossi MJ: The 2022 orthopaedic surgery residency match leaves many qualified candidates unmatched. *Arthrosc J Arthroscopic Relat Surg* 2022;38:1755-1757

13. Rothfusz CA, Emara AK, Ng MK, et al.: The orthopaedic interview spreadsheet: Classification and comparison to the national resident matching program. *J Surg Educ* 2022;79:112-121

14. Exam Stats: *ABOS*.Available at: https://www.abos.org/certification/part-i/examination-statistics/ Accessed April 12, 2023.

15. 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

16. Subramani M, Jaleel I, Krishna Mohan S: Evaluating the performance of ChatGPT in medical physiology university examination of phase I MBBS. *Adv Physiol Educ* 2023;47:270-271

17. Morreel S, Mathysen D, Aye VerhoevenV: AI! ChatGPT passes multiple-choice family medicine exam. *Med Teach* 2023:1

18. Sinha RK, Deb Roy A, Kumar N, Mondal H: Applicability of ChatGPT in assisting to solve higher order problems in pathology. *Cureus* 2023;15:e35237

19. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PLCielo D, Oyelese AA, Doberstein CE, Telfeian AE, Gokaslan ZL, Asaad WFet al. Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. *Neurosurgery*2023 Jun 12. doi:10.1227/neu.0000000000002551. Epub ahead of print. PMID: 37306460.

20. Incrocci M. Orthopaedic In-Training Examination (OITE) Technical Report 2021.

21. Biswas S: ChatGPT and the future of medical writing. *Radiology* 2023;307:e223312