Title: ChatGPT-4 as a Board-Certified Surgeon: A Pilot Study

Authors and Affiliations: Joshua Roshal MD[1,2], Caitlin Silvestri MD[3], Tejas Sathe MD[3,4], Courtney Townsend MD MAMSE FACS[1], V. Suzanne Klimberg MD PhD MSHCT MAMSE FACS[1], Alexander Perez MD MSHCT FACS[1]

[1]University of Texas Medical Branch, Galveston, TX
[2]Brigham and Women's Hospital/Harvard Medical School, Boston, MA
[3]New York Presbyterian/Columbia University Irving Medical Center, Department of Surgery, New York, NY
[4]University of California San Francisco, San Francisco, CA

Corresponding Author Information:
Joshua Roshal, MD
Mailing Address: 301 University Blvd, Galveston, TX 77551
Email: jaroshal@utmb.edu

Authors' ORCID IDs:
JR 0009-0004-5117-1878
CS 0009-0004-2124-0473
TS 0000-0003-0449-4469
CT N/A
VSK 0000-0003-2167-0698
AP 0000-0003-1416-5730

Abstract: (*224 words*)

Purpose:

Large language models (LLMs), such as GPT-4 (OpenAI; San Francisco, CA), are promising tools for surgical education. However, skepticism about their accuracy and reliability remains a significant barrier to their widespread adoption. Although GPT-4 has demonstrated a remarkable ability to pass multiple-choice tests, its general surgery knowledge and clinical judgment in complex oral-based examinations are less clear. This study aims to evaluate GPT-4's general surgery knowledge using written and oral board-style examinations to drive improvements that will enable the tool to revolutionize surgical education and practice.

Methods:

We tested GPT-4's ability to answer 250 random multiple-choice questions (MCQs) from the Surgical Council on Resident Education (SCORE) question bank and navigate four oral board scenarios derived from the Entrustable Professional Activities (EPA) topic list. Two former oral board examiners assessed the responses independently for accuracy.

Results:

On MCQs, GPT-4 answered 197 out of 250 (78.8%) correctly, corresponding to a 99% probability of passing the American Board of Surgery Qualifying Examination (ABS QE). On oral board scenarios, GPT-4 committed critical failures in three of four (75%) clinical cases. Common reasons for failure were incorrect timing of intervention and incorrect suggested operation.

Conclusions:

While GPT-4's high performance on MCQs mirrored prior studies, the model struggled to generate accurate long-form content in our mock oral board examination. Future efforts should use specialized datasets and advanced reinforcement learning to enhance GPT-4's contextual understanding and clinical judgment.

Keywords:
artificial intelligence (AI), oral boards, surgical education, board certification, simulation, ChatGPT

Acknowledgments:
None

Disclosures:
Joshua Roshal is a consultant for McGraw Hill, an American publishing company whose mission is the education of current and future healthcare professionals.

Introduction:

Large language models (LLMs) such as GPT-4 (OpenAI; San Francisco, CA) represent a subset of generative artificial intelligence (AI) tools that can create human-like content from text-based prompts. Various education technology companies already harness LLMs to facilitate cost-effective content creation. For example, Duolingo (NASDAQ: DUOL), the most popular education app worldwide, incorporates language learning exercises generated by GPT-4 [1, 2]. However, concerns about accuracy and reliability have limited the adoption of LLMs in health professions education, given the detrimental effects of misinformation to patient safety [3]. To address these concerns, several groups have sought to validate GPT-4's knowledge in specialized domains. In the absence of validated scoring rubrics, early researchers have relied on multiple-choice questions (MCQ) because of their simplicity and efficiency in evaluating the accuracy of LLM outputs. In these studies, GPT-4 achieved passing scores on all three parts of the United States Medical Licensing Exam (USMLE) as well as MCQ banks designed for neurosurgery written and oral board preparation [4, 6].

However, GPT-4's specific knowledge of General Surgery remains unknown. In the United States (US), general surgeons demonstrate their mastery of surgical principles through two examinations. The American Board of Surgery Qualifying Examination (ABS QE) is an eight-hour, multiple-choice test designed to evaluate knowledge of general surgical principles and applied science. Since MCQs only capture a narrow aspect of the knowledge and skills required for surgical practice, candidates also complete an oral examination. The ABS Certifying Examination (CE) is a dynamic and interactive testing experience in which the examinee is orally assessed on their ability to sequentially evaluate, plan, treat, and manage general surgical problems [7, 8]. While one prior study has highlighted the feasibility of using GPT-4 to simulate oral board examinations in anesthesiology, the model's accuracy in answering MCQs and free-response questions in general surgery contexts has not been assessed [9].

This study aims to evaluate GPT-4's general surgery knowledge using written and oral examinations modeled after the ABS QE and CE. By identifying GPT-4's knowledge gaps, these findings serve to inform the development of generative AI tools that will enable surgeon-educators to prioritize quality review over content creation in the instruction and assessment of trainees.

Methods:

Chatbot Creation and Prompt Generation

To simulate the mock written and oral examinations in this study, we built a custom generative AI chatbot based on the GPT-4 LLM. We modified a Vercel (Vercel, San Francisco, CA) template to provide our chatbot with a standard user interface and Supabase (Supabase Pvt. Ltd, Singapore) to store completed chats. Supplementary domain-specific clinical knowledge was not provided, and internet browsing capabilities were disabled, which ensured that generated outputs were based on the pretraining data of the GPT-4 model (last update: September 2021). Our code is available here: https://github.com/JRoshal/chatgpt-facs.

For the oral exam, in an effort to mimic the characteristic dialogue of oral general surgery examinations and shape the quality and relevance of responses to approximate the expertise of a surgeon eligible for board certification, we created a 594-word initial prompt and included the *Essential Attributes of a Certifiable Surgeon*, which are used to inform the assessment of surgical residents on oral examinations by ABS examiners (Figure 1) [8].

ChatGPT, you are about to engage in a Mock American Board of Surgery Certifying Examination, simulating the dynamics of a real-world oral examination with human examiners. You are going to be assessed on difficult oral board surgical scenarios. You are the primary surgeon responsible for managing the patient's surgical needs and complications. You will first prioritize obtaining a specific history and physical examination before considering labs and imaging. No consults from other surgical specialties are allowed; use your own expertise. The scenarios will be complex and involve patients with comorbidities or other not straightforward features. You will not be prompted for the next steps or given any hints or help unless you ask for it. Each brief clinical vignette will begin with the patient's age, sex, chief complaint, and care setting, and you will be asked about initial management. I will not provide vitals unless you specifically ask for them. I will only provide the specific history findings that you ask for. I will not provide a generalized history. I will only provide the specific exam findings that you ask for. I will not provide a generalized exam. I will only provide lab results for tests you specifically ask for. I do not provide all labs. I only provide the imaging study results you specifically ask for. I will not provide results for imaging you don't ask for. When requesting vitals or labs, use a markdown table. You will be asked to describe the preoperative workup if an operation is required. Then, you will be asked to detail the steps of the operation. You will be asked about post-operative care. You will be thrown for a loop by asking how you would manage a possible intra-op or post-op operation complication.

1. Scenarios:
    a. You will be presented with four clinical cases during each 30-minute session, with approximately 7 minutes dedicated to each case. Be mindful of the time to ensure each case is addressed adequately.
2. Response Expectations:
    a. Engage in Conversation: Treat this as a dialogue with the examiners. Listen (or "read") carefully to their queries and respond promptly and decisively in a step-by-step manner.
    b. Clinical Actions: Describe the steps you'd take with clarity and confidence while maintaining a conversational tone.
    c. Rationale: Provide concise yet clear reasoning behind your actions and decisions. Explain not just the "what" but also the "why."
    d. Communication: Be clear, concise, and decisive. Tailor your answers directly to the questions asked without overloading them with unnecessary details. Be prepared to expand if prompted.
3. Evaluation Criteria:
    a. Examiners are evaluating your ability to handle real-world situations in a conversational manner as you would in your own practice. Be authentic in your responses, focusing on patient-centered care, safety, and optimal outcomes. It's crucial to back up your decisions with logical and evidence-based reasoning.

Remember, the goal is to provide the 'right' answer and demonstrate sound clinical judgment, clear decision-making, and a thorough understanding of the subject matter.

You will be assessed according to the following essential attributes of a certifiable surgeon:

- Demonstrates an organized approach and solid rationale for planned actions.
- Rapidly determines and interprets key findings in a clinical presentation.
- Effectively and efficiently uses clinical knowledge to solve clinical problems; effectively addresses key management points.
- Avoids errors and critical fails (omission and commission) associated with the case.
- Recognizes personal limitations in knowledge and expertise when diagnosing and treating clinical problems.
- Reacts promptly but flexibly to alterations in the patient's course, e.g., disease or treatment complications.
- Overall, demonstrates appropriate surgical judgment, clinical reasoning skills, and problem-solving ability.

Figure 1: Initial Prompt for Mock Oral Examination

Examination Administration and Data Analysis

To simulate the ABS QE, we selected 250 MCQs at random from the Surgical Council on Resident Education (SCORE) online question bank based on the distribution percentage of categories as published by the ABS. We individually prompted the chatbot with each question once. We calculated the total number and percentage of questions the chatbot answered correctly. From the percent scores, we derived percentile rankings using a formula from a prior study that examined the influence of SCORE-based simulated ABSITE exam performance on true ABSITE performance [10]. Finally, we referenced a published relationship between ABSITE scores and ABS QE pass rates to determine the chatbot's likelihood of passing the ABS QE [11].

For the oral exam, access to validated scenarios and scoring rubrics is restricted by the ABS to maintain testing integrity. Additionally, there was a notable lack of standardized or validated frameworks for assessing the accuracy of long-form LLM outputs within general surgery contexts. As such, the authors recruited two former general surgery oral board examiners and Members of the Academy of Master Surgeon Educators (MAMSE). The internationally recognized MAMSE designation is bestowed upon individuals who have made significant contributions to surgical education in their lifetimes after a rigorous peer-review process by the American College of Surgeons (ACS) [12].

Dr. Courtney M. Townsend, Jr., MD, FRCSEd(hon), MAMSE, FACS is a distinguished Professor of Surgery at the University of Texas Medical Branch (UTMB) and holds the Robertson-Poth Distinguished Chair in General Surgery. With a career spanning nearly five decades in general surgery and surgical oncology, Dr. Townsend has held various leadership roles, including the John Woods Harris Distinguished Chairman of Surgery at UTMB, President of the ACS, Chair of the ABS, and Editor-in-Chief of the Sabiston Textbook of Surgery. He has been actively involved in numerous committees and editorial boards, received many accolades for his contributions to surgery and medicine, and has a prolific research and editorial record with numerous publications, book chapters, and grant-funded research projects.

Dr. V. Suzanne Klimberg, MD, PhD, MSHCT, MAMSE, FACS is a renowned Professor and Division Chief of Surgical Oncology and Colorectal Surgery at UTMB, specializing in breast surgical oncology with over three decades of experience. She holds the Courtney M. Townsend, Jr., MD Distinguished Chair in General Surgery and leads as the Clinical Director of the UTMB Cancer Center and Medical Director of the Breast Cancer Service Line. Dr. Klimberg is a pioneer in breast cancer treatment and has contributed extensively to the field with over 260 publications, 49 book chapters, several patents, and editorship of 13 surgical textbooks. Dr. Klimberg also holds prestigious positions in various medical organizations and editorial boards, and her work has been widely recognized and honored for its impact and excellence in surgical oncology.

The MAMSEs evaluated ChatGPT's clinical decision-making skills on four mock clinical case scenarios with topics that were randomly selected from the ABS's Entrustable Professional Activities (EPA) topic list: acute abdomen, gallbladder disease, right lower quadrant pain and appendicitis, and benign and malignant breast disease. The EPA project, announced by the ABS in 2022, redefined the learning objectives of contemporary graduate surgical education and aims to steer the certification standards of general surgery residents in the United States toward a competency-based assessment paradigm [13]. Competencies are defined as abilities that blend knowledge, skills, values, and attitudes, and the EPAs are an assessment framework and tool used to evaluate competency [13]. A resident is deemed ready for independent practice when faculty members consistently assess them as capable of safely executing EPA-specific learning objectives without supervision.

The MAMSE-chatbot dialogue was facilitated by an external microphone that captured MAMSE voices and speech-to-text transcription software. Each MAMSE led two clinical cases each, starting with an improvised history

of a fictional patient presenting with a common chief complaint related to the EPA-in-question, which prompted the chatbot to respond. The conversations continued until both MAMSEs independently determined whether the chatbot's responses demonstrated readiness for independent practice or not.

Results:

Written Examination:

Our chatbot answered 197 out of 250 (78.8%) SCORE MCQs correctly. This percent score corresponds to the 90th percentile of fifth-year US general surgery residents (Figure 2) and approximates the 99% probability of passing the ABS QE.

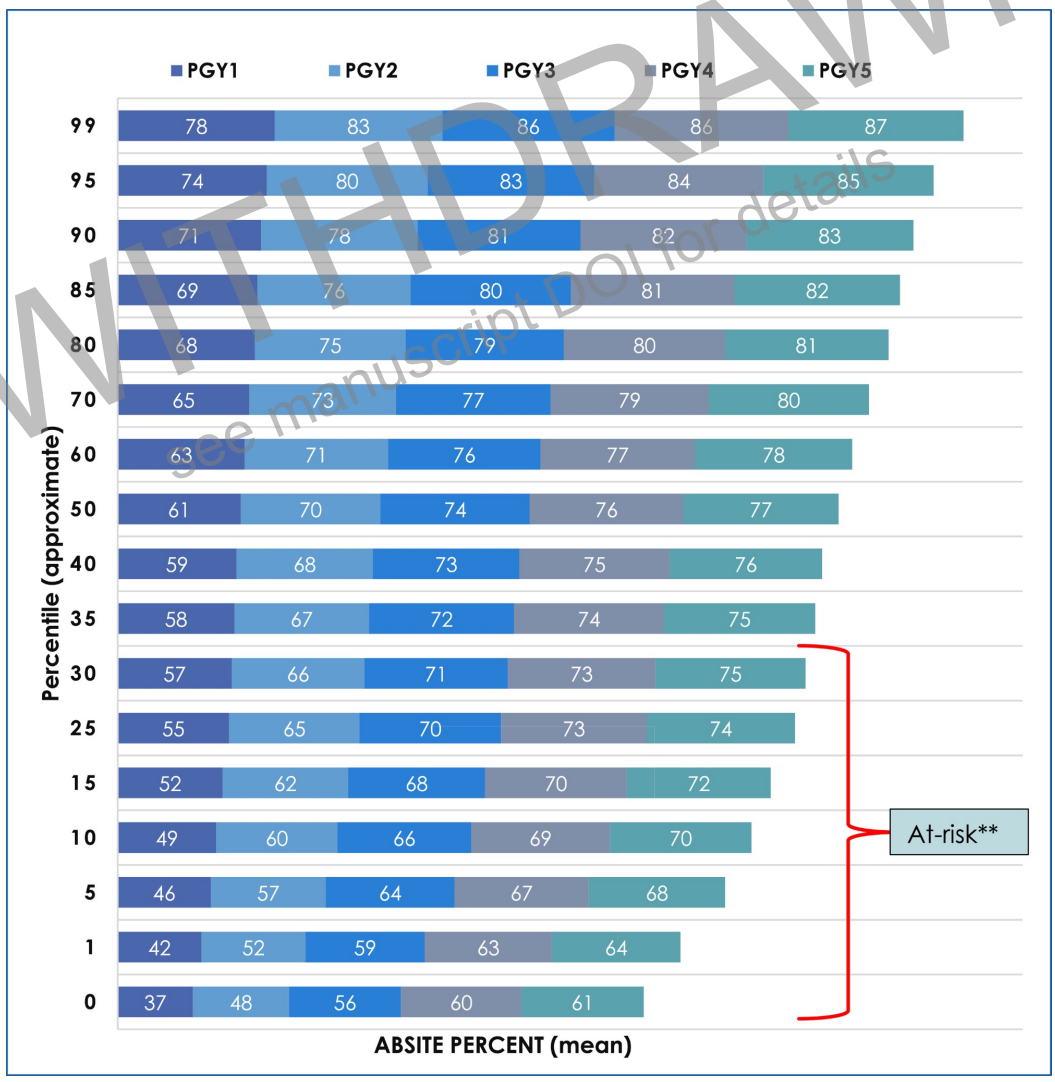

Figure 2: ABSITE Percentile Corresponding to the Average ABSITE Percent for each PGY level* (Shebrain et al., 2021).

*Abstracted from the ABSITE Score Graphs & Percentiles Reports (2014-2020). Numbers in the bars represent the mean value of ABSITE percent scores for that percentile during this period (e.g., if a PGY-1 resident answered 42% of questions correctly, this corresponds to the 1st percentile).

Oral Examination:

Our chatbot demonstrated readiness for independent practice in only one of four (25%) scenarios. Common reasons for failure were incorrect choice of operation or timing of operation suggested. Detailed information on each scenario is provided here:

1. **(Acute Abdomen):** *A 25-year-old waitress experiences an acute onset of the worst abdominal pain she's ever had while walking to her car after finishing her shift at 3:00 PM. She faints. A friend finds her, picks her up, and takes her to the emergency room. Upon arrival, her blood pressure is 90/60 mmHg, pulse 120 BPM, and respiratory rate 30 breaths/min. She is complaining of terrible abdominal pain.*

The MAMSEs presented the chatbot with a case of a 25-year-old woman with acute-onset severe lower abdominal pain, fainting, hypotension, and tachycardia. The chatbot appropriately recognized the patient's symptoms as concerning for an acute intra-abdominal pathology requiring immediate attention. The chatbot requested a history and physical examination, which were focused and relevant to the differential diagnosis, including life-threatening conditions like ectopic pregnancy and ovarian torsion. The chatbot prioritized laboratory testing like beta-hCG and appropriate imaging studies based on the patient's pregnancy status. The proposed management plan was tailored to the potential diagnoses, emphasizing the need for immediate surgical intervention for possible ectopic pregnancy or ovarian torsion. Overall, the chatbot's approach demonstrated a clear understanding of the urgency of the situation and the necessary steps for evaluation and treatment. As such, both MAMSEs independently determined that the chatbot demonstrated readiness for independent practice in the management of this patient with an acute abdomen.

2. **(Gallbladder Disease):** *A 60-year-old male presents to the emergency room with right upper quadrant pain, fever, and appears jaundiced on physical examination.*

The MAMSEs presented the chatbot with a case of a 60-year-old man with right upper quadrant pain, fever, and jaundice, a presentation concerning for acute cholangitis and concomitant pancreatitis. The chatbot failed to request a serum lipase level, which is a crucial diagnostic test in the evaluation of suspected pancreatitis. Furthermore, the chatbot did not contemplate the need for an emergent procedure, such as an endoscopic retrograde cholangiopancreatography with possible sphincterotomy or surgical common bile duct exploration, which are critical intervention options for patients presenting with acute cholangitis. While the chatbot correctly identified the need for a cholecystectomy, the recommended timing was incorrect. The chatbot suggested waiting 2-6 weeks after discharge, but at the time of writing the standard of care is to perform it during the same hospital admission once the episode of pancreatitis has resolved [14]. These oversights demonstrate a lack of understanding of the optimal timing and key components of management for a patient presenting with acute cholangitis. As such, both MAMSEs independently determined that the chatbot did not demonstrate readiness for independent practice in the management of this patient with gallbladder disease.

3. **(Benign and Malignant Breast Disease):** *A 30-year-old woman presents to your clinic with a breast lump.*

The MAMSEs presented the chatbot with a case of a 30-year-old woman with a painful breast lump and skin changes, a presentation concerning for a breast abscess. The chatbot initially gathered information about the lump's duration, associated symptoms, personal and family cancer history, menstrual and obstetric history, and any relevant medical conditions or medications. It then ordered a bilateral mammography and breast ultrasound as initial imaging studies. However, when the ultrasound showed a mixed hypoechoic mass, the chatbot jumped to recommend a core needle biopsy. It overlooked less invasive office-based interventions like an ultrasound-guided needle aspiration that could have immediately differentiated an abscess from a solid mass. Even when the fictional radiologist later aspirated pus, increasing the suspicion for an abscess, the chatbot failed to fully adapt its plan, remaining focused on cancer rather than abscess management. It struggled to synthesize information, refine its differential diagnosis, and tailor its

approach based on new findings. Despite generating some relevant points, the chatbot's limitations in integrating information and adaptive reasoning undermine its reliability for clinical decision-making. As such, both MAMSEs independently determined that the chatbot did not demonstrate readiness for independent practice in the management of this patient with benign breast disease.

4. **(Right Lower Quadrant Pain and Appendicitis):** *A 55-year-old woman presents to the emergency room with a 48-hour history of abdominal pain. The pain initially began in the mid-abdomen and then localized to the right lower quadrant.*

The MAMSEs presented the chatbot with a case of a 55-year-old woman with a 48-hour history of worsening abdominal pain that began in the mid-abdomen and localized to the right lower quadrant. Although the patient's vital signs were normal, her physical examination revealed exquisite generalized abdominal tenderness with right-sided rigidity. The chatbot appropriately gathered information about her symptoms (e.g., nausea, fever) and surgical history, which included an appendectomy. It then ordered laboratory studies (leukocytosis count of 15,000/μL) and cross-sectional imaging, which revealed a complex mass involving the right colon, suggesting a potential diagnosis of complicated right-sided diverticulitis. Despite these findings, the chatbot failed to recommend expedited surgical intervention. It instead suggested antibiotics and possibly a biopsy, and inappropriately pursued a colonoscopy. The chatbot generated relevant considerations but overlooked the patient's critical physical examination findings, thereby delaying the recognition of an urgent surgical condition. As such, both MAMSEs independently determined that the chatbot did not demonstrate readiness for independent practice in the management of this patient with right lower quadrant abdominal pain.

Discussion:

Despite the enticing potential of generative AI in surgical education, rigorous testing is still required to validate its clinical accuracy and mitigate safety risks arising from incorrect or misleading information. In this study, we demonstrated the GPT-4 LLM, without additional knowledge retrieval or fine-tuning, achieves high performance in multiple-choice exams but lower performance in oral examinations. GPT-4's predicted percent correct on the ABSITE, which is derived from its percent correct on our SCORE-based exam, surpasses 90% of scores achieved by fifth-year general surgery residents in the US and correlates to a 99% probability of passing the ABS QE [10, 11]. This performance suggests that GPT-4 possesses a level of knowledge in general surgical principles and applied science that meets the benchmarks established by the ABS. However, it is important to interpret these results with caution because of the inherent limitations of using MCQs as a measure of LLMs' capabilities. One study comparing GPT-4's performance on multiple-choice questions (MCQs) across various knowledge domains suggested that the LLM may select answers based on their physical position within the input prompt [15]. This phenomenon aligns with prior research demonstrating that shifting relevant information within the input prompt can decrease LLM question-answering performance. Notably, LLMs appear to perform most poorly when required to use information embedded within lengthy input prompts [16]. Given these challenges, GPT-4's predicted performance on our SCORE-based exam in this study may not accurately represent its general surgery knowledge base.

In the mock oral examination component of our study, GPT-4 demonstrated readiness for independent practice in a mere one out of four clinical scenarios (25%). The MAMSEs identified two themes of errors in GPT-4's responses: inaccuracies and omissions. For instance, the chatbot recommended pursuing a biopsy for a patient presenting with signs and symptoms of a breast abscess and a colonoscopy for a patient presenting with signs and symptoms of acute complicated diverticulitis, which are both deviations from standards of care [17, 18]. On the other hand, GPT-4 failed to consider the necessity of an emergent endoscopic retrograde cholangiopancreatography (ERCP) and possible sphincterotomy or surgical common bile duct exploration to adjudicate the common bile duct for a patient presenting with acute cholangitis. An interesting observation emerged from the acute abdomen scenario—the only case GPT-4 managed to navigate successfully—where a shorter exchange with the MAMSEs appeared to facilitate

better information integration by the chatbot, suggesting a potential limitation in handling long-form dialogues. Despite the limited number of cases in this study, this finding reiterates the importance of the physical arrangement of relevant information within prompts on LLM performance. More recent work has highlighted GPT-4's ability to accurately mimic the cognitive processes of physicians, which offers promising implications for the integration of LLMs into clinical workflows [19]. However, another study of ChatGPT-generated management recommendations for 362 breast cancer cases highlighted the inconsistency of the model in generating uniform long-form content in response to identical inputs and its difficulty in managing more advanced stages of cancer [20]. Therefore, future research should explore the parameters and characteristics of prompts that optimally enable LLMs to engage in clinical reasoning within long-form content generation. This will be crucial for the safe and effective incorporation of LLMs into surgical education and practice.

The intersection of surgical education and generative AI indeed represents a pivotal opportunity in the evolution of healthcare training and delivery. Surgical knowledge has traditionally been transmitted via the "see one, do one, teach one" method, in which learners are expected to conceptualize knowledge through repetition and pattern recognition on live patients. In response to public concerns about patient safety, simulation education has emerged as a tool to better prepare trainees for patient encounters and high-stakes procedures in surgery [21]. However, traditional simulation methods, including computer-based virtual environments, high-fidelity mannequins, task trainers, and standardized patients, can be expensive [22, 23]. For instance, one study found that it can cost upwards of $4.2 million to implement the American College of Surgeons (ACS)/Association of Program Directors in Surgery (APDS)-based surgical skills curriculum [24]. In another study of pharmacy students, using standardized patients to cultivate interpersonal communication skills can cost about $100 per student [25, 26]. These expenses often overlook the valuable time and expertise that surgeon-educators contribute to developing and leading learning sessions, which is estimated to be $30,000 in relative teaching value units per participating faculty member per year [24]. Unfortunately, the sustainability of the national instructor workforce is also at risk as programs continue to struggle to recruit faculty members to meet education demands and must instead rely on a small, committed group of instructors to shoulder the increasing burden of training the next generation of surgeons. One study from the University of Minnesota Department of Surgery found that between 2006 and 2014, the number of hours that department faculty had dedicated annually to resident and medical student simulation events surged from 81 to 365 [27]. These temporal and financial barriers can preclude the democratization of simulation-based surgical education in resource-poor institutions and regions [28, 29].

However, generative AI presents a promising avenue for more effective and cost-efficient tools in simulation education in various knowledge domains [30]. In surgery, AI systems are already being used to generate performative feedback from live operative videos and have informed the design of tailored educational interventions to strengthen interprofessional teamwork in the operating room [31–36]. With the ability to generate "unique" text-based clinical scenarios, LLMs may offer a cost-effective alternative to cultivating select competencies in graduate surgical education, such as diagnostic reasoning and communication skills [37]. Generative AI chatbots, such as ChatGPT (OpenAI; San Francisco, CA), are actively being explored as tools to simulate mock oral examinations in preparation for board certification in general surgery, given that in-person sessions replicating real-life testing conditions are infrequent, geographically limited, and can cost $35 per resident examinee [38, 39]. Another proof-of-concept study demonstrated the utility of using ChatGPT to teach emergency physicians how to break bad news [40]. In the future, these technologies may offer an opportunity to democratize access to quality surgical education in low- and middle-income countries such as Ukraine, where the scarce surgeon workforce is desperate for an improved surgical education infrastructure in response to recent influxes of patients with traumatic injuries [41–43].

This study has several key limitations. First, the MCQ portion of the exam did not involve prompting GPT-4 with multiple iterations of the same question, which limits our assessment of the model's consistency and reliability. For the oral examination component, the sample size of clinical cases was small, and we did not employ a pre-validated rubric to assess the LLM's responses, instead relying solely on the judgment of MAMSEs, who are recognized for their teaching excellence through the ACS's peer-review process. The clinical cases were also formulated on the spot

by the MAMSEs, rather than being pre-designed and validated, which may not have optimally challenged GPT-4's clinical decision-making abilities. Furthermore, a traditional oral examination typically consists of three consecutive 30-minute sessions, each featuring four case-based scenarios; our study condensed this to a single session of four cases, which may have influenced the model's performance, particularly in terms of handling prolonged, complex dialogues. In addition, while our 594-word initial prompt aimed to simulate the dialogue of a real oral examination, its length and structure may have inadvertently affected the model's ability to generate accurate long-form content [16]. Finally, this study only examined GPT-4, and the findings may not generalize to other LLMs. These limitations underscore the need for future research with larger sample sizes, validated evaluation rubrics, and diverse, pre-validated clinical scenarios to more comprehensively assess the potential and pitfalls of AI in surgical education.

Looking ahead, several techniques to increase the likelihood of extracting desired LLM outputs have been described. In addition to optimized prompt engineering, experts advocate for training LLMs on extensive private datasets meticulously annotated by human experts, a process known as fine-tuning [44]. Of note, this technique is only possible with particular LLMs, not including GPT-4. The creation of such a health dataset would demand significant resources and may blur the lines between private innovation and public welfare by jeopardizing the confidentiality of individually identifiable health information and perpetuating systemic biases arising from seemingly objective data [45]. Retrieval-augmented generation (RAG), which instead integrates external data sources (e.g., textbooks, website data) to anchor LLM responses, maybe a more effective and cost-efficient approach for enhancing accuracy within specialized domains [46]. In other words, fine-tuning a model is akin to enrolling in a college course; the model is trained on a specific dataset to enhance its performance in a particular area. RAG resembles taking an open-book test; the model accesses external information sources to generate more accurate responses. Combining vector embeddings with RAG is like also having a tutor during that test; the model is directed to the exact information needed, which improves retrieval speed and accuracy.

As generative AI technologies continue to improve, validation studies and ethical oversight are essential. In fact, some research groups have started to develop frameworks for the manual evaluation of generative AI outputs in health professions education [38]. However, methods that rely on human expert review tend to be expensive and struggle to keep pace with rapid advancements in AI technologies. Fortunately, automated evaluation metrics such as the bilingual evaluation understudy (BLEU) algorithm, which measures the accuracy of machine-translated text against human translations, offer powerful and efficient potential solutions for assessing LLM outputs in medical contexts [47]. Therefore, the development of safe, effective, generative AI-powered tools and scalable, cost-efficient evaluation systems in surgery will require the collaborative expertise and commitment of engineers, surgeons, and educators.

Conclusion:

The integration of generative AI into surgical education heralds a new era of training and practice. On one hand, generative AI may significantly reduce the cost of developing educational content, and thus make it more accessible. At the same time, the limited clinical reasoning abilities of some untrained models necessitate a cautious approach to their use in surgical education and care delivery. As this field evolves, ongoing research, ethical consideration, and collaboration between educators, engineers, and clinicians will be essential to harness AI's full potential while safeguarding the integrity of surgical education and patient care. Future research should focus on validating models and enhancing contextual understanding and clinical judgment through targeted prompt engineering and the integrating specialized datasets.

References:

1. Transcribing MF. Duolingo (DUOL) Q1 2024 Earnings Call Transcript. The Motley Fool. Published May 8, 2024. Accessed May 22, 2024. https://www.fool.com/earnings/call-transcripts/2024/05/08/duolingo-duol-q1-2024-earnings-call-transcript/
2. Duolingo lays off staff as language learning app shifts toward AI | CNN Business. https://www.cnn.com/2024/01/09/tech/duolingo-layoffs-due-to-ai/index.html
3. Sallam M. The Utility of ChatGPT as an Example of Large Language Models in Healthcare Education, Research and Practice: Systematic Review on the Future Perspectives and Potential Limitations. Published online February 21, 2023:2023.02.19.23286155. doi:10.1101/2023.02.19.23286155
4. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations. Neurosurgery. 2023;93(6):1353-1365. doi:10.1227/neu.0000000000002632
5. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. Neurosurgery. 2023;93(5):1090. doi:10.1227/neu.0000000000002551
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. General Surgery Examinations. American Board of Surgery. Accessed May 22, 2024. https://www.absurgery.org/get-certified/general-surgery/exams/
8. General Surgery Certifying Examination. American Board of Surgery. Accessed May 22, 2024. https://www.absurgery.org/get-certified/general-surgery/certifying-exam/
9. Cureus | Pioneering the Integration of Artificial Intelligence in Medical Oral Board Examinations | Article. Accessed May 22, 2024. https://www.cureus.com/articles/220972-pioneering-the-integration-of-artificial-intelligence-in-medical-oral-board-examinations#!/
10. Shebrain S, Folkert K, Baxter J, Leinwand M, Munene G, Sawyer R. SCORE-Based Simulated ABSITE Exam Performance as a Predictor of Performance on the ABSITE. J Surg Educ. 2021;78(5):1692-1701. doi:10.1016/j.jsurg.2021.03.011
11. ABSITE-QE. Accessed May 22, 2024. https://absurgerydata.shinyapps.io/relationqeabsite/
12. About the Academy | ACS. Accessed May 22, 2024. https://www.facs.org/for-medical-professionals/education/programs/academy-of-master-surgeon-educators/about-the-academy/
13. Montgomery KB, Mellinger JD, Lindeman B. Entrustable Professional Activities in Surgery: A Review. JAMA Surg. 2024;159(5):571-577. doi:10.1001/jamasurg.2023.8107
14. Management of acute pancreatitis - UpToDate. Accessed May 23, 2024. https://www.uptodate.com/contents/management-of-acute-pancreatitis
15. [2403.17752] Can multiple-choice questions really be useful in detecting the abilities of LLMs? Accessed May 23, 2024. https://arxiv.org/abs/2403.17752
16. Liu NF, Lin K, Hewitt J, et al. Lost in the Middle: How Language Models Use Long Contexts. Published online November 20, 2023. doi:10.48550/arXiv.2307.03172
17. Acute colonic diverticulitis: Surgical management - UpToDate. Accessed May 24, 2024. https://www.uptodate.com/contents/acute-colonic-diverticulitis-surgical-management
18. Primary breast abscess - UpToDate. Accessed May 24, 2024. https://www.uptodate.com/contents/primary-breast-abscess
19. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. Npj Digit Med. 2024;7(1):1-7. doi:10.1038/s41746-024-01010-1
20. Liao N, Li C, Gradishar WJ, et al. Accuracy and Reproducibility of ChatGPT Responses to 362 Breast Cancer Tumor Board Cases. Published online February 15, 2024. doi:10.2139/ssrn.4724374
21. Scott DJ. Patient safety, competency, and the future of surgical simulation. Simul Healthc J Soc Simul Healthc. 2006;1(3):164-170. doi:10.1097/01.sih.0000244453.20671.f2
22. Ker J, Hogg G, Maran N. Cost-effective simulation. In: Cost Effectiveness in Medical Education. CRC Press; 2011.
23. Zendejas B, Wang AT, Brydges R, Hamstra SJ, Cook DA. Cost: the missing outcome in simulation-based medical education research: a systematic review. Surgery. 2013;153(2):160-176. doi:10.1016/j.surg.2012.06.025
24. Danzer E, Dumon K, Kolb G, et al. What is the cost associated with the implementation and maintenance of an ACS/APDS-based surgical skills curriculum? J Surg Educ. 2011;68(6):519-525. doi:10.1016/j.jsurg.2011.06.004
25. Gillette C, Stanton RB, Rockich-Winston N, Rudolph M, Anderson HG. Cost-Effectiveness of Using

Standardized Patients to Assess Student-Pharmacist Communication Skills. Am J Pharm Educ. 2017;81(10):6120. doi:10.5688/ajpe6120

26. Willson MN, McKeirnan KC, Yabusaki A, Buchman CR. Comparing trained student peers versus paid actors as standardized patients for simulated patient prescription counseling. Explor Res Clin Soc Pharm. 2021;4:100081. doi:10.1016/j.rcsop.2021.100081
27. Acton RD, Chipman JG, Lunden M, Schmitz CC. Unanticipated teaching demands rise with simulation training: strategies for managing faculty workload. J Surg Educ. 2015;72(3):522-529. doi:10.1016/j.jsurg.2014.10.013
28. Alkire BC, Raykar NP, Shrime MG, et al. Global access to surgical care: a modelling study. Lancet Glob Health. 2015;3(6):e316-323. doi:10.1016/S2214-109X(15)70115-4
29. Surgical Simulation and Technology-Enabled Learning in Resource Limited Countries: A Review. J Surg. Published online 2021. doi:10.29011/2575-9760.001355
30. Aihua Z. New Ecology of AI-Assisted Language Education. J Phys Conf Ser. 2021;1861(1):012040. doi:10.1088/1742-6596/1861/1/012040
31. 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
32. Morris MX, Fiocco D, Caneva T, Yiapanis P, Orgill DP. Current and future applications of artificial intelligence in surgery: implications for clinical practice and research. Front Surg. 2024;11. doi:10.3389/fsurg.2024.1393898
33. Using the Operating Room Black Box to Assess Surgical Team M... : Annals of Surgery. Accessed May 23, 2024. https://journals.lww.com/annalsofsurgery/abstract/9900/using_the_operating_room_black_box_to_assess.737.aspx
34. Møller KE, Sørensen JL, Topperzer MK, et al. Implementation of an Innovative Technology Called the OR Black Box: A Feasibility Study. Surg Innov. 2023;30(1):64-72. doi:10.1177/15533506221106258
35. Xue Y, Hu A, Muralidhar R, Ady JW, Bongu A, Roshan U. An AI system for evaluating pass fail in fundamentals of laparoscopic surgery from live video in realtime with performative feedback. In: IEEE Computer Society; 2023:4167-4171. doi:10.1109/BIBM58861.2023.10385428
36. Bian Y, Xiang Y, Tong B, Feng B, Weng X. Artificial Intelligence-Assisted System in Postoperative Follow-up of Orthopedic Patients: Exploratory Quantitative and Qualitative Study. J Med Internet Res. 2020;22(5):e16896. doi:10.2196/16896
37. Amri MM, Hisan UK. Incorporating AI Tools into Medical Education: Harnessing the Benefits of ChatGPT and Dall-E. J Nov Eng Sci Technol. 2023;2(02):34-39. doi:10.56741/jnest.v2i02.315
38. Silvestri C. Creation and Evaluation of a Novel Artificial Intelligence (AI)-Powered Chatbot for Oral-Boards Scenarios. Plenary presented at: Association of Surgical Education 2024 Annual Meeting; April 2024; Orlando, FL.
39. Subhas G, Yoo S, Chang YJ, et al. Benefits of mock oral examinations in a multi-institutional consortium for board certification in general surgery training. Am Surg. 2009;75(9):817-821.
40. Webb JJ. Proof of Concept: Using ChatGPT to Teach Emergency Physicians How to Break Bad News. Cureus. 2023;15(5). doi:10.7759/cureus.38755
41. Botelho F, Tshimula JM, Poenaru D. Leveraging ChatGPT to Democratize and Decolonize Global Surgery: Large Language Models for Small Healthcare Budgets. World J Surg. 2023;47(11):2626-2627. doi:10.1007/s00268-023-07167-2
42. Wireko AA, Ng JC, David L, Abdul-Rahman T, Sikora V, Isik A. Calling for continuous surgical support in Ukraine. Int J Surg Lond Engl. 2023;110(1):571-573. doi:10.1097/JS9.0000000000000040
43. Dzhemiliev A, Kizub D, Welten VM, et al. Patient Care and Surgical Training During Armed Conflict: Experiences and Perspectives of Surgical Residents in Ukraine. Ann Surg. 2023;278(1):19-21. doi:10.1097/SLA.0000000000005873
44. Church KW, Chen Z, Ma Y. Emerging trends: A gentle introduction to fine-tuning. Nat Lang Eng. 2021;27(6):763-778. doi:10.1017/S1351324921000322
45. Suresh S, Tavabi N, Golchin S, et al. Intermediate Domain Finetuning for Weakly Supervised Domain-adaptive Clinical NER. In: Demner-fushman D, Ananiadou S, Cohen K, eds. The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. Association for Computational Linguistics; 2023:320-325. doi:10.18653/v1/2023.bionlp-1.29
46. Ovadia O, Brief M, Mishaeli M, Elisha O. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. Published online January 30, 2024. doi:10.48550/arXiv.2312.05934
47. Wołk K, Marasek K. Enhanced Bilingual Evaluation Understudy. Published online September 30, 2015. doi:10.48550/arXiv.1509.09088