



Contents lists available at ScienceDirect

Journal of Hand Surgery Global Online

journal homepage: www.JHSGO.org

Original Research

ChatGPT's Performance on the Hand Surgery Self-Assessment Exam: A Critical Analysis

Yuri Han, BA, * Hassaam S. Choudhry, BA, † Michael E. Simon, MD, * Brian M. Katt, MD *

* Rutgers Robert Wood Johnson Medical School, New Brunswick, NJ

† Rutgers New Jersey Medical School, Newark, NJ



ARTICLE INFO

Article history:

Received for publication July 7, 2023

Accepted in revised form November 21, 2023

Available online January 2, 2024

Key words:

AI
ChatGPT
Certification
Education
Self-assessment

Purpose: To assess the performance of Chat Generative Pre-Trained Transformer (ChatGPT) when answering self-assessment exam questions in hand surgery and to compare correct results for text-only questions to those for questions that included images.

Methods: This study used 10 self-assessment exams from 2004 to 2013 provided by the American Society for Surgery of the Hand (ASSH). ChatGPT's performance on text-only questions and image-based questions was compared. The primary outcomes were ChatGPT's total score, score on text-only questions, and score on image-based questions. The secondary outcomes were the proportion of questions for which ChatGPT provided additional explanations, the length of those elaborations, and the number of questions for which ChatGPT provided answers with certainty.

Results: Out of 1,583 questions, ChatGPT answered 573 (36.2%) correct. ChatGPT performed better on text-only questions than image-based questions. Out of 1,127 text-only questions, ChatGPT answered 442 (39.2%) correctly. Out of the 456 image-based questions, it answered 131 (28.7%) correctly. There was no difference between the proportion of elaborations among text-only and image-based questions. Although there was no difference between the length of elaborations for questions ChatGPT got correct and incorrect, the length of elaborations provided for image-based questions were longer than those provided for text-only questions. Out of 1,441 confident answers, 548 (38.0%) were correct; out of 142 unconfident answers, 25 (17.6%) were correct.

Conclusions: ChatGPT performed poorly on the ASSH self-assessment exams from 2004 to 2013. It performed better on text-only questions. Even with its highest score of 42% for the year 2012, the AI platform would not have received continuing medical education credit from ASSH or the American Board of Surgery. Even when only considering questions without images, ChatGPT's high score of 44% correct would not have "passed" the examination.

Clinical relevance: At this time, medical professionals, trainees, and patients should use ChatGPT with caution as the program has not yet developed proficiency with hand subspecialty knowledge.

Copyright © 2023, THE AUTHORS. Published by Elsevier Inc. on behalf of The American Society for Surgery of the Hand. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Chat Generative Pre-Trained Transformer (ChatGPT), a language model by OpenAI, has received significant attention since its public release in 2022. The artificial intelligence (AI) platform converses with the user and answers inquiries by drawing textual data from the internet through the year 2021.¹ ChatGPT has

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

Corresponding author: Brian M. Katt, MD, Department of Orthopaedic Surgery, Rutgers Robert Wood Johnson Medical School, 125 Paterson St., New Brunswick, New Jersey 08901.

E-mail address: briankatt@gmail.com (B.M. Katt).

<https://doi.org/10.1016/j.jhsg.2023.11.014>

2589-5141/Copyright © 2023, THE AUTHORS. Published by Elsevier Inc. on behalf of The American Society for Surgery of the Hand. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

already been used in medical education for a variety of purposes, ranging from automatic exam scoring to personalized learning.¹ ChatGPT has also been used in the research process, such as supporting researchers in designing and conducting epidemiology studies.² ChatGPT has previously been found to perform at the level of a third-year medical student when answering questions of the United States Medical Licensing Examination Step 1 and 2 examinations.³ ChatGPT has also been studied in the context of clinical and surgical practices, from streamlining appointments for surgical patients to connecting patients from remote or rural areas to providers and providing patients details of their operative care.^{4,5}

Previous research has evaluated ChatGPT's performance on specialty-specific knowledge. Mihalache et al⁶ found that ChatGPT failed to provide substantial assistance in board certification preparation when using the free trial from OphthoQuestions for ophthalmic board certification preparation due to poor performance. On the other hand, Kalantar et al⁷ advocated for the use of ChatGPT in teaching orthopedic trauma to residents due to the AI's ability to interact with residents, identify areas of difficulty, and personalize its guidance to each user. On the whole, web-based tools are becoming increasingly common during resident education and in the preparation for certification exams. For instance, competency-based education for orthopedic surgery has used smartphone-optimized web-based technology to bolster the use of learning programs.⁸

Similarly, board certification and recertification processes have also increased their reliance on web-based platforms, including a variety of online modules and self-assessments through which physicians can continue to learn and stay up-to-date with current medical knowledge.⁹ Lum et al¹⁰ found that ChatGPT performed at a second-year orthopedic resident level when answering 3,900 questions based on the Orthopedic In-Training Examination (OITE), yet fell short of passing for the American Board of Orthopedic Surgery Part I certification exam. It performed the best in basic science and the worst in anatomy, hand surgery, and hip and knee reconstruction. However, this study excluded image-based questions from the analysis due to the AI system's inability to process visual data.

The American Society for Surgery of the Hand (ASSH) has developed the yearly self-assessment examination, an educational tool that gauges clinical knowledge in the subspecialty. A score of 50% is required for continuing medical education credits for all test-takers except general surgeons for whom a 75% is required for credit.¹¹ Although many prior studies on the performance of AI on medical examinations excluded image-based questions, the present study aims to assess ChatGPT's performance comprehensively on self-assessment questions pertaining to hand surgery while also comparing its performance on text-only and image-based questions. We decided to include image-based questions to determine whether ChatGPT can use the text in the question stems of such questions to select the correct answer. We hypothesized that ChatGPT would not pass hand surgery self-assessment exams, but that the AI platform would perform better on text-only questions compared to questions with images. A potential strength of ChatGPT, especially when used as an educational tool, is its ability to provide elaborations, explanations, and context. Therefore, this study also aimed to explore ChatGPT's elaborations to examination questions. We hypothesized that ChatGPT would provide longer explanations on those questions it answered correctly.

Materials and Methods

This study used the ASSH self-assessment exams from 2004 to 2013. Questions with image-based answer choices, tables, videos, and multiple-part questions were excluded. However, questions with images in the question stem were included. Out of the 1,605 questions available, 1,583 were included after exclusion of questions involving image-based answer choices, tables and/or videos, and linked questions. There were 1,127 (71.2%) text-only questions and 456 (28.8%) image-based questions.

The primary outcomes of this study were total score, score on text-only questions, and score on image-based questions. The score was measured as the proportion of correct answers. Secondary outcomes included the proportion of questions for which ChatGPT provided elaborations and the character length of those elaborated

answers. The percentages of confident and unconfident answers were also measured and compared. ChatGPT 3.5 was used.

Questions were entered into ChatGPT individually with site refresh before each question from May 21, 2023 to May 25, 2023. The total number of questions inputted into ChatGPT for data collection were roughly divided equally per day within this time-frame, with all questions from each self-assessment exam inputted in one sitting. Conversation histories were cleared before each question to avoid ChatGPT's use of previous questions for subsequent questions. Questions involving images were entered into ChatGPT without alteration to the text of the question; however, the image itself could not be included because ChatGPT cannot process image files.

If ChatGPT reported it could not answer a question but took a likely guess that was the correct answer, it was considered "correct." If ChatGPT reported it could not answer an image-based question because it was not provided a question, it was considered "incorrect." If ChatGPT provided additional details beyond its selection of an answer choice—or, for image-based questions, anything additional to its statement of inability to answer—those additional details were considered "elaboration." ChatGPT's answers were considered "unconfident" based on whether the AI platform referred to medical professional knowledge in addition to the answer it provided; otherwise, the answers were considered "confident" (Table 1).

Student *t* tests were performed to assess for differences between ChatGPT's performance on text-only and image-based questions, the proportion of answers with elaborations in response to text-only and image-based questions, the proportion of answers with elaborations for correctly and incorrectly answered questions, the difference in elaboration length between text-only and image-based questions, the difference in elaboration length between correctly and incorrectly answered questions, the proportion of confident and unconfident answers, and the difference between the proportion of correct confident and correct unconfident answers. *P* values were 2-tailed, and a *P* value < .05 was considered statistically significant. Pearson's correlation statistic was also used to determine the difference in ChatGPT's performance on the self-assessment exams based on exam year (2004–2013).

Results

In total, ChatGPT answered 573 (36.2%) questions correctly (Table 2). There was no significant difference in ChatGPT's performance across the different exams by year (*P* = .512).

Of the 1,127 text-only questions, ChatGPT answered 442 (39.2%) correctly (Table 3). Of the 456 image-based questions, ChatGPT answered 131 (28.7%) correctly (Table 4). ChatGPT performed better on text-only questions than image-based questions, with a difference of 9.10% (*P* = .007) (Fig. 1).

Of the 1,583 questions included in the study, ChatGPT provided elaborations for 934 (59.0%), with the majority of elaborations provided for questions from the 2007–2013 exams (Table 2, with example in Table 5). Elaborations were provided for 616 (54.7%) of the 1,127 text-only questions and 318 (69.7%) of the 456 image-based questions. There was no statistically significant difference between the proportion of elaborations for text-only and image-based questions (*P* = .541). Similarly, there was no statistically significant difference between the proportions of elaborations for questions ChatGPT answered correctly and incorrectly (*P* = .715).

The average length of elaborations was 736 characters. The average character count for answers to image-based questions was greater than that of text-only questions (difference: 128; *P* = .005). There was no statistically significant difference in elaboration

Table 1
ChatGPT's Answer Types With Examples

ChatGPT's Answer Type	Example
ChatGPT admitted it could not answer a question but took a likely guess that was the correct answer	"Since the figure is not provided, it is difficult to make an accurate assessment. However, based on the options provided, the most likely answer to the prompt would be: E. Flexion Thumb flexion involves the movement of the thumb toward the palm, which is an essential function for grasping objects. Tendon transfers can be performed to restore thumb flexion if there is damage or loss of function in the original muscles responsible for this movement."
ChatGPT could not answer an image-based question because it was not provided figures	"I'm sorry, but as an AI text-based model, I don't have the capability to view or interpret visual content such as Figure 1 . In order to identify the clinical sign depicted in the patient's left hand, I would need a verbal description or more information about the sign. Please provide additional details or describe the sign, and I'll do my best to assist you."
ChatGPT's elaboration	"Without the visual reference to Figure 1 , it is difficult to provide a definitive answer. However, I can provide some general information on the potential complications following suture ligation of abnormalities. A. Recurrent ulceration: This complication may occur if there is poor wound healing or if there is ongoing pressure or friction on the area after the sutures are placed. It could be a possible complication depending on the nature of the abnormality in Figure 1 . B. Recurrent infection: If the sutures are not properly sterilized or if there is poor wound care after the procedure, there is a risk of developing a recurrent infection. C. Rudimentary digit: This complication refers to the formation of an underdeveloped or non-functional digit. It may be relevant if the abnormality in Figure 1 involves the digits or limbs. D. Bony overgrowth: Sometimes, after the sutures are placed, there can be abnormal bone growth in the area. This could occur if the abnormality involves bones or joints. E. Digital neuroma: A digital neuroma is a painful condition caused by the formation of a nerve bundle. This may be relevant if the abnormality in Figure 1 involves nerves or if there is nerve damage during the procedure. To accurately determine the most common complication following suture ligation of the abnormality in Figure 1 , a visual assessment and a specific understanding of the abnormality are necessary. It is best to consult with a healthcare professional or refer to relevant medical literature for an accurate answer based on the specific details of Figure 1 ."
ChatGPT's "unconfident" answer	"Since Figure 1 is not provided, I am unable to directly analyze the malunion and provide a specific answer. However, I can provide you with general information on the treatment options for malunion and their potential effects. Please keep in mind that this information may not be specific to the given case without visual reference. A malunion refers to a situation where a fractured bone heals in an abnormal position, resulting in misalignment. The appropriate treatment for a malunion depends on various factors, such as the location and severity of the malunion, the patient's symptoms, and functional limitations. Based on the given answer choices, the best option for correcting a malunion and its potential effects would be: B. Improve arc of motion and decrease symptoms Surgical intervention or other corrective measures for malunion aim to improve the alignment of the fractured bone and restore its normal position. By doing so, it can lead to improved range of motion and decreased symptoms. The realignment of the bone can relieve pain, improve function, and restore normal biomechanics. It is important to note that the specific outcome of a treatment for malunion can vary depending on the individual case and the chosen treatment approach. Consulting with a medical professional who can evaluate the specific malunion and provide appropriate treatment recommendations would be the best course of action."

length between questions ChatGPT answered correctly versus incorrectly ($P = .379$).

Of the 1,583 questions, ChatGPT provided 1,441 (91.0%) confident answers and 142 (8.97%) unconfident answers. There was a significant difference between the percent of confident answers and the percent of unconfident answers ($P < .001$). Out of the 1441 confident answers, 548 (38.0%) were correct; out of 142 unconfident answers, 25 (17.6%) were correct ($P = .008$).

Discussion

Although ChatGPT's use in medical education is increasing and promising, it is important to recognize that ChatGPT is performing poorly on questions related to subspecialty knowledge in hand surgery.

Previous studies found that ChatGPT was able to answer correctly 51.3% of the time on practice questions for the OITE.¹⁰ Among the subspecialties it performed the worst in were anatomy (32.6%), hand surgery (36.8%), and hip and knee reconstruction (39.6%). Specifically, it answered 53 out of 144 questions in hand surgery correctly.¹⁰ In this study, ChatGPT answered 36.2% of

1,583 questions correctly, a finding consistent with previous results.

Prior studies have excluded image-based questions in their analysis of ChatGPT's performance on medical knowledge assessments because the AI platform cannot process visual input. Although ChatGPT can answer some image-based questions correctly, this study demonstrates that the AI system's performance is significantly better on text-only questions compared to image-based questions. An assessment of ChatGPT's true performance on image-based questions will require the development of an image-interpretation program. ChatGPT's most recent version of its visual program DALL-E3 was made available in October 2023. Although this program cannot interpret images, it can generate images based off text descriptions.¹¹ Image interpretation remains an emerging field in the AI realm. Although progress has been made in creating platforms that automatically interpret medical images, a model for generating natural language text (like ChatGPT) in image interpretation has yet to be developed.¹² The results of our present study can serve as a comparison for future work on AI image-interpreting and generation investigations.

Table 2
Question Characteristics and ChatGPT's Total Performance Based on Year

Exam Year	Total No. of Questions	No. of Correct Responses (%)	No. of Text-Only Questions	No. of Image-Based Questions	No. of Elaborated Answers	Mean Elaboration Length (SD)
2004	120	49 (40.8)	80	40	8	481.5 (191.7)
2005	120	40 (33.3)	89	31	2	652 (128.7)
2006	113	38 (33.6)	102	11	1	383 (N/A)
2007	143	52 (36.3)	104	39	98	736.6 (440.8)
2008	150	52 (34.7)	95	55	112	660.7 (345.3)
2009	148	45 (30.4)	95	53	118	770.7 (488.2)
2010	197	60 (30.5)	128	69	164	758.3 (437.7)
2011	198	75 (37.9)	127	71	151	733.1 (462.9)
2012	199	84 (42.2)	163	36	141	712.7 (464.9)
2013	195	78 (40.0)	144	51	139	769.4 (507.6)

Table 3
ChatGPT's Performance on Text-Only Questions

Exam Year	No of Questions	No of Correct Answers (%)	No of Elaborated Answers (%)	Mean Elaboration Length (SD)
2004	80	34 (42.5)	7 (8.75)	487.3 (212.5)
2005	89	34 (38.2)	1 (1.12)	561 (N/A)
2006	102	34 (33.3)	1 (0.980)	383 (N/A)
2007	104	39 (37.5)	69 (66.3)	742.6 (408.0)
2008	95	36 (37.9)	68 (71.6)	617.8 (299.0)
2009	95	37 (38.9)	72 (75.8)	784.3 (473.5)
2010	128	43 (33.6)	103 (80.5)	699.2 (404.9)
2011	127	56 (44.1)	88 (69.3)	674.6 (451.2)
2012	163	69 (42.3)	109 (63.2)	674.0 (428.4)
2013	144	60 (41.7)	98 (68.1)	725.7 (477.7)

Table 4
ChatGPT's Performance on Image-Based Questions

Exam Year	No of Questions	No of Correct Answers (%)	No of Elaborated Answers (%)	Mean Elaboration Length (SD)
2004	40	15 (37.5)	1 (2.50)	441 (N/A)
2005	31	6 (19.4)	1 (3.23)	743 (N/A)
2006	11	4 (36.4)	0	N/A
2007	39	13 (33.3)	29 (74.4)	813.6 (515.0)
2008	55	16 (29.1)	44 (80.0)	727.1 (401.4)
2009	53	8 (15.1)	46 (86.8)	744.0 (519.5)
2010	69	17 (24.6)	61 (88.4)	858.1 (474.9)
2011	71	19 (26.8)	63 (88.7)	814.7 (470.1)
2012	36	15 (41.7)	32 (88.9)	844.5 (559.8)
2013	51	18 (35.3)	41 (80.4)	873.8 (565.3)

In a similar study assessing ophthalmic knowledge, ChatGPT provided additional explanations to 63% of questions with no significant difference between correctly and incorrectly answered questions.⁶ Likewise, in this study, ChatGPT provided elaborations for 59% of questions with no statistically significant difference between correctly and incorrectly answered questions. Because elaborations are generally viewed as one of ChatGPT's more helpful skills, our results suggest that the AI platform may provide inaccurate elaborations in support of incorrect answers. Given that ChatGPT provided elaborations for questions it answered incorrectly just as often as it did for questions it answered correctly, such elaborations may be harmful. Medical professionals and trainees must take this into consideration when using ChatGPT for hand surgery self-assessment preparation.

Although ChatGPT provided elaborations to image-based questions just as often as it did to text-only questions in this study, the elaborations for image-based questions were significantly longer. For some image-based questions that ChatGPT answered incorrectly because it could not access the figures, the AI platform still provided additional details relevant to each answer choice that could guide the user to make their own educated selection

(Table 5). This may have contributed to the longer elaboration character counts for image-based questions.

The majority of the answers ChatGPT provided were considered confident. While ChatGPT encouraged the user to consult medical professionals before taking its selected answers to heart, it only did so 8.97% of the time, despite only getting 36.2% of the questions correct. For the questions to which ChatGPT provided confident answers, only 38% were correct. This has significant implications not only for the medical professionals who may use this program but also for the patients and their family members who may turn to such software to educate themselves about their medical conditions and to help guide their decision making when seeking medical care. Regarding the case for "Dr. Google," approximately half of those who use web-based health information would like further assistance finding health information (navigational needs), and such consumers look for 1) more types of information and 2) information from a greater variety of sources compared to those without navigational needs.¹³ This, along with the increased use of ChatGPT in health contexts as well as the changing nature of trust placed on online health resources, emphasizes the importance of ensuring accurate dissemination of knowledge through this new

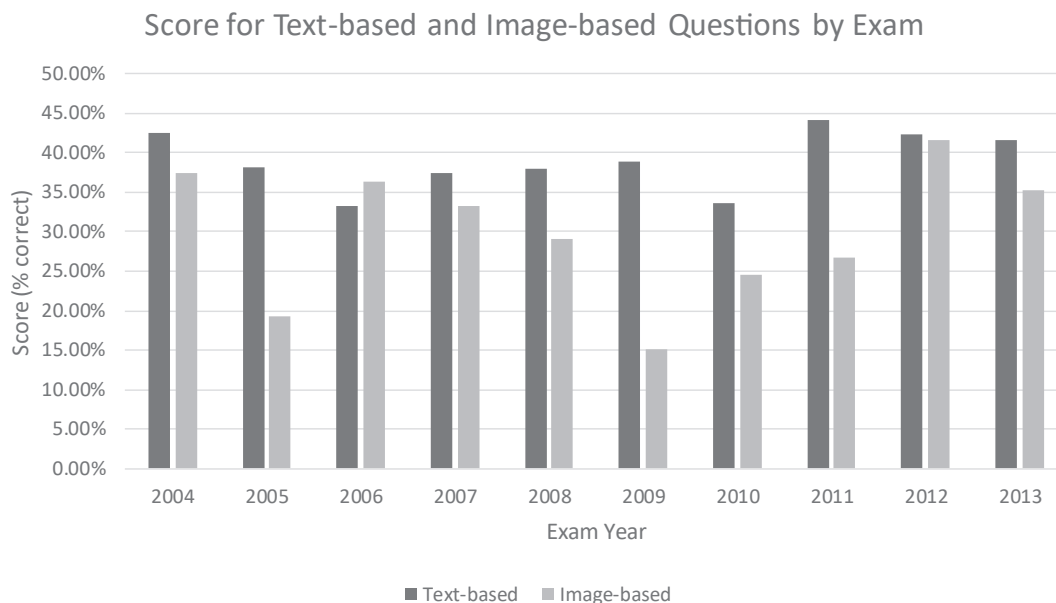


Figure 1. ChatGPT's score for text-based and image-based questions for each exam based on year.

Table 5
Example of ChatGPT Providing Additional Details for Each Answer Choice to a Question It Could Not Answer Because It Did Not Have Access to the Relevant Figures

Question	ChatGPT's answer
<p>A 30-year-old woman presents with the injury shown in Figure 1. The distal phalanx is fractured with exposed bone. Based on the available evidence in the literature the treatment that will predictably yield satisfactory results is:</p> <p>A. Moberg flap B. Skin graft C. V-Y advancement flap D. Thenar flap E. Microvascular free toe flap</p>	<p>"As an AI text-based model, I don't have the capability to view or analyze images. Therefore, I cannot refer to Figure 1 or evaluate the specific injury mentioned in the prompt. However, I can provide general information about the treatment options listed and their applications for certain types of injuries.</p> <p>A. Moberg flap: A Moberg flap is a surgical procedure that involves transferring a local skin flap from the base of the thumb to reconstruct the defect. It is commonly used for soft tissue defects of the thumb.</p> <p>B. Skin graft: A skin graft involves taking a piece of skin from one area of the body (donor site) and placing it onto the injured area (recipient site). Skin grafts can be used to cover wounds or areas with exposed bone.</p> <p>C. V-Y advancement flap: A V-Y advancement flap is a technique that involves creating a triangular-shaped flap of tissue and advancing it to cover a defect. It is often used for soft tissue defects that require local tissue rearrangement.</p> <p>D. Thenar flap: A thenar flap is a surgical procedure that involves transferring a flap of tissue from the thenar eminence (the fleshy part at the base of the thumb) to reconstruct a defect in the hand or thumb region.</p> <p>E. Microvascular free toe flap: A microvascular free toe flap is a complex procedure that involves transferring a flap of tissue, along with its blood supply, from the toe to the injured area. It is typically used for large defects that require a substantial amount of tissue for reconstruction.</p> <p>Without further details about the specific injury or the available evidence in the literature, it is difficult to determine the most appropriate treatment option. It is recommended to consult with a qualified healthcare professional who can assess the injury, consider the individual's medical history and other factors, and provide personalized treatment recommendations based on the specific case."</p>

and popular AI platform for both physician and patient medical knowledge.¹⁴

This study has several limitations. The majority of the 2004–2006 self-assessment exam questions were answered without elaboration, whereas the majority of ChatGPT's responses to the 2007–2013 questions included additional explanations. Given that questions were entered into ChatGPT one at a time in numerical order and chronologically by exam year, the reasons for this discrepancy may have included internet conditions (ie, traffic due to the number of users on ChatGPT at a given time, location where ChatGPT was used, and broadband access for high-speed internet) or the fact that ChatGPT's responses to questions are unique every time it is asked again. This may have

skewed the secondary outcomes, and there may have been more elaborations than measured in our study. In addition, there are limitations to the AI platform itself, such as the validity of the data it is trained on. ChatGPT uses data available on the internet to answer questions; therefore, ChatGPT's responses are subject to the biases of that data. Furthermore, because ChatGPT's answers to the same questions may vary each time, further research should be performed to determine the degree of intrareliability of answers by asking the AI platform the same questions at separate timepoints.

In sum, this study demonstrates that ChatGPT does not earn passing scores on the ASSH self-assessment examinations in hand surgery. Medical professionals, trainees, and patients should take

this into consideration when using ChatGPT and other language learning models.

References

1. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. *Pak J Med Sci.* 2023;39(2):605–607.
2. Sanmarchi F, Bucci A, Nuzzolese AG, et al. A step-by-step researcher's guide to the use of an AI-based transformer in epidemiology: an exploratory analysis of ChatGPT using the STROBE checklist for observational studies. *Z Gesundh Wiss.* 2023;1–36.
3. 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.
4. Bhattacharya K, Bhattacharya AS, Bhattacharya N, Yagnik VD, Garg P, Kumar S. ChatGPT in surgical practice—a new kid on the block. *Indian J Surg.* 2023.
5. 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):269–273.
6. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmol.* 2023;141(6):589–597.
7. Kalantar SH, Khabiri SS, Bagheri N. ChatGPT: A useful tool for teaching orthopedic trauma to residents. *Iran J EmergMed.* 2023;10(1):e9.
8. Van Heest AE, Armstrong AD, Bednar MS, et al. American Board of Orthopaedic Surgery's initiatives toward competency-based education. *JBJS Open Access.* 2022;7(2):e21.00150.
9. Adkinson JM, Chung KC. Lifelong learning for the hand surgeon. *J Hand Surg Am.* 2015;40(9):1915–1918.
10. Lum ZC, Collins D, Dennison S, et al. Can artificial intelligence pass the American Board of Orthopaedic Surgery? An analysis of 3900 questions. Accessed December 29, 2023. <https://openai.com/blog/dall-e-3-is-now-available-in-chatgpt-plus-and-enterprise>
11. OpenAI. DALL-E 3 is Now Available in ChatGPT Plus and Enterprise. Accessed October 27, 2023. <https://openai.com/blog/dall-e-3-is-now-available-in-chatgpt-plus-and-enterprise>
12. Ayesha Hareem, Iqbal Sajid, Tariq Mehreen, et al. Automatic medical image interpretation: State of the art and future directions. *Pattern Recognition.* 2021;114:107856.
13. American Society for Surgery of the Hand. Self-Assessment Examination. Accessed July 3, 2023. <https://www.assh.org/s/self-assessment-examination>
14. Lee K, Hoti K, Hughes JD, Emmerton LM. Consumer use of "Dr Google": a survey on health information-seeking behaviors and navigational needs. *J Med Internet Res.* 2015;17(12):e288.