OPEN

PRS GLOBAL OPEN

ORIGINAL ARTICLE

Technology

Artificial Intelligence Augmentation: Performance of GPT-4 and GPT-3.5 on the Plastic Surgery In-service Examination

Daniel Najafali, BS[*]
Erik Reiche, MD[†]
Sthefano Araya, MD[‡]
Manuel Orellana, MD[§]
Farrah C. Liu, MD[¶]
Justin M. Camacho, MBA[‡]
Sameer A. Patel, MD[‡]
Justin M. Broyles, MD[†]
Amir H. Dorafshar, MD[‖]
Shane D. Morrison, MD, MS[**][††]
Leonard Knoedler, MD[‡‡]
Paige M. Fox, MD, PhD[¶]

Background: ChatGPT-3.5 scored in the 52nd percentile of the Plastic Surgery In-service Examination, making its knowledge equivalent to a first-year integrated resident. The updated GPT-4 may have improved performance given its more expansive training set. We hypothesized that GPT-4 would outperform its predecessor, making it a more valuable potential asset to surgical education.

Methods: Questions from the 2022 Plastic Surgery In-service Examination were given to GPT-4 and GPT-3.5. Both were prompted using 3 different structures. The 2022 American Society of Plastic Surgeons Norm Tables were used to compare the performance of the chatbot to national metrics from plastic surgery residents.

Results: GPT-4 answered a total of 237 questions with an overall accuracy of 63% across all 3 strategies. The accuracy was as follows for the prompting schemes: 54% for open ended, 67% for multiple choice (MC), and 68% for MC with explanation. The section with the highest accuracy (74%) among all strategies was Section 4: Breast and Cosmetic. GPT-4's highest scoring methodology (MC with explanation, 68%) placed it in the following national integrated percentiles: 93rd percentile for the first year, 76th percentile for the second year, 52nd percentile for the third year, 34th percentile for the fourth year, 17th percentile for the fifth year, and 15th percentile for the sixth year. GPT-3.5 scored 58% overall.

Conclusions: GPT-4 outperformed its predecessor but only scored in the 15th percentile compared with postgraduate year-6 residents. More refinement is needed to achieve performance metrics equivalent to an attending plastic surgeon and become a valuable tool for surgical education. *(Plast Reconstr Surg Glob Open 2025;13:e6645; doi: 10.1097/GOX.0000000000006645; Published online 10 April 2025.)*

INTRODUCTION

Large language models (LLMs) have been applied to multiple use cases relevant to plastic and reconstructive surgery (PRS) and other fields of medicine.[1–12] Recently, the Plastic Surgery In-service Examination (PSISE) was taken by ChatGPT (GPT-3.5).[13,14] It scored in the 49th percentile compared with first-year integrated residents based on the analysis by Humar et al[13]. GPT-4 is an improved version from OpenAI that is trained on more expansive data.[15] The PSISE is challenging and assesses trainee's knowledge objectively in 5 core competency areas.[16]

To measure the degree of potential improvement from the previous iteration of OpenAI's chatbot (GPT-3.5 → GPT-4) with respect to plastic surgery, we sought to challenge it

*From the *Carle Illinois College of Medicine, University of Illinois Urbana-Champaign, Urbana, IL; †Division of Plastic and Reconstructive Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; ‡Division of Plastic and Reconstructive Surgery, Fox Chase Cancer Center, Temple University Philadelphia, PA; §Department of Surgery, Harbor-UCLA Medical Center, Torrance, CA; ¶Division of Plastic and Reconstructive Surgery, Department of Surgery, Stanford University School of Medicine, Stanford, CA; ‖Division of Plastic and Reconstructive Surgery, Rush University Medical Center, Chicago, IL; **Division of Plastic and Reconstructive Surgery, University of Washington at Harborview Medical Center, Seattle, WA; ††Department of Urology, University of Washington Medical Center, Seattle, WA; and ‡‡Department of Plastic, Hand and Reconstructive Surgery, University Hospital Regensburg, Regensburg, Germany.*

DOI: 10.1097/GOX.0000000000006645

Received for publication April 25, 2024; accepted January 27, 2025.

Mr. Najafali and Dr. Reiche contributed equally to this work.

Dr. Knoedler and Dr. Fox contributed equally to this work.

The data are available from this article upon reasonable request to the corresponding author.

with the PSISE. To clarify the relevance of chatbot performance on the PSISE for practicing surgeons, we emphasize that understanding its potential and limitations in enhancing clinical decision-making support, providing continuing medical education, and ultimately assisting patients with their questions will promote more informed use of this dynamic technology. This study aims to measure GPT-4's and GPT-3.5's performance metrics on the PSISE and compare them with national resident metrics. We hypothesized that GPT-4 would outperform its predecessor GPT-3.5 on the PSISE, making it a more valuable potential asset to surgical education.

METHODS

Natural Language Processing Artificial Intelligence

GPT-4 and -3.5 Architecture

GPT-4 (OpenAI, San Francisco, CA) was released on March 13, 2023. It is an update to the GPT-3.5 architecture, which has been used in multiple applications. The second LLM utilized in this study was OpenAI's 2023 ChatGPT 3.5 February release.

Plastic Surgery Examination

The 2022 American Society of Plastic Surgeons (ASPS) PSISE was used. The questions that were excluded by the test writing committee were also excluded in this study, along with other questions that could not be prompted based on structure. This study received approval from ASPS Education leadership, and institutional review board approval was not required.

Prompting Strategies

The questions were prompted in 3 ways similar to the methodology by Kung et al[17]. The method of prompting GPT-3.5 and GPT-4 was as follows: (1) open-ended (OE) format that removed all answer choices, (2) multiple choice (MC) format with answer choices after the phrase "please select the correct answer," and (3) MC format with answer choices and prompted reasoning (MC with explanation [MCE]) after the phrase "please select the correct answer and provide an explanation."

Each question was asked in a new chat. The 2022 PSISE had 5 sections which consisted of Section 1: Comprehensive; Section 2: Hand and Lower Extremity; Section 3: Craniomaxillofacial; Section 4: Breast and Cosmetic; and Section 5: Core Surgical Principles.

Performance Evaluation and Grading of Responses for Accuracy

For the OE, MC, and MCE strategies, a standardized rubric similar to Kung et al was used with correct, incorrect, and indeterminate answers tabulated for each methodology. (**See figure, Supplemental Digital Content 1,** which displays the rubric for grading adapted from Kung et al, **http://links.lww.com/PRSGO/D935**).[17] Accuracy was defined by a correct answer. Concordance was determined by reviewers by examining the explanation of the chatbot for its chosen answer, similar to Kung et al. GPT-3.5 was found to have very high answer-explanation concordance for the United States Medical Licensing Examinations by Kung et al. Chatbot performance was stratified by examination section. ASPS Norm Tables were used to determine the performance metrics on a national level. Analysis included metrics on accuracy, concordance, and insight, which were assessed by 2 independent reviewers and tabulated. Interrater agreement was measured using weighted Kappa scores and demonstrated excellent agreement among our team (scores between 0.81 and 1.00).

Takeaways

Question: How does GPT-4 compare with its predecessor, GPT-3.5, with respect to plastic and reconstructive surgery questions?

Findings: Utilizing a variety of prompting strategies and comparing the performances of GPT-3.5 and GPT-4 on the Plastic Surgery In-service Examination, the study found GPT-4 outperformed its predecessor.

Meaning: GPT-4 can be a valuable educational tool for plastic surgery residents but still requires additional refinement.

Statistical Analysis

Data were collected using a standardized Microsoft Excel spreadsheet (Microsoft Corp., Redmond, WA). Descriptive statistics were captured with mean (± SD), median (interquartile range), or frequency (percentage). Normality was determined using the Shapiro–Wilk test. Variables continuous in nature were compared via a Student *t* test or via a Mann-Whitney U test as appropriate. Chi-square test with Yates correction or Fisher exact test was used for categorical variables. All statistical analysis was performed in the R (version 4.1.0) software in the RStudio (version 1.4.1717) environment. Statistical significance was met via 2-sided *P* values less than 0.05.

RESULTS

GPT-4

From the initial questions (N = 250), 7 questions were removed by the examination committee due to ambiguity or poor statistical performance (N = 243). Another 6 questions were removed (N = 237), 5 of which pertained to graphics and 1 that had multiple columns that needed to be evaluated in the answer. A total of 237 questions were analyzed using GPT-4. Table 1 summarizes the performance of GPT-4 in comparison to GPT-3.5 stratified by residency training type and year. GPT-4's highest scoring methodology (MCE, 68%) placed it in the following national percentiles compared with plastic surgery

Disclosure statements are at the end of this article, following the correspondence information.

Related Digital Media are available in the full-text version of the article on www.PRSGlobalOpen.com.

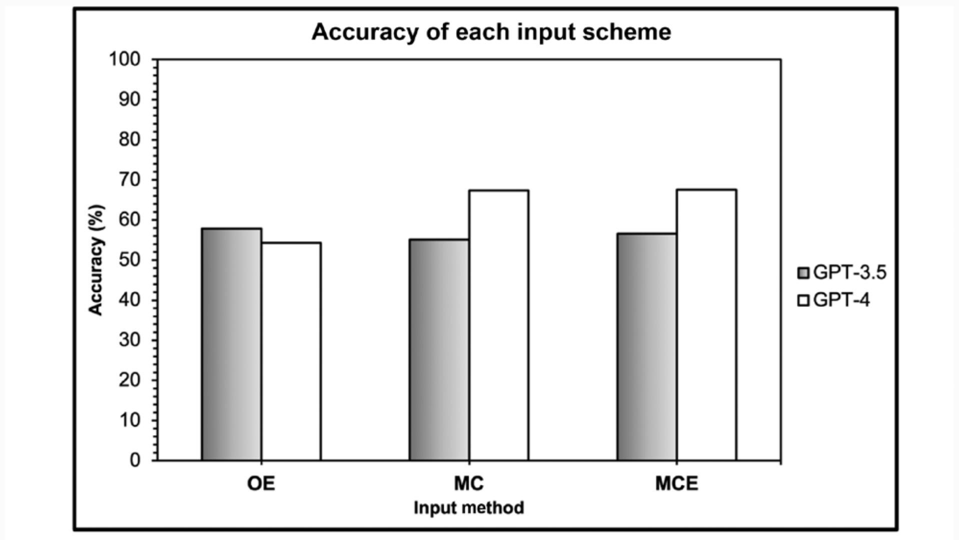
integrated residents who took the examination: 93% for the first year, 76% for the second year, 52% for the third year, 34% for the fourth year, 17% for the fifth year, and 15% for the sixth year. GPT-3.5 scored 58% overall. Table 2 summarizes accuracy and concordance for each section of the examination. Accuracy for GPT-4 based on OE questions was 54%, MC was 67%, and MCE was 68% (Fig. 1). Accuracy was highest using the MCE method. A total of 4 questions were indeterminate (3 OE and 1 MC). The method of prompting GPT-4 demonstrated that answer

Table 1. Norm Table Corresponding to the Highest Accuracy Achieved by GPT-4 (68%) and GPT-3.5 (58%)

	Total Test	Independent Program				Integrated Program					
Chatbot	% Correct	All	First Year	Second Year	Third Year	First Year	Second Year	Third Year	Fourth Year	Fifth Year	Sixth Year
GPT-4	68%	52	89	67	48	93	76	52	34	17	15
GPT-3.5	58%	14	29	11	10	52	16	7	4	0	0

Table 2. Performance Metrics of GPT-4 Stratified by PSISE Section

Variables	Section 1: Comprehensive	Section 2: Hand and Lower Extremity	Section 3: Craniomaxillofacial	Section 4: Breast and Cosmetic	Section 5: Core Surgical Principles	*P*
Accuracy OE, (%)						
Inaccurate	32.7	51.0	47.9	41.3	53.3	0.28
Accurate	63.3	49.0	50.0	58.7	46.7	
Indeterminate	4.1	0.0	2.1	0.0	0.0	
Concordance OE, (%)	100.0	98.0	100.0	97.8	100.0	0.55
Accuracy MC, (%)						
Inaccurate	32.7	42.9	31.2	26.1	28.9	0.49
Accurate	65.3	57.1	68.8	73.9	71.1	
Indeterminate	2.0	0.0	0.0	0.0	0.0	
Concordance MC, (%)	100.0	100.0	100.0	97.8	97.8	0.52
Accuracy MCE, (%)						
Inaccurate	34.7	32.7	27.1	34.8	33.3	0.93
Accurate	65.3	67.3	72.9	65.2	66.7	
Indeterminate	0.0	0.0	0.0	0.0	0.0	
Concordance MCE, (%)	98.0	100.0	100.0	93.5	95.6	0.52



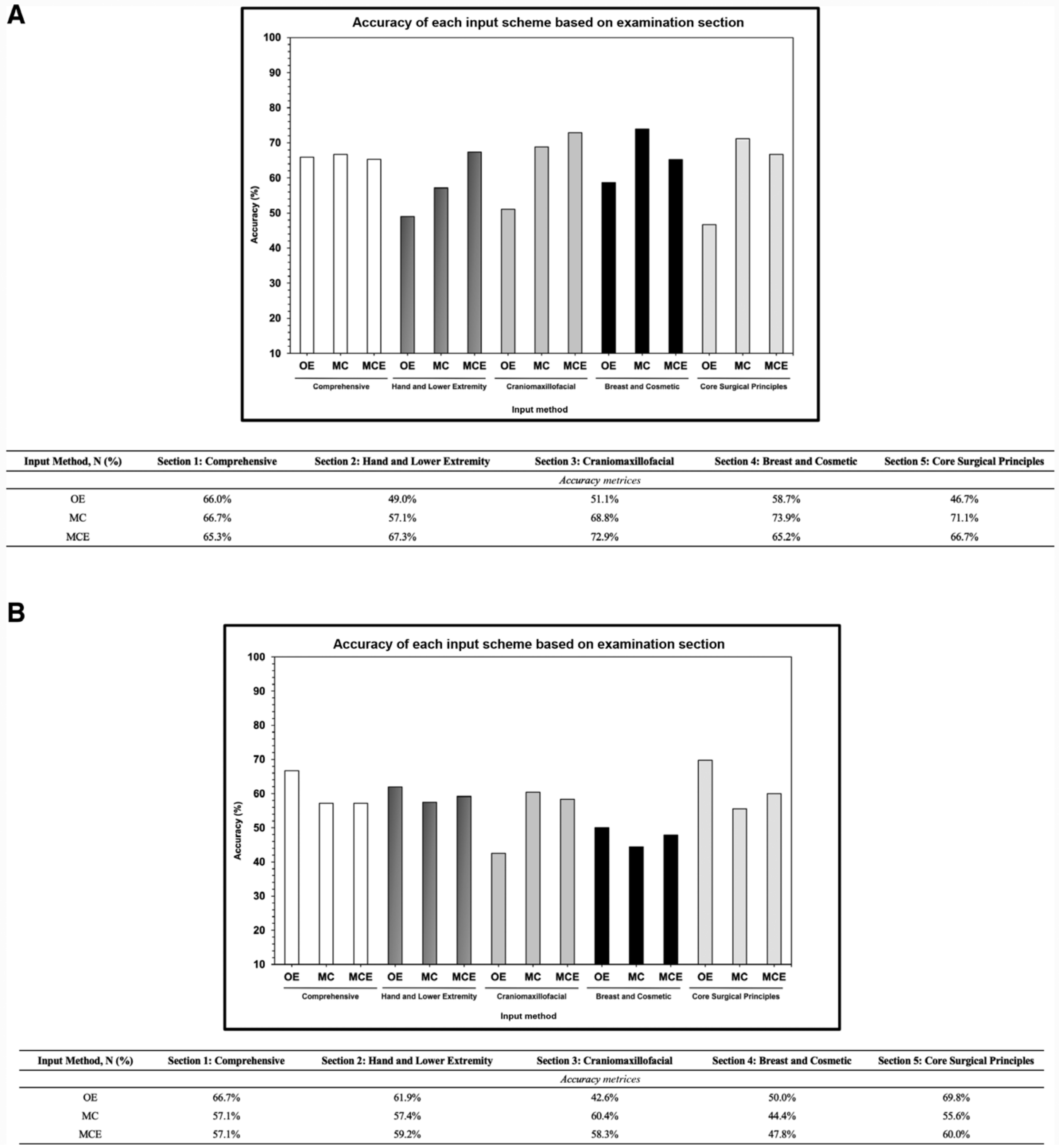
Performance variables	OE	MC	MCE
Accuracy GPT-3.5, N (%)*	57.8% (129/223)	55.0% (129/234)	56.5% (134/237)
Accuracy GPT-4, N (%)**	54.3% (127/234)	67.4% (159/236)	67.5% (160/237)

Fig. 1. Performance based on input strategy presented to the chatbots for the 2022 PSISE. *OE had 14 questions and MC had 3 questions that were indeterminate and were not included in the final accuracy calculation. **OE had 3 questions and MC had 1 question that were indeterminate and were not included in the final accuracy calculation.

responses for the MCE method did not yield any indeterminate responses. Figure 2A depicts performance of each strategy based on the section of the examination and strategy used. Section 4: Breast and Cosmetic had the best performance using the MC strategy (74%).

GPT-3.5

A total of 237 questions were included in the final analysis of GPT-3.5 (Table 1). The performance metrics stratified by the ASPS In-service Examination sections are reported in Table 3. Accuracy for OE questions, MC, and MCE were 57.8%, 55.0%, and 56.5%, respectively (Fig. 1). The OE approach was the most accurate for GPT-3.5. A total of 17 (7%) questions were indeterminate amongst all strategies used (14 OE and 3 MC). Output and concordance were highest for OE questions at 100%, followed by 97.5% for MCE and 96.6% for MC. GPT-3.5 was insightful on 82.7% (196 of 237) of questions across all structured



Input Method, N (%)	Section 1: Comprehensive	Section 2: Hand and Lower Extremity	Section 3: Craniomaxillofacial	Section 4: Breast and Cosmetic	Section 5: Core Surgical Principles
			Accuracy metrices		
OE	66.0%	49.0%	51.1%	58.7%	46.7%
MC	66.7%	57.1%	68.8%	73.9%	71.1%
MCE	65.3%	67.3%	72.9%	65.2%	66.7%

Input Method, N (%)	Section 1: Comprehensive	Section 2: Hand and Lower Extremity	Section 3: Craniomaxillofacial	Section 4: Breast and Cosmetic	Section 5: Core Surgical Principles
			Accuracy metrices		
OE	66.7%	61.9%	42.6%	50.0%	69.8%
MC	57.1%	57.4%	60.4%	44.4%	55.6%
MCE	57.1%	59.2%	58.3%	47.8%	60.0%

Fig. 2. Input strategy-related performance metrics. A, Performance based on input strategy presented to GPT-4 for the 2022 PSISE stratified by section. B, Performance based on input strategy presented to GPT-3.5 for the 2022 PSISE stratified by section.

Table 3. Performance Metrics of GPT-3.5 Stratified by PSISE Section

Variables	Section 1: Comprehensive	Section 2: Hand and Lower Extremity	Section 3: Craniomaxillofacial	Section 4: Breast and Cosmetic	Section 5: Core Surgical Principles	*P*
Images, (%)	12.2	24.5	27.7	2.2	0.0	**<0.001**
Accuracy OE, (%)						
Inaccurate	30.6	32.7	56.2	50.0	28.9	**0.007**
Accurate	61.2	53.1	41.7	50.0	66.7	
Indeterminate	8.2	14.3	2.1	0.0	4.4	
Concordance OE, (%)	100.0	100.0	100.0	100.0	100.0	0.43
Insight OE, (%)	98.0	100.0	93.8	97.8	97.8	0.40
OE character count, median [IQR]	365.00 [299.0–405.00]	305.00 [191.00–381.00]	264.00 [196.25–337.00]	320.00 [247.75-429.00]	384.00 [241.00–541.00]	**<0.001**
OE word count, median [IQR]	56.00 [45.00–65.00]	46.00 [30.00–59.00]	39.50 [29.75–52.00]	48.50 [36.75–61.00]	58.00 [37.00–81.00]	**<0.001**
Accuracy MC, (%)						
Inaccurate	42.9	40.8	39.6	54.3	44.4	0.43
Accurate	57.1	55.1	60.4	43.5	55.6	
Indeterminate	0.0	4.1	0.0	2.2	0.0	
Concordance MC, (%)	98.0	100.0	97.9	87.0	100.0	**0.002**
Insight MC, (%)	93.9	91.8	72.9	84.8	84.4	**0.031**
MC character count, median [IQR]	544.00 [473.00–647.00]	481.00 [377.00–592.00]	445.00 [391.00–544.50]	500.00 [419.00–621.50]	617.00 [441.00–701.00]	**0.002**
MC word count, median [IQR]	86.00 [71.00–95.00]	76.00 [59.00–91.00]	66.00 [57.75–80.25]	76.00 [63.25–95.00]	92.00 [68.00–108.00]	**0.001**
Accuracy MCE, (%)						
Inaccurate	42.9	41.0	42.0	52.0	40.0	0.76
Accurate	57.1	59.2	58.3	47.8	60.0	
Indeterminate	0.0	0.0	0.0	0.0	0.0	
Concordance MCE, (%)	98.0	100.0	100.0	93.5	95.6	0.19
Insight MCE, (%)	98.0	100.0	81.2	95.7	97.8	**<0.001**
MCE character count, median [IQR]	586.00 [500.00–694.00]	508.00 [404.00–619.00]	472.00 [418.00–571.50]	527.00 [446.00–648.50]	644.00 [468.00–728.00]	**0.002**
MCE word count, median [IQR]	90.00 [75.00–100.00]	80.00 [63.00–95.00]	70.00 [61.75–84.25]	80.00 [67.25–99.00]	96.00 [72.00–112.00]	**0.001**

IQR, interquartile range.
Bold values indicate statistical significance ($P < 0.05$).

formats. When faced with a question relating to graphics or images GPT-3.5 was most accurate for images using an MCE approach 59.4%.

The method of prompting GPT-3.5 demonstrated that answer responses for the MCE method did not yield any indeterminate responses. Performance metrics of each input method stratified by each of the examination section are depicted in Figure 2B. Core surgical principles had the highest accuracy for the OE and MCE method across all sections. The OE strategy was lowest for the craniomaxillofacial section. The worst performing strategy across all sections was the MC approach for breast and cosmetic.

Based on the 2022 ASPS Norm Table and the highest scoring method of input (OE, 57.8%, ~58%), GPT-3.5 performed at the 52nd percentile compared with first-year residents from integrated programs (N = 185).

DISCUSSION

This study measured the performance of OpenAI's latest chatbot, GPT-4, on the PSISE and compared it with its predecessor GPT-3.5. GPT-4 outperformed GPT-3.5 on overall performance on the PSISE. The metrics for GPT-4 also corresponded to a higher percentile than GPT-3.5 across all residency training years when looking at the national trends for the examination. Its performance metrics are at a level equivalent to the 93rd percentile for first-year integrated plastic surgery residents.

The most plausible explanation for GPT-4's superior performance in comparison to GPT-3.5 is its advanced reasoning capabilities.[18] Other studies evaluating the GPT-3.5 architecture on important medical examinations such as the United States Medical Licensing Examination, neurosurgical boards, and the PSISE found that the platform was capable of generating outputs with correct answers to these complex multidisciplinary questions.[13,14,17,19,20] The aforementioned article evaluating the neurosurgical oral boards found that GPT-4 outperformed Google's Bard. PRS literature has also investigated Google's Bard, prompting it with PSISE questions and measuring its performance.[21]

A major limitation of previous studies on PSISE is that the chatbot performed decently compared with first-year integrated residents, but in some cases scored in the lowest percentile (0%) when compared with metrics of more senior residents. Humar et al[13] concluded that ChatGPT did not have the complex ability to drive context-dependent medical decision-making required throughout a surgical residency. The overall "performance" of the LLM is influenced by its training architecture and design to predict the next word in a document using publicly available data (eg, internet data, licensed data) combined with reinforcement learning with human feedback.[15] Evaluating national performance and placing the percentile of GPT-4 into the context of resident metrics, its 68% correct with MCE makes it similar to the median of third-year residents. Given the percentile similarity to residents in their third year and the fashion of OpenAI's training, this may reflect that residents at this level of training have a broad understanding of the specialty with a vast amount of knowledge found across the internet (eg, journal articles, reports from national societies) or could be explained by other factors such as increased experience with the expectations of the in-service examination, more effective learning strategies, or other factors beyond their fund of knowledge. It is important to compare GPT-4's performance to residents in their later years of training too. The highest score was observed with MCE prompting. Based on our study's findings, we found that GPT-4 scored in the 15th percentile compared with the most senior residents (postgraduate year 6). This places GPT-4's performance below the thresholds determined by Girotto et al[22] when evaluating all available in-service examination scores (cutoff percentile for all residents = 24th; area under the curve = 81.4%) and when evaluating in-service examination scores taken in the final year of training (cutoff percentile for all residents = 31st; area under the curve = 84.6%). GPT-4 and other LLMs will likely continue to face scrutiny, particularly when their overall score on such standardized examinations falls in a lower percentile compared with humans. Although current capabilities of chatbots are impressive, these technologies and LLMs will likely continue to remain as adjuncts to human expertise. Serving alongside trainees and physicians with a supportive role in patient care is a more probable function for LLMs, rather than functioning as autonomous replacements. Their performance can certainly be augmented with greater specificity and tailored training that is supplemented with more robust medical data. Applying chatbots to the PRS written boards and observing their "performance" will be a possible next step.[23] Generating LLMs that are trained on medical data, literature, and relevant examinations to a specialty could be the path forward in improving these technologies for medical applications.

Recently, ChatGPT was found in a cross-sectional study to exhibit empathy at a higher degree than humans: "…evaluators preferred chatbot responses to physician responses in 78.6% of the 585 evaluations."[24] Chatbots such as GPT-4 and its predecessor, GPT-3.5, demonstrate remarkable abilities; however, it should be noted that the perspective of evaluating humans versus chatbots (eg, in this case plastic surgery residents versus GPT-4) is potentially a flawed perspective. Chatbots should be seen as a tool that can potentially improve resident education and the quality of care for our patients. Being able to seamlessly integrate chatbots into medicine to make us more efficient, decrease resource utilization, and increase access to care could be a tremendous benefit.

The role of LLMs in education naturally extends from the convenience of their question-and-answer format from the prompting environment that handles queries followed by a generated output. These virtual interactions can take the form of a conversation. Huang et al evaluated the performance of chatbots on a family medicine test, which led them to propose potential applications for medical education such as generating examination questions and scenarios, and serving as a resource for medical information.[25] Grigorian et al[26] evaluated GPT-4 in answering surgery shelf questions and determined that chatbots can empower surgical educators by generating high-quality MC questions that encompass a range of difficulties. Moreover, they suggest, with additional refinements to the model, that trainees may use GPT-4 to create questions that reinforce their learning instead of seeking additional resources. Mohapatra et al[27] evaluated the role that LLMs could have in PRS residency training by playing the role of a "teaching assistant." However despite the multiple applications and ease of use, trainees and educators must be cautious as chatbots have limitations. Users should be aware of the possibility of inaccurate information, incorrect references, and "hallucinations" of chatbots that are not typically associated with traditional vetted resources. An area that has been less explored is the ability of LLMs and generative artificial intelligence to generate anatomically relevant images with appropriate descriptions for applications in medical education. Midjourney was investigated for medical anatomy by Buzzaccarini et al,[28] with the conclusion that their lack of accuracy renders them ineffective for medical education and can lead to potential misconceptions. In a response to the article, Ozmen et al[29] suggested that it is too premature to write off generative artificial intelligence and its utility in medical illustrations based on a single tool like Midjourney that was not designed for the specific purpose of medical imagery, but for artistic purposes. We agree that these tools are powerful, and with refinement towards medical applications, they can be valuable assets. GPT-4 can be leveraged to develop highly specific prompts for artificial intelligence-based image generators. OpenAI's DALL·E text-to-image models along with other ones on the market such as Stable Diffusion may become more geared toward medical imagery in the future. Knoedler et al[30] demonstrated the power of a generative adversarial network that is trained with preoperative and postoperative images for rhinoplasty outcomes. Tailoring text-to-image generators, generative adversarial networks, and chatbots with PRS-specific features could drastically improve these models in the future.[31] "Synthetic" data may also be augmented with real data to produce additional data frames that can be leveraged for medical education and applications.

For artificial intelligence to reach a state where it assists plastic surgeons in decision-making, it must demonstrate that its predictability capacity is at or beyond that of board-certified clinicians. A plastic surgery resident is knowledgeable in a wide array of domains beyond written examination skills, including technical skills (eg, intricate flap techniques, anatomy) and nontechnical skills (eg, developing rapport with patients). Clinical decision-making is multifaceted and needs to consider the patient's wishes. With the advancement of artificial intelligence and the introduction of new chatbots strictly focused on medical applications, the performance and improvements of these systems should open the doors to a reality where chatbots are integrated to improve medical care but not designed to replace the surgeon. A blend of artificial intelligence and the surgeon is possible, as demonstrated by the da Vinci Surgical System, improving surgical maneuvers with surgeon direction.[32] There are also ethical concerns that trouble the medical community that need to be further explored with proper safeguards if chatbots are to be widely adopted.[33,34]

Chatbots and similar artificial intelligence should be leveraged to increase the quality of education for plastic surgery residents and integrated into medicine to assist healthcare delivery, quality, and access. Chatbots have many possibilities to improve plastic surgery trainee education via studying, question generation, and evaluation of the available literature. Trainees use a variety of resources when studying for the PSISE, and no single resource is ideal. The ability to combine multiple resources through chatbot technology could create a valuable educational tool.

Several recent studies have elucidated the benefit of integrating chatbots into medical education. Li et al[35] created a custom artificial intelligence–powered chatbot that aided medical students in learning anatomy. Students reported feeling more comfortable making mistakes conversing with the artificial intelligence compared with conversations with human instructors.[35] Furthermore, chatbots have been recently utilized to improve patient medical education. Gortz et al[36] designed a user-friendly chatbot trained on evidence-based content to accompany prostate cancer patients through their diagnosis and treatment decision process. The study revealed that despite being comprised of older patients (mean age = 68 y), participants managed to utilize the chatbot without much assistance and expressed the desire to increase the use of chatbots in healthcare.[36] Regardless of a desire to adopt the use of chatbots and artificial intelligence in healthcare for both patients and healthcare professionals, medical students currently still lack educational opportunities to increase their knowledge in this evolving field.[37,38]

Another intriguing application of chatbot technology would be to improve standardized examination question writing. Reevaluating questions to which the chatbot cannot find answers and determining the reason the answer is not correct or available may help eliminate ambiguous or unclear questions.

Chatbot Limitations

There are also limitations to the chatbot's answers, which tend to defer some aspects of offering a treatment or diagnosis. Inherently, GPT-4 is limited by its training data. PRS literature, like other fields of medicine, has a considerable number of high-quality articles as subscription based and not open access. Textbooks and their chapters are also likely to be behind paywalls, which may not have been available to LLMs such as GPT-3.5 and GPT-4 during their training period. Democratizing such data with collaboration among software engineers and stakeholders (eg, journal editorial boards, authors, domain experts) may increase the overall quality of these platforms and the available parameters of the model used. The complexity behind the task of answering these questions highlights that perhaps the parameters it has been trained on, despite being extensive, is a limitation for this use case.

Study Limitations

This study is limited by the ability to prompt certain PSISE questions based on their style and by the number of prompts given to the chatbot. One year of the PSISE was used, but this was the most recent examination available, and the number of prompts was a sufficient sample size to observe a difference in performance between the 2 chatbots. Future studies should investigate the answers that the chatbot incorrectly provided to determine its reasoning for selecting those answer choices. This study did not evaluate persona-based prompting (eg, respond like a board-certified plastic surgeon) which should also be investigated.[39,40]

CONCLUSIONS

GPT-4 outperformed its predecessor but only scored in the 15th percentile compared with postgraduate year-6-integrated plastic surgery residents. Increased refinement is needed to allow chatbots to become a more powerful tool that truly enhances patient care. Surgical educators should continue to explore chatbots to determine if they can improve resident education through collation of educational materials and quality question generation for assessment of knowledge acquisition and retention.

Paige M. Fox, MD, PhD
Division of Plastic and Reconstructive Surgery
Department of Surgery, Stanford University School of Medicine
770 Welch Road, Suite 400
Palo Alto, CA 94304
E-mail: pfox@stanford.edu
Instagram: @drpaigefox

Leonard Knoedler, MD
Department of Plastic, Hand and Reconstructive Surgery
University Hospital Regensburg
Universität Regensburg
Zentraler Rechnungseingang oa@ur.de
ggf. Universitätsstraße 31, nicht Franz-Josef-Strauß-Allee etc.
93040 Regensburg, Germany
E-mail: lknoedler@mgh.harvard.edu

DISCLOSURE

Dr. Dorafshar is entitled to a technology he invented based on a licensing agreement between KLS Martin (Jacksonville, FL) and the Johns Hopkins University (Baltimore, MD), is entitled to receive royalties on a book he published under a licensing agreement with Elsevier (Amsterdam, the Netherlands), receives honorarium *for educational lectures from De Puy Synthes (Raynham, MA), receives indirect research support from KLS Martin and De Puy Synthes, and has received payment from KCI (Sparks, MD) as a consultant. The other authors have no financial interest to declare in relation to the content of this article.*

ACKNOWLEDGMENT

The authors would like to acknowledge the use of OpenAI's GPT-3.5 and GPT-4 in answering questions related to the in-service examination.

REFERENCES

1. Najafali D, Galbraith LG, Camacho JM, et al. Addressing the rhino in the room: ChatGPT creates "novel" patent ideas for rhinoplasty. *Eplasty*. 2024;24:e13. https://www.hmpgloballearningnetwork.com/site/eplasty/original-research/addressing-rhino-room-chatgpt-creates-novel-patent-ideas-rhinoplasty.
2. Najafali D, Hinson C, Camacho JM, et al. Can chatbots assist with grant writing in plastic surgery? Utilizing ChatGPT to start an R01 grant. *Aesthet Surg J*. 2023;43:NP663–NP665.
3. Najafali D, Camacho JM, Galbraith LG, et al. Ask and you shall receive: OpenAI ChatGPT writes us an editorial on using chatbots in gender affirmation surgery and strategies to increase widespread adoption. *Aesthet Surg J*. 2023;43:NP715–NP717.
4. Najafali D, Hinson C, Camacho JM, et al. Artificial intelligence knowledge of evidence-based recommendations in gender affirmation surgery and gender identity: is ChatGPT aware of WPATH recommendations? *Eur J Plast Surg*. 2023;46:1169–1176.
5. Aljindan FK, Shawosh MH, Altamimi L, et al. Utilization of ChatGPT-4 in plastic and reconstructive surgery: a narrative review. *Plast Reconstr Surg Glob Open*. 2023;11:e5305.
6. Najafali D, Dorafshar AH. Commentary on: evaluating chatbot efficacy for answering frequently asked questions in plastic surgery: a ChatGPT case study focused on breast augmentation. *Aesthet Surg J*. 2023;43:1136–1138.
7. Arif F, Safri MK, Shahzad Z, et al. Exploring the application of CHATGPT in plastic surgery: a comprehensive systematic review. *J Pak Med Assoc*. 2024;74:S17–S28.
8. Patel V, Deleonibus A, Wells MW, et al. Distinguishing authentic voices in the age of ChatGPT: comparing AI-generated and applicant-written personal statements for plastic surgery residency application. *Ann Plast Surg*. 2023;91:324–325.
9. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery*. 2023;93:1353–1365.
10. Lee H. The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ*. 2023;17:926.
11. Frosolini A, Franz L, Benedetti S, et al. Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. *Eur Arch Otorhinolaryngol*. 2023;280:5129–5133.
12. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. 2023;29:721–732.
13. Humar P, Asaad M, Bengur FB, et al. ChatGPT is equivalent to first year plastic surgery residents: evaluation of ChatGPT on the Plastic Surgery In-Service exam. *Aesthet Surg J*. 2023;43:NP1085–NP1089.
14. Gupta R, Herzog I, Park JB, et al. Performance of ChatGPT on the plastic surgery inservice training examination. *Aesthet Surg J*. 2023;43:NP1078–NP1082.
15. OpenAI. GPT-4 technical report. 2023. . Available at https://ui.adsabs.harvard.edu/abs/2023arXiv230308774O. Accessed March 01, 2023.
16. Frojo G, Tadisina KK, Kaswan S, et al. Preparing for the Plastic Surgery In-Service Exam: evidence-based essentials for the plastic surgery resident. *Plast Reconstr Surg*. 2019;143:256e–257e.
17. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2:e0000198.
18. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388:1233–1239.
19. 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*. 2023;139:904–911.
20. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *medRxiv*. 2023;93:2023.04.06.23288265.
21. Najafali D, Reiche E, Araya S, et al. Bard versus the 2022 American Society of Plastic Surgeons In-Service Examination: performance on the examination in its intern year. *Aesthet Surg J Open Forum*. 2024;6:ojad066.
22. Girotto JA, Adams NS, Janis JE, et al. Performance on the Plastic Surgery In-Service Examination can predict success on the American Board of Plastic Surgery written examination. *Plast Reconstr Surg*. 2019;143:1099e–1105e.
23. The American Board of Plastic Surgery. Statistics. Available at https://www.abplasticsurgery.org/about-us/statistics/. Accessed June 19, 2024.
24. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183:589.
25. Huang RS, Lu KJQ, Meaney C, et al. Assessment of resident and AI chatbot performance on the University of Toronto family medicine residency progress test: comparative study. *JMIR Med Educ*. 2023;9:e50514.
26. Grigorian A, Shipley J, Nahmias J, et al. Implications of using chatbots for future surgical education. *JAMA Surg*. 2023;158:1220–1222.
27. Mohapatra DP, Thiruvoth FM, Tripathy S, et al. Leveraging Large Language Models (LLM) for the plastic surgery resident training: do they have a role? *Indian J Plast Surg*. 2023;56:413–420.
28. Buzzaccarini G, Degliuomini RS, Borin M, et al. The promise and pitfalls of AI-generated anatomical images: evaluating midjourney for aesthetic surgery applications. *Aesthetic Plast Surg*. 2024;48:1874–1883.
29. Ozmen BB, Schwarz GS. Letter to the editor: the promise and pitfalls of AI-generated anatomical images-evaluating midjourney for aesthetic surgery applications [published online ahead of print May 3, 2024]. *Aesthetic Plast Surg*. 2024.
30. Knoedler S, Alfertshofer M, Simon S, et al. Turn your vision into reality-AI-powered pre-operative outcome simulation in rhinoplasty surgery. *Aesthetic Plast Surg*. 2024;48:4833–4838.
31. Ozmen BB, Schwarz GS. Future of artificial intelligence in plastic surgery: toward the development of specialty-specific large language models. *J Plast Reconstr Aesthet Surg*. 2024;93:70–71.
32. Pandya A. ChatGPT-enabled daVinci surgical robot prototype: advancements and limitations. *Robotics*. 2023;12:97.
33. Cheng K, Sun Z, He Y, et al. The potential impact of ChatGPT/GPT-4 on surgery: will it topple the profession of surgeons? *Int J Surg*. 2023;109:1545–1547.
34. Gupta R, Bagdady K, Mailey BA. Response to: truth or lies? The pitfalls and limitations of ChatGPT in systematic review creation. *Aesthet Surg J*. 2023;43:NP656–NP657.
35. Li YS, Lam CSN, See C. Using a machine learning architecture to create an AI-powered chatbot for anatomy education. *Med Sci Educ*. 2021;31:1729–1730.

36. Gortz M, Baumgartner K, Schmid T, et al. An artificial intelligence-based chatbot for prostate cancer education: design and patient evaluation study. *Digit Health*. 2023;9:20552076231173304.

37. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. 2023;11:887.

38. Pucchio A, Rathagirishnan R, Caton N, et al. Exploration of exposure to artificial intelligence in undergraduate medical education: a Canadian cross-sectional mixed-methods study. *BMC Med Educ*. 2022;22:815.

39. Milicka J, Marklova A, VanSlambrouck K, et al. Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *PLoS One*. 2024;19:e0298522.

40. Deshpande A, Murahari V, Rajpurohit T, et al. Toxicity in ChatGPT: analyzing persona-assigned language models. 2023. Available at http://arxiv.org/abs/230405335. Accessed January, 2024.