Evaluating Capabilities of Large Language Models: Performance of GPT4 on Surgical Knowledge Assessments

Brendin R Beaulieu-Jones, MD MBA[1,2]; Sahaj Shah, BS[3]; Margaret T Berrigan, MD[1]; Jayson S Marwaha, MD MBI[4]; Shuo-Lun Lai, MD[4]; Gabriel A Brat, MD, FACS, MPH[1-2]

[1]Department of Surgery, Beth Israel Deaconess Medical Center, Boston, MA
[2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA
[3]Geisinger Commonwealth School of Medicine, Scranton, PA
[4]Division of Colorectal Surgery, National Taiwan University Hospital, Taipei, Taiwan

Dr. Beaulieu-Jones is supported by National Library of Medicine/NIH grant [T15LM007092]

Disclosure Information: Nothing to disclose**.**

Corresponding Author:
Gabriel A Brat, MD, FACS, MPH
Department of Surgery, Beth Israel Deaconess Medical Center
Department of Biomedical Informatics, Harvard Medical School
110 Francis Street, Suite 2G, Boston, MA 02215

Keywords: ChatGPT; artificial intelligence; language models; surgical education; surgery

Short Title: Performance of ChatGPT-4 on Surgical Knowledge Assessments

Abstract

Background: Artificial intelligence (AI) has the potential to dramatically alter healthcare by enhancing how we diagnosis and treat disease. One promising AI model is ChatGPT, a large general-purpose language model trained by OpenAI. The chat interface has shown robust, human-level performance on several professional and academic benchmarks. We sought to probe its performance and stability over time on surgical case questions.

Methods: We evaluated the performance of ChatGPT-4 on two surgical knowledge assessments: the Surgical Council on Resident Education (SCORE) and a second commonly used knowledge assessment, referred to as Data-B. Questions were entered in two formats: open-ended and multiple choice. ChatGPT output were assessed for accuracy and insights by surgeon evaluators. We categorized reasons for model errors and the stability of performance on repeat encounters.

Results: A total of 167 SCORE and 112 Data-B questions were presented to the ChatGPT interface. ChatGPT correctly answered 71% and 68% of multiple-choice SCORE and Data-B questions, respectively. For both open-ended and multiple-choice questions, approximately two-thirds of ChatGPT responses contained non-obvious insights. Common reasons for inaccurate responses included: inaccurate information in a complex question (n=16, 36.4%); inaccurate information in fact-based question (n=11, 25.0%); and accurate information with circumstantial discrepancy (n=6, 13.6%). Upon repeat query, the answer selected by ChatGPT varied for 36.4% of inaccurate questions; the response accuracy changed for 6/16 questions.

Conclusion: Consistent with prior findings, we demonstrate robust near or above human-level performance of ChatGPT within the surgical domain. Unique to this study, we demonstrate a substantial inconsistency in ChatGPT responses with repeat query. This finding warrants future consideration and presents an opportunity to further train these models to provide safe and consistent responses. Without mental and/or conceptual models, it is unclear whether language models such as ChatGPT would be able to safely assist clinicians in providing care.

Background

Artificial intelligence (AI) models have the potential to dramatically alter healthcare by enhancing how we diagnosis and treat disease. These models could lead to increased efficiency, improved accuracy and personalized patient care. Successful healthcare-related applications have been widely reported.[1–10] Within surgery, machine learning approaches that include natural language processing, computer vision, and reinforcement learning have each shown promise to advance care.[1,11–14] Still, despite the promise of AI to revolutionize healthcare, its use within the field is markedly limited compared to other industries. The severe implications of errors and empathy concerns regarding the use of AI in healthcare have led to cautious adoption.[7,11,15–18]

One promising recent model for use in healthcare is ChatGPT, a publicly-available, large language model trained by OpenAI.[19] Released in November 2022, ChatGPT received unprecedented attention,[20] given its notable performance across a range of medical and non-medical domains.[21] ChatGPT has shown robust, human-level performance on several professional and academic benchmarks, including a simulated bar exam, the graduate record examination (GRE), numerous Advanced Placement (AP) examinations, and the Advanced Sommelier knowledge assessment.[19] With regard to medical knowledge, an earlier version of ChatGPT was shown to perform at or near the passing threshold of 60% accuracy on the United States Medical Licensing Exam (USMLE).[22,23] In addition, ChatGPT has demonstrated robust performance on knowledge assessments in family medicine,[24] neurosurgery,[25] hepatology,[26] and a combination of all major medical specialties.[27] Moreover, ChatGPT has shown promise as a clinical decision support tool in radiology,[28] pathology,[29] and orthodontics.[30] ChatGPT has also performed valuable clinical tasks,[31–34] such as writing patient clinic letters, composing inpatient discharge summaries, suggesting cancer screening, and conveying patient education.[35] Lastly, several studies have highlighted the potential impact of ChatGPT on medical education and research, with roles ranging from supporting nursing education to advancing data analysis and streamlining the writing of scientific publications.[32,36–38] The emergence of ChatGPT has reignited interest in exploring AI applications in healthcare; however, it has also provoked numerous concerns, regarding bias, reliability, privacy, and governance.[21,26,32,36–41]

In the current study, we evaluate ChatGPT-4's performance on two commonly used surgical knowledge assessments: The Surgical Council on Resident Education (SCORE) curriculum and a second case-based question bank for general surgery residents and practicing surgeons – which is referred to as Data-B and not identified due to copyright restrictions. SCORE is an educational resource and self-assessment used by many US residents throughout residency training.[42–45] Data-B is principally designed for graduating surgical residents and fully-trained surgeons in preparation for the American Board of Surgery (ABS) Qualifying Exam (QE). These assessments were selected as their content represents the knowledge expected of surgical residents and board-certified surgeons, respectively. As such, it was thought that Form-B, while based on the same content area as SCORE, should include more higher-order management or multi-step reasoning questions, and that we may observe differential ChatGPT performance. The performance of ChatGPT on each of these assessments may provide important insights regarding ChatGPT-4's capabilities at this point in time. Perhaps more importantly, in addition to assessing performance, this study investigates reasons for ChatGPT errors and assesses its performance on repeat queries. This latter objective represents a significant contribution to our current understanding of large language models – and a critical domain for research for safe and effective use of AI in healthcare.

Methods

Artificial Intelligence

ChatGPT (Open AI, San Francisco, CA) is a publicly-available, subscription-based AI chatbot that first launched in November 2022. It was initially derived from GPT-3 (Generative Pretrained Transformer) language models, which are pre-trained transformer models designed primarily to generate text via next word prediction. To improve performance for ChatGPT, initial GPT-3 models were further trained using a combination of supervised and reinforcement learning techniques.[50] In particular, ChatGPT was trained using Reinforcement Learning from Human Feedback (RLHF), in which a reward model is trained from human feedback. To create a reward model, a dataset of comparison data was created, which was comprised of two or more model responses ranked by quality by a human AI trainer. This data could then be used to fine-tune the model using Proximal Policy Optimization.[46]

ChatGPT-PLUS is the latest development from Open-AI and employs GPT-4, which is the fourth iteration of the GPT family of language models.[19] Details regarding the architecture and development of GPT-4 are not publicly available. It is generally accepted that GPT-4 was trained in a similar fashion as GPT-3 via RLHF. While specific technical details are unknown, Open AI states in their technical report that one of the main goals in developing GPT-4 was to improve the language model's ability to understand and generate natural text, particularly in complex and nuanced scenarios. The report highlights improved performance of GPT-4, relative to GPT-3.5.

For example, GPT-4 passed a simulated bar exam with a score in the 90th percentile, whereas GPT-3.5 achieved a score in the 10th percentile of exam takers. GPT-4 was officially released on March 13, 2023 and is currently available via the ChatGPT Plus paid subscription.

Input Sources

Input questions were derived from two commonly used surgical educational resources:

1. SCORE: The Surgical Council on Resident Education (SCORE) is a nonprofit organization established in 2004 by the principal organizations involved in surgical education in the United States, including the American Board of Surgery (ABS) and the Association for Surgical Education (ASE). SCORE maintains a curriculum for surgical trainees, which includes didactic educational content and more than 2400 multiple-choice questions for self-assessment. A total of 175 self-assessment questions were obtained from the SCORE question bank. Access to the SCORE question bank was obtained through the research staff's institutional access. SCORE was not part of the research team and did not participate in the study design and completion of research. Using existing functionality within SCORE, study questions were randomly selected from all topics, except systems-based practice; surgical professionalism and interpersonal communication education; ethical issues in clinical surgery; biostatistics and evaluation of evidence; and quality improvement. Fellowship-level questions were not excluded from study inclusion. Questions containing images were excluded from analysis. After exclusion, a total of 167 questions from SCORE were included in the study analysis.
2. Data-B: Data-B is an educational resource for practicing surgeons and senior surgical trainees, which includes case-based, multiple choice questions across a range of general

surgical domains, including endocrine, vascular, abdomen, alimentary tract, breast, head and neck, oncology, perioperative care, surgical critical care, and skin/soft issue. A total of 120 questions were randomly selected for inclusion in the study. Questions containing images were excluded from analysis. After exclusion, 119 questions were included.

Encoding

For input into ChatGPT, all selected questions were formatted two ways:

1. Open-ended (OE) prompting: Constructed by removing all answer choices and translating the existing question into an open-ended phrase. Examples include: “What is the best initial treatment for this patient?”; “For a patient with this diagnosis and risk factor, what is the most appropriate operative approach?”; or “What is the most appropriate initial diagnostic test to determine the cause of this patient’s symptoms?”
2. Multiple choice (MC) single answer without forced justification: Created by replicating the original SCORE or Data-B question verbatim. Examples include: “After appropriate cardiac workup, which of the following surgeries should be performed?”; “Which of the following laboratory values would most strongly suggest the presence of ischemic bowel in this patient?”; or “Which of the following options is the best next step in treatment?”

Open-ended prompts were deliberately varied to avoid systemic errors. For each entry, a new chat session was started in ChatGPT to avoid potential bias. To reiterate, all questions were inputted twice (once with open-ended prompting and once via the multiple choice format). Presenting questions to ChatGPT in two formats provided some insight regarding the capacity of a predictive language model to generate accurate, domain-specific responses without prompting.

Assessment

Outputs from ChatGPT-4 were assessed for Accuracy, Internal Concordance and Insight by two surgical residents using the criteria outlined by Kung *et al.* in their related work on the performance of ChatGPT on the USMLE exam.[23] Response accuracy (i.e., correctness) was assessed based on the provided solution and explanation by SCORE and Data-B, respectively. Internal response concordance refers to the internal validity and consistency of ChatGPT's output—specifically whether the explanation affirms the answer and negates remaining choices without contradiction. Insight refers to text that is non-definitional, non-obvious and/or valid.[23]

Each reviewer adjudicated 100 SCORE questions and 70 Data-B questions, with 30% and 28% overlap, respectively. For overlapping questions, residents were blinded to each other's assessment. Interrater agreement was evaluated by computing the Cohen kappa (κ) statistic for each question type (Supplemental Table 1).

For the combined set of 167 SCORE questions included in the study, the median performance for all human SCORE users was 65%, as reported in the SCORE dashboard. Reference data for Data-B is not available, preventing an exact comparison between ChatGPT and surgeon users.

In addition, we reviewed all inaccurate ChatGPT responses to multiple choice SCORE questions to determine and classify the reason for the incorrect output. The classification system for inaccurate responses was created by study personnel and designations were made by consensus. Reasons for inaccurate responses included: inaccurate information in complex question, inaccurate information in fact-based question; accurate information, circumstantial discrepancy; inability to differentiate relative importance of information; imprecise application of detailed information; and imprecise application of general information. A description of each error type, as well as representative examples are shown in **Table 1**.

To further assess the performance and reproducibility of GPT-4, all responses to SCORE questions (MC format) that were initially deemed inaccurate were re-queried. Second, ChatGPT responses were compared to the initial output to determine if the answer response changed and if it changed, whether the response was now accurate or if it remained inaccurate.

Results

Accuracy of ChatGPT Responses

A total of 167 SCORE and 112 Data-B questions were presented to ChatGPT. The accuracy of ChatGPT responses for OE and MC SCORE and Data-B questions is presented in **Figure 1**. ChatGPT correctly answered 71% and 68% of MC SCORE and Data-B questions, respectively. The proportion of accurate responses for OE questions was lower than for MC, particularly for SCORE questions, which is largely due to an increase in responses that were deemed indeterminate by study adjudicators in the setting of the open-ended format.

Internal Response Concordance of ChatGPT Responses

Internal Concordance was adjudicated by review of the entire ChatGPT response (**Table 2**). Overall internal response concordance was very high: 85.6% and 100% for OE SCORE and Data-B questions, respectively, and 88.6% and 97.3% for MC SCORE and Data-B questions.

Among OE SCORE questions, internal response concordance was also assessed by accuracy subgroup (**Figure 2**). Concordance was nearly 100% (79/80) for accurate responses. Internally discordant responses were more frequently observed for inaccurate responses (33%, 31/75).

Insights within ChatGPT Responses

For both OE and MC questions, approximately two-thirds of ChatGPT responses contained nonobvious insights (**Table 2**). Insights were more frequently observed for OE questions (SCORE: 66.5% versus 63.5%; Data-B: 77.7% versus 62.7%).

Classification of Inaccurate ChatGPT Responses to MC SCORE Questions

Reasons for inaccurate responses are shown (**Table 3).** The most common reasons were: inaccurate information in a complex question (36.4%); inaccurate information in fact-based question (25.0%); and accurate information, circumstantial discrepancy (13.6%**).**

Outcome of Repeat Question for Initially Inaccurate ChatGPT Responses

For all inaccurate ChatGPT responses to MC SCORE questions, the exact MC SCORE question was re-presented to the ChatGPT on a separate encounter, using a new chat. The accuracy of the response was assessed as prior. In total, the answer selected by ChatGPT varied between iterations for 16 questions (36.4% of inaccurate questions). The response remained inaccurate in 10/16 questions and was accurate on the second encounter for 6/16 questions. No change in the selected MC answer was observed in nearly two-thirds of cases (n=28, 63.6%).

Discussion

To assess ChatGPT's capabilities within the surgical domain, we assessed the performance of ChatGPT-4 on two surgical knowledge self-assessments. Consistent with prior findings in other domains, ChatGPT exhibited robust accuracy and internal concordance, near or above human-level performance. The study highlights the accuracy of ChatGPT within a highly specific and sophisticated field without specific training or fine-tuning in the domain. The findings also underscore some of the current limitations of AI including variable performance on the same task and unpredictable gaps in the model's capabilities. In addition, the non-tiered performance of ChatGPT on SCORE and Data-B suggests a distinctiveness between human knowledge and/or learning and the development of language models. Nonetheless, the robust performance of a language model within the surgical domain – and potential to enhance its performance domain-specific training (i.e., high-yield surgical literature) – highlights its potential value to support and advance human tasks in clinical decision-making and healthcare. While human context and high-level conceptual models are needed for certain decisions and tasks within surgery, understanding the performance large language models will direct their future development such that AI tools are complementarily positioned within healthcare, offloading extraneous cognitive demands.

Foremost, within the surgical domain, ChatGPT demonstrated near or above human-level performance, with an accuracy of 71% and 68% on MC SCORE and Data-B questions, respectively. This is consistent with ChatGPT performance in other general and specific knowledge domains, including law, verbal reasoning, mathematics, and medicine.[19,23] The current findings are consistent with a study by Hopkins *et al.*, in which ChatGPT was tested on and achieved near human-level performance on a subset of questions from the Congress of Neurological Surgeons (CNS) Self-Assessment Neurosurgery (SANS).[25]

The current study utilizes two knowledge assessments, which are generally accepted to be tiered in difficulty, with SCORE principally designed for residents and Data-B targeted for senior residents and board-certified attending general surgeons. This design provides additional insight into the performance of ChatGPT relative to humans; we would anticipate that ChatGPT would perform superiorly on SCORE relative to Data-B. However, the near equivocal relative performance of ChatGPT on SCORE and Data-B suggests that its capabilities do not parallel those of surgical trainees. Learners progressively attain greater layers of context and understanding to expand their knowledge. A predictive language processing model such as ChatGPT does not improve in a similar manner, given the nature of its corpus of information and reinforcement-based training. It is an informal observation that SCORE questions often require more precise delineation of similar answer choices (e.g., distal pancreatectomy with splenectomy versus distal pancreatectomy alone), and Data-B generally requires a broader knowledge set to answer each question. The near equivocal performance suggests that a probabilistic algorithm like ChatGPT can function at a high level in both tasks, but it also highlights that the mental and conceptual models that providers use to develop their expertise should not be attributed to these models. Mental models have acknowledged limitations, but they allow physicians to think broadly during clinical encounters where information is limited. Importantly, such differences may lead to errors by the language model that experienced providers would consider basic or unlikely given the way we learn. Thus, it is still too early to assume language models can safely assist clinicians in providing care. Future research into how large language models perform, with specific attention to end-points beyond accuracy, is needed to direct further development and application of language models and related AI in surgery and healthcare.

Two additional findings warrant consideration. First, our analysis highlighted the kind of errors that ChatGPT makes on surgical knowledge questions. In 11 of 44 inaccurate responses (25%), the incorrect response related to a straightforward, fact-based query (e.g., What is the second most common location of blunt thoracic vascular injury after the aorta?). Second, we observed inconsistencies in ChatGPT responses. When erroneous responses were re-presented to the language model interface, output varied in one-third of instances, and responses were different and incorrect (e.g. select another multiple choice response) in two-thirds. These two findings highlight a substantial limitation of current predictive language models when the response changes over several days. Future performance metrics should include a measure of consistency as well as initial capability. Fine-tuning to the specific domain may improve the confidence of the model and subsequent consistency; this type of finding underscores the importance of implementing AI tools in a complementary fashion in healthcare, given the high costs of errors.

To our knowledge, this is the first study testing the performance of ChatGPT on knowledge assessments over multiple instances. The extraordinary results of a general-purpose model like ChatGPT highlight both the incredible opportunity that exists and the value of additional domain-specific fine tuning and reinforcement learning. In particular, future research is needed to assess ChatGPT's performance within clinical encounters, rather than standardized knowledge assessments. Large language models such as ChatGPT lack a conceptual model, and this is fundamentally different from how humans diagnose and treat, and may be a major limitation of

ChatGPT's performance in clinical settings—as highlighted in a recent blog post by an emergency medicine physician who tested ChatGPT's diagnostic capacity for a subset of recent clinical encounters.[52] Without these mental or conceptual models, correct responses to deterministic questions, like the questions within SCORE and Data-B, do not necessarily imply that the model would be able to assist clinicians in providing care in its current form.

The current study has notable limitations. First, a relatively small bank of questions was used, which may not accurately reflect the broader surgical knowledge domain. Second, the assessment of accuracy and internal concordance for the open-ended responses may be biased, but we found significant inter-rater reliability to calm this concern. Third, and most importantly, it is possible that some of the questions and/or answers are available in some form online and may allow the model to draw on previous answers. While the content is easily accessible online, the specific questions are less likely to be available online as both SCORE and Data-B are not open-source assessments. Finally, a metric of human performance on Data-B is not readily available, though median performance is likely equivalent to SCORE, given the reported data on both the American Board of Surgery In-Training and Qualifying Examinations.

Conclusion

Consistent with prior findings, the current study demonstrates the robust performance of ChatGPT within the surgical domain. Unique to this study, we demonstrate inconsistency in ChatGPT responses and answer selections upon repeat query. This finding warrants future consideration and demonstrates an opportunity for further research to develop tools for safe and reliable implementation in healthcare. Without mental models, it is unclear whether language models such as ChatGPT would be able to safely assist clinicians in providing care.

References

1. Khalsa RK, Khashkhusha A, Zaidi S, Harky A, Bashir M. Artificial intelligence and cardiac surgery during COVID-19 era. *J Card Surg*. 2021;36(5):1729-1733. doi:10.1111/JOCS.15417

2. Mehta N, Pandit A, Shukla S. Transforming healthcare with big data analytics and artificial intelligence: A systematic mapping study. *J Biomed Inform*. 2019;100. doi:10.1016/J.JBI.2019.103311

3. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *J Am Med Inform Assoc JAMIA*. 2020;27(7):1173-1185. doi:10.1093/JAMIA/OCAA053

4. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc JAMIA*. 2018;25(10):1419-1428. doi:10.1093/JAMIA/OCY068

5. Luh JY, Thompson RF, Lin S. Clinical Documentation and Patient Care Using Artificial Intelligence in Radiation Oncology. *J Am Coll Radiol JACR*. 2019;16(9 Pt B):1343-1346. doi:10.1016/J.JACR.2019.05.044

6. Johnson SP, Wormer BA, Silvestrini R, Perdikis G, Drolet BC. Reducing Opioid Prescribing After Ambulatory Plastic Surgery With an Opioid-Restrictive Pain Protocol. *Ann Plast Surg*. 2020;84(6S Suppl 5):S431-S436. doi:10.1097/SAP.0000000000002272

7. Makhni EC, Makhni S, Ramkumar PN. Artificial Intelligence for the Orthopaedic Surgeon: An Overview of Potential Benefits, Limitations, and Clinical Applications. *J Am Acad Orthop Surg*. 2021;29(6):235-243. doi:10.5435/JAAOS-D-20-00846

8. Hammouda N, Neyra JA. Can Artificial Intelligence Assist in Delivering Continuous Renal Replacement Therapy? *Adv Chronic Kidney Dis*. 2022;29(5):439-449. doi:10.1053/J.ACKD.2022.08.001

9. McBee MP, Awan OA, Colucci AT, et al. Deep Learning in Radiology. *Acad Radiol*. 2018;25(11):1472-1480. doi:10.1016/J.ACRA.2018.02.018

10. Rashidi HH, Tran NK, Betts EV, Howell LP, Green R. Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods. *Acad Pathol*. 2019;6. doi:10.1177/2374289519873088

11. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial Intelligence in Surgery: Promises and Perils. *Ann Surg*. 2018;268(1):70-76. doi:10.1097/SLA.0000000000002693

12. Mumtaz H, Saqib M, Ansar F, et al. The future of Cardiothoracic surgery in Artificial intelligence. *Ann Med Surg 2012*. 2022;80. doi:10.1016/J.AMSU.2022.104251

13. Raffort J, Adam C, Carrier M, Lareyre F. Fundamentals in Artificial Intelligence for Vascular Surgeons. *Ann Vasc Surg*. 2020;65:254-260. doi:10.1016/J.AVSG.2019.11.037

14. Stumpo V, Staartjes VE, Regli L, Serra C. Machine Learning in Pituitary Surgery. *Acta Neurochir Suppl*. 2022;134:291-301. doi:10.1007/978-3-030-85292-4_33

15. Petch J, Di S, Nelson W. Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. *Can J Cardiol*. 2022;38(2):204-213. doi:10.1016/J.CJCA.2021.09.004

16. Jarrett D, Stride E, Vallis K, Gooding MJ. Applications and limitations of machine learning in radiation oncology. *Br J Radiol*. 2019;92(1100). doi:10.1259/BJR.20190001

17. Cheng JY, Abel JT, Balis UGJ, McClintock DS, Pantanowitz L. Challenges in the Development, Deployment, and Regulation of Artificial Intelligence in Anatomic Pathology. *Am J Pathol*. 2021;191(10):1684-1692. doi:10.1016/J.AJPATH.2020.10.018

18. Sarno L, Neola D, Carbone L, et al. Use of artificial intelligence in obstetrics: not quite ready for prime time. *Am J Obstet Gynecol MFM*. 2023;5(2). doi:10.1016/J.AJOGMF.2022.100792

19. OpenAI. GPT-4 Technical Report. Published online March 15, 2023.

20. Zhang C, Zhang C, Li C, Qiao Y. One Small Step for Generative AI, One Giant Leap for AGI: A Complete Survey on ChatGPT in AIGC Era. *ArXiv 1013140RG222478970883*. Published online April 4, 2023.

21. Quick uptake of ChatGPT, and more - this week's best science graphics. *Nature*. Published online February 28, 2023. doi:10.1038/D41586-023-00603-2

22. 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. doi:10.2196/45312

23. 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

24. Morreel S, Mathysen D, Verhoeven V. Aye, AI! ChatGPT passes multiple-choice family medicine exam. *Med Teach*. Published online 2023. doi:10.1080/0142159X.2023.2187684

25. Hopkins BS, Nguyen VN, Dallas J, et al. ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board-style questions. *J Neurosurg*. Published online March 1, 2023:1-8. doi:10.3171/2023.2.JNS23419

26. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. Published online March 22, 2023. doi:10.3350/CMH.2023.0089

27. 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. *Res Sq*. Published online 2023. doi:10.21203/RS.3.RS-2566942/V1

28. Ismail A, Ghorashi NS, Javan R. New Horizons: The Potential Role of OpenAI's ChatGPT in Clinical Radiology. *J Am Coll Radiol JACR*. Published online March 2023. doi:10.1016/J.JACR.2023.02.025

29. Sinha RK, Deb Roy A, Kumar N, Mondal H. Applicability of ChatGPT in Assisting to Solve Higher Order Problems in Pathology. *Cureus*. 2023;15(2). doi:10.7759/CUREUS.35237

30. Strunga M, Urban R, Surovková J, Thurzo A. Artificial Intelligence Systems Assisting in the Assessment of the Course and Retention of Orthodontic Treatment. *Healthc Basel Switz*. 2023;11(5). doi:10.3390/HEALTHCARE11050683

31. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. *Lancet Digit Health*. 2023;5(4). doi:10.1016/S2589-7500(23)00048-1

32. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthc Basel Switz*. 2023;11(6). doi:10.3390/HEALTHCARE11060887

33. Rao A, Pang M, Kim J, et al. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow. *MedRxiv Prepr Serv Health Sci*. Published online February 26, 2023. doi:10.1101/2023.02.21.23285886

34. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. *Radiology*. Published online April 4, 2023:230424. doi:10.1148/RADIOL.230424

35. Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectr*. 2023;7(2). doi:10.1093/JNCICS/PKAD010

36. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*. 2023;15(2). doi:10.7759/CUREUS.35179

37. 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). doi:10.1007/S10916-023-01925-4

38. Thomas SP. Grappling with the Implications of ChatGPT for Researchers, Clinicians, and Educators. *Issues Ment Health Nurs*. 2023;44(3):141-142. doi:10.1080/01612840.2023.2180982

39. Vaishya R, Misra A, Vaish A. ChatGPT: Is this version good for healthcare and research? *Diabetes Metab Syndr*. 2023;17(4):102744. doi:10.1016/J.DSX.2023.102744

40. Dahmen J, Kayaalp ME, Ollivier M, et al. Artificial intelligence bot ChatGPT in medical research: the potential game changer as a double-edged sword. *Knee Surg Sports Traumatol Arthrosc Off J ESSKA*. 2023;31(4). doi:10.1007/S00167-023-07355-6

41. Will ChatGPT transform healthcare? *Nat Med*. 2023;29(3). doi:10.1038/S41591-023-02289-5

42. American Board of Surgery. SCORE - Surgical Council on Resident Education. *https://www.absurgery.org/default.jsp?scre_booklet*.

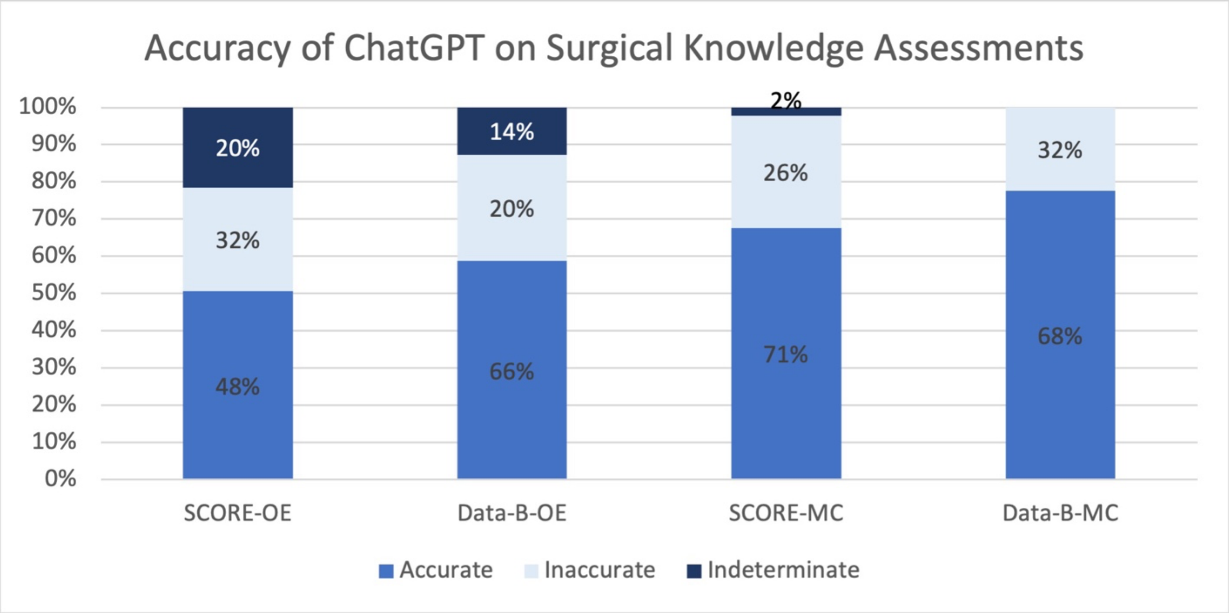
43. Bell RH. Surgical council on resident education: a new organization devoted to graduate surgical education. *J Am Coll Surg*. 2007;204(3):341-346. doi:10.1016/J.JAMCOLLSURG.2007.01.002

44. Klingensmith ME, Malangoni MA. SCORE provides residents with web-based curriculum for developing key competencies. *Bull Am Coll Surg*. 2013;98(10):10-15.

45. Moalem J, Edhayan E, Darosa DA, et al. Incorporating the SCORE curriculum and web site into your residency. *J Surg Educ*. 2011;68(4):294-297. doi:10.1016/J.JSURG.2011.02.010

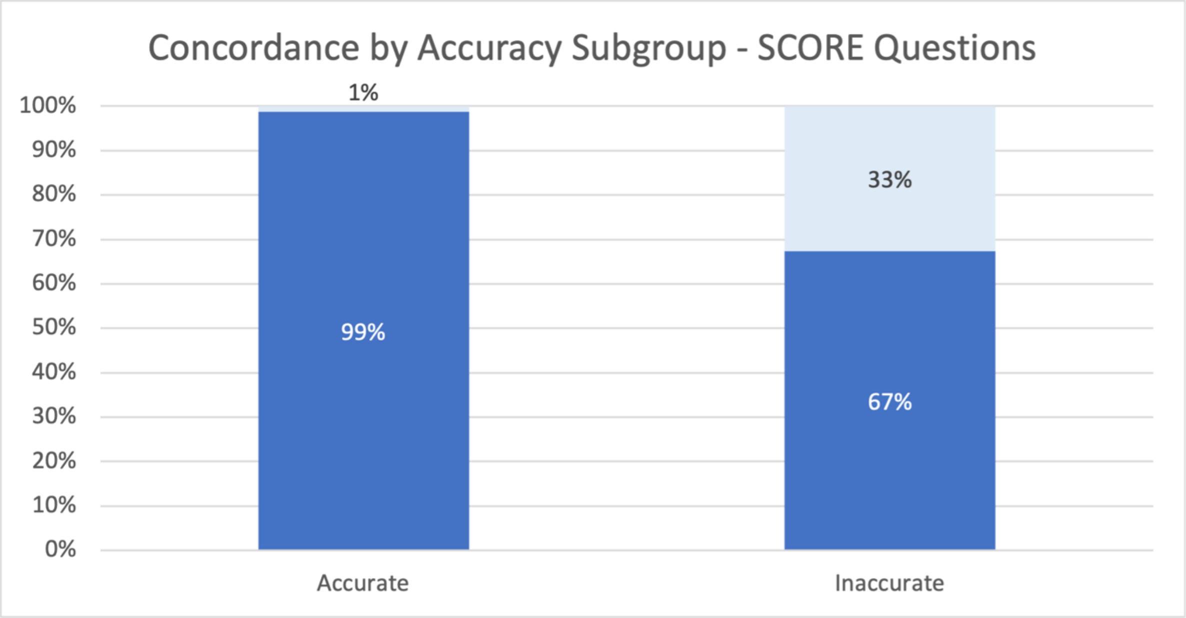
46. Gao L, Schulman J, Hilton J. Scaling Laws for Reward Model Overoptimization. Published online October 19, 2022.

Figure 1: Accuracy of ChatGPT Output for Open-Ended and Multiple-Choice Questions



Legend: Surgical knowledge questions from SCORE and Data-B were presented to ChatGPT via two formats: open-ended (OE; left sided and multiple-choice (MC; right side). ChatGPT's outputs were assessed for accuracy by surgeon evaluators. A total of 167 SCORE and 112 Data-B questions were presented to the ChatGPT interface. ChatGPT correctly answered 71% and 68% of multiple choice SCORE and Data-B questions, respectively.

Figure 2: Internal Concordance by Accuracy Subgroup among SCORE Questions



Legend: SCORE questions were presented to ChatGPT via two formats: open-ended and multiple-choice. ChatGPT's outputs to open-ended SCORE questions were assessed for internal concordance by accuracy subgroup. A total of 167 SCORE questions were presented to the ChatGPT interface. Concordance was nearly 100% (79/80) for accurate responses. Internally discordant responses were more frequently observed for inaccurate responses (33%, 31/75).

Table 1: Classification of Error Type: Description and Examples

Error Type	**Description and Theoretical Example(s) of Error**
Imprecise application of detailed information	Description: Answer selection was based on detailed clinical information, which was applied imprecisely or inaccurately to the clinical context Example: • Recommend medical management rather than surgery as first-line treatment for specific diagnosis, which is accurate, unless symptoms are medically-refractory, which is the case in the question
Imprecise application of general knowledge	Description: Answer selection was based on general knowledge, which was either incompletely accurate or out of scope given context of the question Example: • Recommend against a secondary procedure in a child to avoid additional anesthesia and potential procedural complications
Inability to differentiate relative importance of information	Description: Answer selection was based on accurate information, but did not delineate between more accurate options Example: • Select a laboratory finding which is present in most patients with a specific condition, when a more characteristic finding was intended
Accurate information; circumstantial discrepancy	Description: Response is based on accurate information, which is incorrect based on question interpretation or other circumstantial factors that unlikely reflect competency of GPT Example: • Select the cost-effective, first-line imaging, rather than the gold standard mechanism for diagnosis
Inaccurate information in fact-based question	Description: Response is based on inaccurate information in the context of a single-part, fact-based question Example: • Incorrectly identify the second most common site of pathology
Inaccurate information in complex question	Description: Response is based on inaccurate information in the context of a complex clinical scenario or multi-part question Example: • Inaccurate selection of most appropriate next step in patient with constellation of symptoms and description of imaging

Table 2: Accuracy, Internal Concordance and Nonobvious Insights of ChatGPT Responses

	Accuracy, N (%)	Internal Concordance	Insights
Open-Ended Format			
SCORE	80 (47.9%)	143 (85.6%)	111 (66.5%)
Data-B	74 (66.1%)	112 (100%)	87 (77.7%)
Combined	154 (55.2%)	255 (91.4%)	198 (71.0%)
Multiple Choice - Single Answer			
SCORE	119 (71.3%)	148 (88.6%)	106 (63.5%)
Data-B	76 (67.9%)	109 (97.3%)	69 (61.6%)
Combined	195 (69.9%)	257 (92.1%)	175 (62.7%)

SCORE: Surgical Council on Resident Education;
Data-B: refers to a second commonly used surgical knowledge assessment and question bank

Table 3: Classification of Inaccurate ChatGPT Responses for SCORE Questions (N=44)

Classification	**N (%)**
Imprecise application of detailed information	3 (6.8)
Imprecise application of general knowledge	4 (9.1)
Inability to differentiate relative importance of information	4 (9.1)
Accurate information; circumstantial discrepancy	6 (13.6)
Inaccurate information in fact-based question	11 (25.0)
Inaccurate information in complex question	16 (36.4)
Total	**44 (100%)**

Table 4: Outcome of Repeat Question for 44 Initially Inaccurate Responses to SCORE

Outcome	**N (%)**
No change in answer/response	28 (63.6%)
Change in answer/response	
Inaccurate to inaccurate	10 (22.7%)
Inaccurate to accurate	6 (13.6%)
Total	**44 (100%)**

Supplemental Material

Supplemental Table 1: Interrater Agreement – Cohen kappa for OE and MC questions

	Open-Ended Questions		Multiple Choice – Single Answer	
	Cohen K	N	Cohen K	N
SCORE	0.720	30	1.0	30
Data-B	0.681	20	1.0	20

SCORE: Surgical Council on Resident Education
Data-B: refers to a second commonly used surgical knowledge assessment and question bank

Additional Supplemental Material

All input to the ChatGPT interface and associated output were recorded. Due to copyright laws, this data is not presented in the current manuscript. However, pending requisite approval from the respective organizations, this data may be shared upon reasonable request.