Original Research

HAND
1–8

DOI: 10.1177/15589447251322914
journals.sagepub.com/home/HAN

Matching Human Expertise: ChatGPT's Performance on Hand Surgery Examinations

Zachary A. Kirschenbaum[1], Yuri Han[1], Kiera L. Vrindten[1], Hanbin Wang[1], Ron Cody[2], Brian M. Katt[1], and David Kirschenbaum[1]

Abstract

Background: The integration of artificial intelligence (AI) into health care witnessed significant advancements, particularly with AI-driven tools like ChatGPT. Initial evaluations indicated that ChatGPT 3.5 did not perform as well as humans on specialized hand surgery self-assessment examinations. The purpose of this study is to evaluate the performance of ChatGPT 4o on American Society for Surgery of the Hand (ASSH) self-assessment questions and whether using enhanced techniques such as better prompts and file search improve accuracy. **Methods:** Using data from the ASSH self-assessment examinations (2008-2013), we explored the impact of ChatGPT model version, prompt, and file search on the accuracy of AI-generated responses. We used OpenAI's application programming interface to automate question input and response scoring. Statistical analysis was conducted using one-way analysis of variance. KR-20 was used to assess the reliability of the test. **Results:** Results indicate that the latest AI models, particularly ChatGPT 4o with enhanced prompting and access to peer-reviewed literature, can achieve performance levels comparable to human examinees, particularly on text-based questions. ChatGPT 4o performed significantly better than ChatGPT 3.5 and showed marked improvement with better prompts and file search capabilities. The KR-20 for the 2013 examination was 0.946, indicating a very reliable test. **Conclusions:** These findings highlight AI's potential to support medical education and practice, demonstrating that ChatGPT can perform at a human-equivalent level on hand surgery self-assessment examinations. Our results suggest potential utility as a supplementary tool in educational settings and as a supportive resource in clinical practice.

Keywords: ChatGPT, AI, education, certification, self-assessment

Introduction

The modern field of artificial intelligence (AI) has seen significant advancements since its inception in the 1950s. Over the past decade, AI has transitioned from a theoretical concept to a practical tool integrated into various aspects of health care. For instance, AI is used to detect anomalies in medical imaging, predict patient outcomes, and personalize treatment plans. Hospitals and clinics increasingly rely on AI-driven systems to enhance patient care and streamline administrative processes.

In November 2022, AI surged in popularity with the release of ChatGPT. As of March 2024, about 1 in 5 American adults have used ChatGPT.[1] Given the excitement around AI, the role of AI in health care is an area of active interest.

Recent literature has evaluated ChatGPT compared with human examinees on standardized medical examinations. Recent examinations evaluated include the United States Medical Licensing Examination (USMLE), Orthopaedic Surgery In-Training Examination, the American Society for Surgery of the Hand (ASSH) and European Board Hand Surgery assessment examinations.[2-7] Although ChatGPT passed the USMLE, results have been questionable for the specialized orthopedic content like the hand surgery subspeciality. For example, Han et al[8] demonstrated that

[1]Rutgers Robert Wood Johnson Medical School, New Brunswick, NJ, USA
[2]SAS Institute, Cary, NC, USA

Corresponding Author:
Brian M. Katt, Department of Orthopaedic Surgery, Rutgers Robert Wood Johnson Medical School, 125 Paterson Street, New Brunswick, NJ 08901, USA.
Email: briankatt@gmail.com

Table 1. Summary of Experiment Setup.

Experiment No.	Experiment description	Years evaluated	Text questions evaluated	Image questions evaluated	Total questions evaluated
1	ChatGPT 3.5	2009-2013	659	0	659
2	ChatGPT 4o	2009-2013	659	286	945
3	ChatGPT 4o with better prompt	2009-2013	659	286	945
4	ChatGPT 4o with file search	2013 only	144	51	195

ChatGPT 3.5 would not have passed any of 10 years of ASSH self-assessment examinations.

This study evaluates the performance of ChatGPT on ASSH self-assessment examination questions over a 5-year period, using enhanced techniques such as improved prompting and file search capabilities. The null hypothesis is that the versions of ChatGPT with and without improved prompt writing and citations from peer-reviewed journals will perform equally. By rigorously evaluating ChatGPT's performance, we aim to explore not only its accuracy but also its potential implications for hand surgery education and clinical practice.

Methods

This study was approved by our institution's institutional review board. After signing a usage agreement with the ASSH, we requested self-assessment examination data for the years 2004 to 2013 because these were the years we evaluated in a previous study.[8] We could not obtain examinations prior to 2008 because the ASSH was unable to export the data in the format we required.

Scoring

We split the study into 4 experiments: ChatGPT 3.5, ChatGPT 4o, ChatGPT 4o with better prompts, and ChatGPT 4o with file search (Table 1). For each experiment, a question was scored as correct if ChatGPT's response matched the correct answer key provided by ASSH. Each experiment was run 10 times to account for the nondeterministic nature of ChatGPT, and the average accuracy was reported with standard deviations.

Experiment 1: ChatGPT 3.5

In the first experiment, we sent questions verbatim to ChatGPT 3.5 without any enhancements. This is how most humans interact with ChatGPT and how we evaluated our previous study.[8] Because ChatGPT 3.5 cannot read images, we only evaluate text questions. The following is an example of the input (the input is called the prompt):

The recommended surgical treatment of a thumb subungual malignant melanoma in a 62-year-old is:

A. Mohs resection
B. Marginal resection
C. Ray resection
D. Amputation at the metacarpal phalangeal joint
E. Amputation at the interphalangeal joint

Experiment 2: ChatGPT 4o

In May 2024, OpenAI released a stronger model named ChatGPT 4o.[9] For experiment 2, we simply upgrade the model from ChatGPT 3.5 to ChatGPT 4o. In addition, we begin evaluating image questions because ChatGPT 4o can read images. We did not evaluate video questions because no version of ChatGPT supports video.

Experiment 3: Better Prompts

It is well understood that writing better prompts can help improve ChatGPT output quality.[10] In our third experiment, we added general instructions, examples of desired behavior, and encouraged the model to think through the answer. For the purpose of this study, we will avoid technical jargon and simply call this "better prompt."

When adding examples, we used 2 text questions from each of the 7 categories (Ancillary, Basic Science, Bone and Joint, Miscellaneous, Neuromuscular, Skin, Vascular) on the 2008 examination. As we included 2008 answers in the prompt, we did not evaluate ChatGPT against 2008. The following is an example of the revised prompt:

General instructions

You are a board-certified hand surgeon. You are taking a multiple-choice exam to test your hand surgery knowledge. You will be presented with a question and a few possible answer choices. You may also be provided with images if the question references a figure. First, think through each of the options. Briefly discuss each option and decide on the best answer choice. Then, write the letter of the answer you have chosen.

Example of desired behavior

To help you understand what I'm looking for, here's a question from the 2008 exam.

A 23-year-old man sustained a jamming injury to his right long finger six months prior to seeking consultation . . . A prerequisite to rebalancing the extensor mechanism requires division of:

A. Transverse retinacular ligaments
B. Lateral bands
C. Sagittal bands
D. Volar plate
E. Central slip

A prerequisite to surgical correction is division of the transverse retinacular ligaments which are foreshortened in chronic boutonniere deformity . . . Central slip reconstruction is required to restore active PIP extension. Therefore, the correct answer is A.

Actual question

Now that I've shared the instructions and an example of the desired behavior, please answer the following question from the 2013 exam.

The recommended surgical treatment of a thumb subungual malignant melanoma in a 62-year-old is:

A. Mohs resection
B. Marginal resection
C. Ray resection
D. Amputation at the metacarpal phalangeal joint
E. Amputation at the interphalangeal joint

Experiment 4: File Search (2013 Only)

When physicians take self-assessment examinations, they might consult peer-reviewed journals. Until recently, ChatGPT could not do this because it had no way of searching user-provided files (even then, it would need to pass a paywall if material was not publicly available). In April 2024, OpenAI released file search,[11] which let ChatGPT 4o search up to 10 000 files before answering a question. This is known as Retrieval-Augmented Generation (often referred to as an "agent" approach in recent AI buzz), but for the purpose of this study, we will call it "file search" to make it easier to understand. We measured the impact of file search on self-examinations.

The ASSH provided a list of references for each question. We manually downloaded each reference, uploaded the files to OpenAI storage, and created an Assistant to read from the storage. Due to access issues, we only downloaded 441 of the 518 references. We only evaluated file search on the 2013 examination due to the time-consuming nature of manually downloading all references. Although providing the relevant references to ChatGPT might appear to "stack the deck," this mirrors the intended real-world use case where physicians seek targeted information to address specific problems. File search enables ChatGPT to identify and retrieve the most relevant material—such as a reference on subungual malignant melanoma—rather than unrelated content, which is precisely the advantage of retrieval-augmented systems in assisting with focused medical queries.

Automation

We wrote python code to automate question input to ChatGPT using the OpenAI application programming interface (API). Automation allowed us to improve the speed and consistency of our experiments. For example, when OpenAI released ChatGPT 4o in May 2024, we could simply rerun our program rather than having to manually enter hundreds of questions into the tool. Beyond that, automation provided stricter data privacy guarantees because OpenAI cannot retrain their models on data sent through the API. In September 2024, OpenAI released an even stronger model named o1-preview[12]; however, we did not include o1-preview in the study because it did not support images at the time of writing. With this said, ChatGPT 4o was the version used in this present study.

Our code is publicly available at https://github.com/zkbaum/handai.[13] However, the data are not publicly available because it is owned by the ASSH.

Statistics

Each row of data contained a question, answer letter, answer explanation, category, figure URLs (if image question), and the percent of human examinees who selected the correct answer. For 2013 only, we obtained a separate sheet of every human examinee's response to every question so that we could compute KR-20.

We computed *P*-values using one-way analysis of variance (ANOVA) in a python script. This measured whether the differences in each version of ChatGPT were statistically significant. However, ANOVA cannot be used to compare ChatGPT to human examinees. Therefore, we used a different python script to compute ChatGPT averages with 95% confidence intervals and compare it to human averages. Because of limitations in the ASSH testing system, we only obtained human averages per-question. Therefore, to obtain a per-examination average, we averaged the score of each question. The human results do not have confidence intervals because each year had a different number of human examinees. We assume the human confidence intervals are small given that each year had more than 1000 human examinees.

We assessed the assumptions of one-way ANOVA using Levine's test for homogeneity of variance, which did not reject the null hypothesis ($P > .05$), confirming equal variances across groups. Box plot visualizations further showed

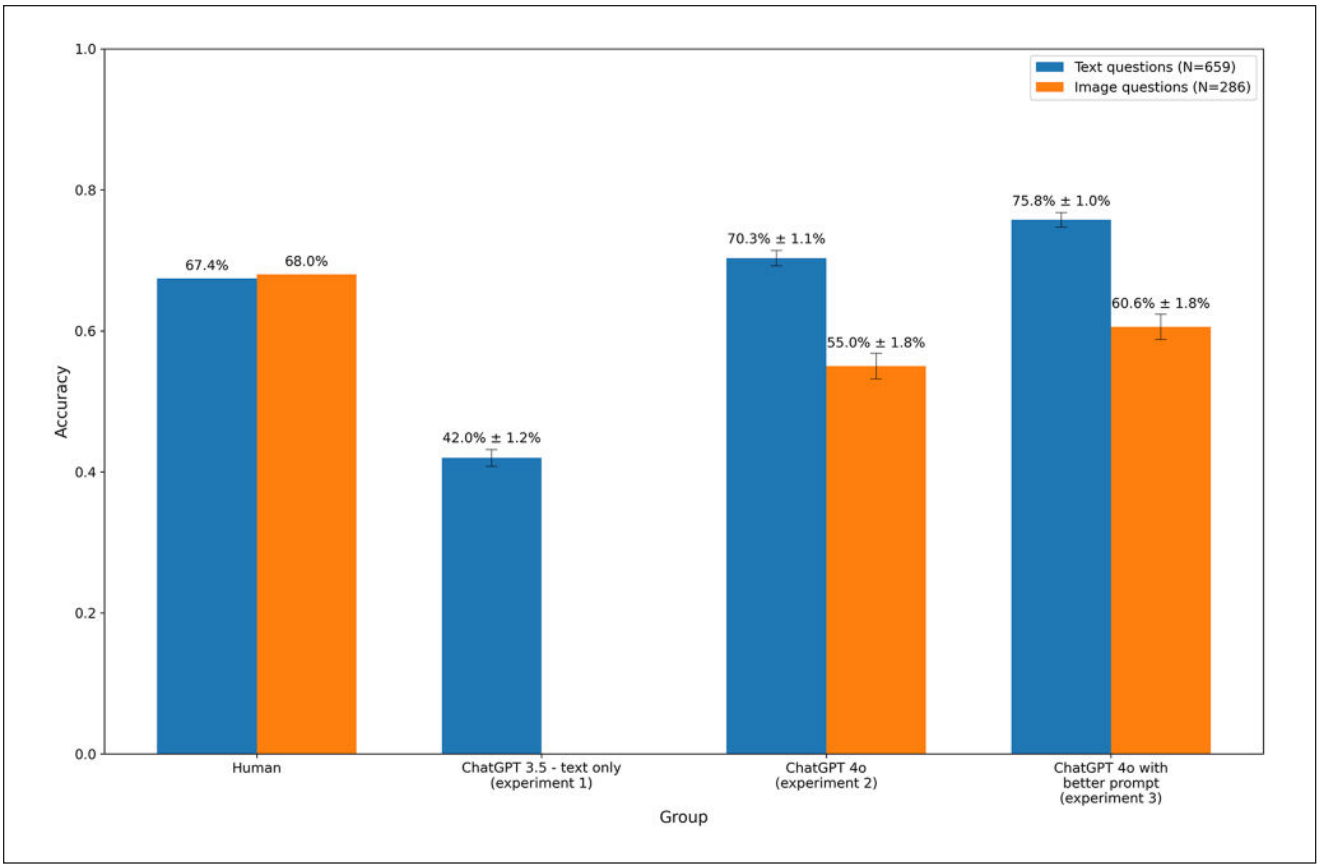

Figure 1. Performance of humans versus ChatGPT on self-assessment examinations (2009-2013), with and without better prompting.

no outliers. Given these findings, and the robustness of ANOVA to minor deviations from normality when variance is equal, we proceeded with ANOVA for group comparisons.

Separately, we used SAS to evaluate the KR-20, a measure of the test's reliability (also called Cronbach's alpha). Test reliability uses the human responses to measure if the test is "doing what it's supposed to do." For instance, if you tested again, would you get the same results?[14,15] This helped us understand if we were using high-quality data—our study would be less valuable if we evaluated on poor-quality tests. We analyzed test reliability on the 2013 examination only because of limitations in the human examinee data. Our analysis required pseudonymized access to each test-taker's response and this was only available post-2012. We also removed the data from the 31 of 1721 humans that left more than 5% of the test blank.

Results

With statistical significance, ChatGPT 4o performed better than ChatGPT 3.5, and ChatGPT 4o with better prompts performed better than ChatGPT 4o (Figure 1, Table 2).

Table 2. One-Way ANOVA Demonstrates That Upgrading the Model Version and Using Better Prompts Improves ChatGPT Performance With Statistical Significance.

Comparison	Question type	*P*-value
ChatGPT 3.5 versus ChatGPT 4o	Text	<.05
ChatGPT 3.5 versus ChatGPT 4o with better prompts	Text	<.05
ChatGPT 4o versus ChatGPT 4o with better prompts	Text	<.05
	Image	<.05

Note. ANOVA = analysis of variance.

ChatGPT 4o achieved a score of 70.3 ± 1.1% on text questions without prompting, whereas the score for human test-takers was 67.4% (Figure 1). With the addition of better prompts, ChatGPT 4o achieved a score of 75.8 ± 1.0% for text-only questions (Figure 1). Similarly, ChatGPT 4o performed better on image-based questions with the addition of better prompting (Table 2).

Furthermore, ChatGPT 4o performed better on the self-assessments with file search for text-only questions (Figure 2,

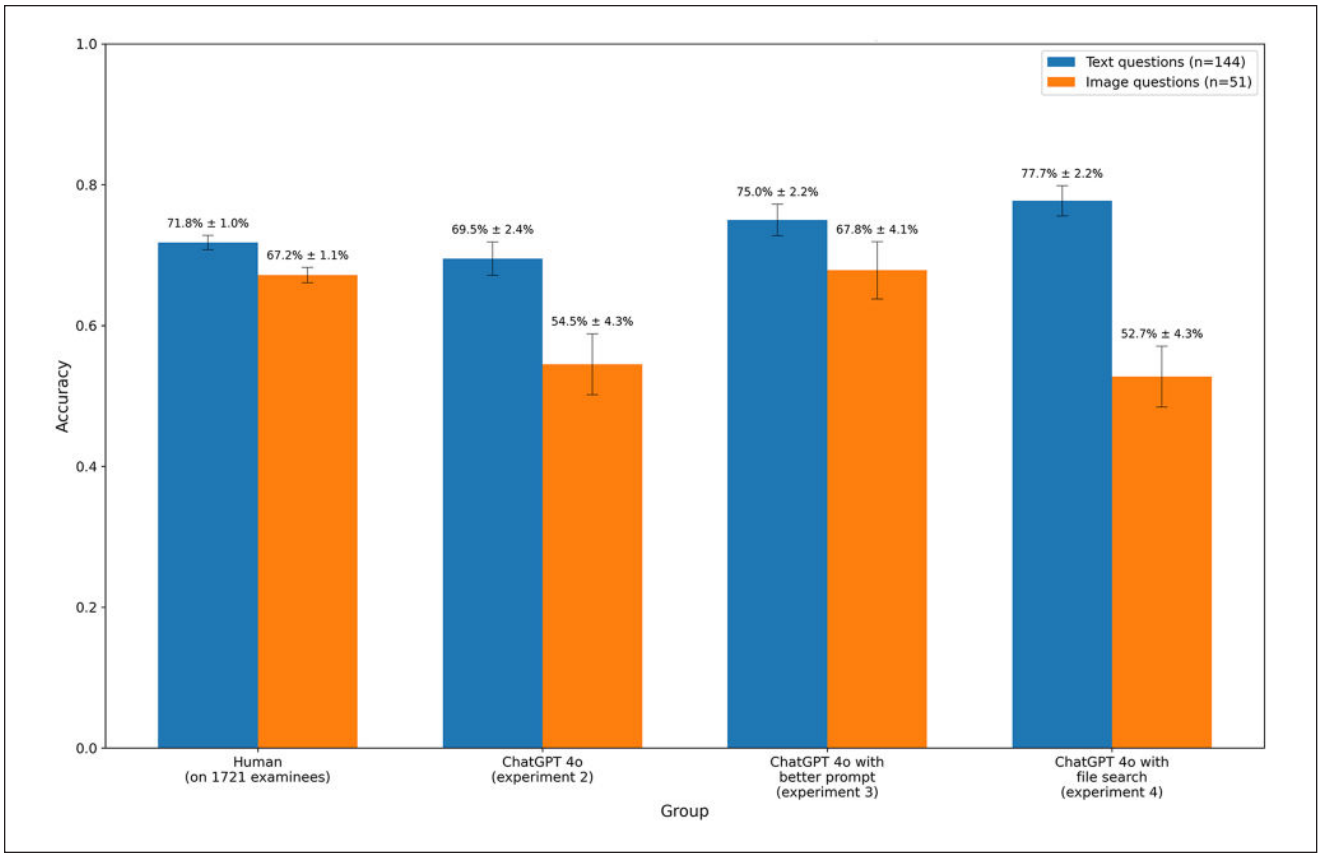

Figure 2. Performance of humans versus ChatGPT on self-assessment examination (2013 only), with and without file search.

Table 3). ChatGPT 4o scored 69.5 ± 2.4% without file search compared with scoring 77.7 ± 1.0% with file search (Figure 2). There was no statistical difference between ChatGPT 4o's performance on image-based questions regardless of file search involvement (Table 3). In addition, there was no statistically significant difference between the performance of ChatGPT 4o with better prompting versus that of ChatGPT 4o with file search for text-only questions (Table 3). However, better prompt yielded a higher score with statistical significance compared with file search for image-only questions (Table 3).

The KR-20 for the 2013 examination was 0.946 and indicated a very reliable test. One-way ANOVA revealed a significant main effect for group comparisons ($P < .05$), confirming differences in ChatGPT performance across conditions.

Table 3. One-Way ANOVA Demonstrates That File Search Improves ChatGPT Performance With Statistical Significance on Text Questions, but Not IMAGE Questions.

Comparison	Question type	*P*-value
ChatGPT 4o versus ChatGPT 4o with better prompts	Text	<.05
	Image	<.05
ChatGPT 4o versus ChatGPT 4o with file search	Text	<.05
	Image	.572
ChatGPT 4o with better prompts versus ChatGPT 4o with file search	Text	.087
	Image	<.05

Note. ANOVA = analysis of variance.

Discussion

ChatGPT Performance Across 5 Years of Examinations

ChatGPT 4o performed much better than ChatGPT 3.5. This is expected as ChatGPT 4o is the newer model.

ChatGPT 4o performed slightly better than humans on text questions. For example, ChatGPT scored 70.3 ± 1.1% on text questions without any sophisticated prompting, compared with humans scoring 67.4%. With better prompts, ChatGPT scored even higher at 75.8 ± 1.0%. This is a huge improvement from our previous study, where ChatGPT struggled to exceed 40%.[7,8]

Although ChatGPT performed well on text questions, it still lags on image questions. Even with better prompts,

ChatGPT still underperformed humans by about 8%. This discrepancy in performance between text and image questions is similar to those of previous work, suggesting that—despite the platform's developments—ChatGPT still harbors limitations within subspecialty fields such as hand surgery.[3]

Practical Strategies for Improving ChatGPT's Performance

It is not surprising that writing better prompts improves the response quality. Previous research by Sakai et al noted this with Japanese Ophthalmology Society examinations.[16,17] Although writing better prompts improves ChatGPT performance, it is not realistic to expect every hand surgeon to become a prompting expert. For this reason, a more practical recommendation at this time is for orthopedic hand physicians to simply transition from using ChatGPT 3.5 to the newer ChatGPT 4o.

Beyond that, file search could also be a reasonable alternative to prompting. In our 2013 experiments, ChatGPT 4o with file search (77.7 ± 2.2%) performed similarly to ChatGPT 4o with better prompts (75.0 ± 2.2%) on text-only questions. For example, suppose hand surgeons could link their ChatGPT accounts to their *HAND*, *Journal of Bone and Joint Surgery*, and *Journal of Hand Surgery* subscriptions so that ChatGPT could read peer-reviewed journals before generating an answer. Alternatively, suppose *HAND* licensed their data to OpenAI similar to what is done with The Atlantic and Reddit.[18,19] That said, file search on image questions still lags behind (52.8 ± 4.3% file search compared with 67.8 ± 4.1% for better prompt). Image file search needs to improve before we can consider it as a valid alternative, and journals need to provide a way for subscribers to link their ChatGPT accounts.

Data Selection and Quality

There are many ways to evaluate ChatGPT on medical tasks, such as through fact-checking, bias detection, and broad medical data sets like MedQA. However, we chose the ASSH self-assessment examinations because they are specific to hand surgery and offer the most objective and measurable assessment. The questions on the ASSH self-examinations were selected by a committee in an attempt to provide a comprehensive review of hand surgery. Most questions are not controversial, and the KR-20 findings confirm the self-assessment examination is very reliable and an excellent choice for comparing different test-takers. In addition, our previous study used the same ASSH self-examinations and were thus the most consistent choice in evaluating ChatGPT's performance on hand surgery questions.

When medical issues are straightforward with typically one good diagnosis and treatment (as is the case in most self-assessment examination questions), AI shows ability similar to or better than humans. The results may be much more divergent when treatment is more controversial and not always based on scientific data or clinical practice guidelines. Classic examples include wrist and shoulder fracture treatment in the elderly where, at times, personal preference overrides best available science.[20]

ChatGPT is trained on a vast amount of information from the internet, including Web sites, books, news articles, and more. Responses are based on a handful of sources which are referenced in the response. Thus, the quality of the response is based on the quality of the sources. Some of the most commonly cited studies, including those in peer-reviewed journals, are poorly designed, biased, and controversial. Corruption of medical science can therefore spread misinformation, but that is no different than a medical provider promoting ideology over science.

As a result of running these experiments, we were able to reject the null hypothesis. Using simple reproducible techniques such as better prompts or file search with ChatGPT 4o, we demonstrated that the AI can now perform at least equally as well as human test-takers on the self-assessment examinations. This finding highlights the potential utility of AI in addressing hand-related issues. Given ChatGPT's improved performance, it is important to consider how these findings translate into practical applications in educational and clinical settings.

Implications for Medical Education and Clinical Practice

Our study demonstrates that ChatGPT can perform at a level comparable to human test-takers on hand surgery self-assessment examinations. Although we did not directly evaluate its use as an educational tool, the ability of ChatGPT to arrive at correct answers as determined by qualified experts suggests potential utility in educational settings. For example, trainees might use ChatGPT to check their understanding of complex topics such as microsurgical techniques or the management of intricate conditions like scaphoid fractures. By posing questions and receiving immediate answers that align with expert consensus, trainees could reinforce their learning and identify areas needing further study. This interactive engagement could complement traditional learning resources in hand surgery education.

In clinical settings, practicing hand surgeons might use ChatGPT as a supportive resource for quick reference or to refresh knowledge on rare conditions. For instance, when encountering an uncommon congenital hand anomaly, a surgeon could consult ChatGPT for an overview of treatment options and recent advancements. By linking their journal subscriptions to ChatGPT—enabling the AI to search through paywalled articles when responding to

questions—surgeons could access up-to-date, specialized information efficiently. Although such integration is not currently available, this functionality was effectively simulated in our file search experiments, where we manually uploaded the PDFs beforehand. This approach mirrors how a surgeon might manually review literature but with increased efficiency, assisting in initial information gathering and allowing them to focus more on patient care and complex decision-making that require their expertise and judgment. ChatGPT might also assist physician extenders working in an Emergency Room or Urgent Care when caring for injuries they know little about like fight bites or cases where the diagnosis eludes them like a potential scaphoid fracture.

Our study has limitations. We focused on whether ChatGPT arrived at the correct answers on self-assessment examinations but did not assess the quality or depth of its explanatory responses. Although ChatGPT's ability to select correct answers aligns with expert consensus, we cannot fully attest to the educational value or reliability of its explanations. Therefore, although its performance is promising, further research is needed to evaluate the comprehensiveness of its responses, especially when used as an educational or clinical decision-making tool.

We acknowledge concerns that AI tools like ChatGPT might replace human roles in medicine. Our intention is not to suggest that ChatGPT can replace hand surgeons but to highlight that its demonstrated performance may assist them. By providing rapid access to information and supporting continuous learning, ChatGPT could augment the capabilities of both trainees and experienced surgeons. Importantly, we are not advocating for the use of ChatGPT to cheat on examinations; the self-assessment examinations were used in this study solely as a consistent method to evaluate ChatGPT's knowledge specific to hand surgery. Overall, integrating AI tools like ChatGPT into hand surgery education and practice may offer benefits when used responsibly and ethically.

Conclusion

ChatGPT shows significant promise in answering hand surgery examination questions. ChatGPT 4o achieves human-level performance on text questions and approaches human-level performance on image questions, and even better with better prompts and file search. The next hurdle is ensuring ChatGPT has access to high-quality literature, which will further enhance its applicability not only in hand surgery but in medical practice overall. These advancements highlight the potential of AI in the medical field, suggesting that ChatGPT could serve as a supplementary tool in hand surgery education and as a supportive resource in clinical practice. Responsible integration of AI like ChatGPT requires further research to evaluate the quality of its explanations and to address challenges such as performance on image-based questions. Embracing AI as an assistive tool may enhance learning and patient care when used ethically and responsibly.

Acknowledgments

The authors would like to express their sincere gratitude to Olivia Williams at the American Society for Surgery of the Hand (ASSH) for providing the data exports. They also extend thanks to Aidan O'Shea for contributions to file search, and to James Cohan for technical advising, which greatly enhanced the quality of the work.

Ethical Approval

This study was approved by our institutional review board.

Statement of Human and Animal Rights

This study did not involve human or animal participation.

Statement of Informed Consent

This study did not require human participation and informed consent.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Zachary A. Kirschenbaum https://orcid.org/0009-0009-0902-570X
Brian M. Katt https://orcid.org/0000-0002-6726-3580

References

1. McClain C. Americans' use of ChatGPT is ticking up, but few trust its election information. March 26, 2024. Accessed July 6, 2024. https://www.pewresearch.org/short-reads/2024/03/26/americans-use-of-chatgpt-is-ticking-up-but-few-trust-its-election-information/#who-has-used-chatgpt
2. Mihalache A, Huang RS, Popovic MM, et al. ChatGPT-4: an assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. *Med Teach*. 2024;46(3):366-372. doi:10.1080/0142159X.2023.2249588
3. Ghanem D, Nassar JE, El Bachour J, et al. ChatGPT earns American Board Certification in Hand Surgery. *Hand Surg Rehabil*. 2024;43(3):101688. doi:10.1016/j.hansur.2024.101688
4. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. *J Am Acad Orthop Surg*. 2023;31(23):1173-1179. doi:10.5435/JAAOS-D-23-00396

5. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep*. 2023;13(1):16492. doi:10.1038/s41598-023-43436-9
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(2):e0000198. doi:10.1371/journal.pdig.0000198
7. Arango SD, Flynn JC, Zeitlin J, et al. The performance of ChatGPT on the American Society for Surgery of the Hand Self-Assessment Examination. *Cureus*. 2024;16(4):e58950. doi:10.7759/cureus.58950
8. Han Y, Choudhry HS, Simon ME, et al. ChatGPT's performance on the hand surgery self-assessment exam: a critical analysis. *J Hand Surg Glob Online*. 2024;6(2):200-205. doi:10.1016/j.jhsg.2023.11.014
9. OpenAI. Hello GPT-4o. May 13, 2024. Accessed June 26, 2024. https://openai.com/index/hello-gpt-4o/
10. Prompt engineering—OpenAI API. Accessed June 26, 2024. https://platform.openai.com/docs/guides/prompt-engineering
11. Assistants Overview—OpenAI API. Accessed April 15, 2024. https://platform.openai.com/docs/assistants/overview
12. OpenAI. Introducing OpenAI o1-preview. September 12, 2024. Accessed September 24, 2024. https://openai.com/index/introducing-openai-o1-preview/
13. Kirschenbaum Z. HandAI; 2024. https://github.com/zkbaum/handai
14. Cody R, Smith JK. Test scoring and analysis using SAS. SAS Institute; 2014. Accessed July 6, 2024. https://support.sas.com/content/dam/SAS/support/en/books/test-scoring-and-analysis-using-sas/67044_excerpt.pdf
15. Nunnally JC. *Psychometric Theory*. McGraw-Hill; 1967.
16. Nori H, Lee YT, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv:231116452. 2023. doi: 10.48550/arXiv.2311.16452
17. Sakai D, Maeda T, Ozaki A, et al. Performance of ChatGPT in board examinations for specialists in the Japanese Ophthalmology Society. *Cureus*. 2023;15(12):e49903. doi:10.7759/cureus.49903
18. Bross A. The Atlantic announces product and content partnership with OpenAI. May 29, 2024. Accessed July 6, 2024. https://www.theatlantic.com/press-releases/archive/2024/05/atlantic-product-content-partnership-openai/678529/
19. Open AI, Reddit. OpenAI and Reddit partnership. May 16, 2024. Accessed July 6, 2024. https://openai.com/index/openai-and-reddit-partnership/
20. Aryee JNA, Frias GC, Haddad DK, et al. Understanding variations in the management of displaced distal radius fractures with satisfactory reduction. *Hand*. Published online March 8, 2024. doi:10.1177/15589447241233709