OPEN

PRS GLOBAL OPEN

ORIGINAL ARTICLE

Education

Unveiling the Potential of AI in Plastic Surgery Education: A Comparative Study of Leading AI Platforms' Performance on In-training Examinations

Nicole DiDonna, BA*
Pragna N. Shetty, MD, MPH†
Kamran Khan, MD†
Lynn Damitz, MD†

Background: Within the last few years, artificial intelligence (AI) chatbots have sparked fascination for their potential as an educational tool. Although it has been documented that one such chatbot, ChatGPT, is capable of performing at a moderate level on plastic surgery examinations and has the capacity to become a beneficial educational tool, the potential of other chatbots remains unexplored.

Methods: To investigate the efficacy of AI chatbots in plastic surgery education, performance on the 2019–2023 Plastic Surgery In-service Training Examination (PSITE) was compared among seven popular AI platforms: ChatGPT-3.5, ChatGPT-4.0, Google Bard, Google PaLM, Microsoft Bing AI, Claude, and My AI by Snapchat. Answers were evaluated for accuracy and incorrect responses were characterized by question category and error type.

Results: ChatGPT-4.0 outperformed the other platforms, reaching accuracy rates up to 79%. On the 2023 PSITE, ChatGPT-4.0 ranked in the 95th percentile of first-year residents; however, relative performance worsened when compared with upper-level residents, with the platform ranking in the 12th percentile of sixth-year residents. The performance among other chatbots was comparable, with their average PSITE score (2019–2023) ranging from 48.6% to 57.0%.

Conclusions: Results of our study indicate that ChatGPT-4.0 has potential as an educational tool in the field of plastic surgery; however, given their poor performance on the PSITE, the use of other chatbots should be cautioned against at this time. To our knowledge, this is the first article comparing the performance of multiple AI chatbots within the realm of plastic surgery education. *(Plast Reconstr Surg Glob Open 2024; 12:e5929; doi: 10.1097/GOX.0000000000005929; Published online 21 June 2024.)*

INTRODUCTION

Artificial intelligence (AI) chatbots have sparked fascination for their potential as an educational tool. Chatbots, also referred to as large language models (LLMs), are taught using extensive data sets and are trained to recognize patterns, giving them the ability to complete human-like tasks, such as engage in text conversation, answer questions, brainstorm ideas, and respond to writing prompts.[1–5] ChatGPT was one of the first publicly available LLMs, and its utility in medical education has been promising based on its passing performance on standardized examinations, including the US Medical Licensing Examination.[6,7] Gupta et al[8] sought to test ChatGPT's efficacy as an educational aid for plastic surgery residents, discovering that the platform could answer 2022 Plastic Surgery In-service Training Examination (PSITE) questions with an accuracy of 54.96%. Similarly, Humar et al[9] reported that ChatGPT scored 57% on the 2022 PSITE, which would place the chatbot in the 49th percentile for first-year integrated plastic surgery residents. Although it is well documented that ChatGPT is capable of performing at a moderate level on the PSITE and has the capacity to be a beneficial educational tool, the potential of newer chatbots remains unexplored.

This study aims to evaluate and compare the performance of seven popular chatbots on the PSITE: ChatGPT-3.5, ChatGPT-4.0, Google Bard, Google PaLM, Microsoft Bing AI, Claude, and My AI by Snapchat. To our knowledge, this is the first article comparing the

*From the *School of Medicine, University of North Carolina, Chapel Hill, N.C.; and †Division of Plastic and Reconstructive Surgery, University of North Carolina, Chapel Hill, N.C.*

Received for publication March 3, 2024; accepted May 1, 2024.

DOI: 10.1097/GOX.0000000000005929

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

performance of multiple AI chatbots within the realm of plastic surgery education.

Takeaways

Question: How do popular artificial intelligence (AI) platforms perform on the Plastic Surgery In-service Training Examination (PSITE)?

Findings: ChatGPT-4.0 outperformed other AI platforms, ranking in the 95th percentile of first-year residents on the 2023 PSITE and reaching accuracy rates up to 79%.

Meaning: ChatGPT-4.0 has potential as an educational tool in the field of plastic surgery; however, given their poor performance on the PSITE, the use of other chatbots in plastic surgery education should be cautioned against at this time.

METHODS

Question Selection

To investigate AI's role in plastic surgery education, we tested the performance of various AI platforms on the PSITE, adapting our protocol from Gupta et al.[8] The PSITE is given to plastic surgery residents annually to assess knowledge across five disciplines: comprehensive plastic surgery, hand and lower extremity surgery, craniomaxillofacial surgery, aesthetic/cosmetic surgery, and core surgical principles.[10] Examination scores allow academic programs to compare their students' performance against their peers and can be used to gauge readiness for the American Board of Plastic Surgery written examination.[10,11] As such, we determined that chatbots' PSITE scores would be an effective indicator of their plastic surgery knowledge.

Examinations and answer keys from the last 5 years (2019–2023) were obtained from the American Society of Plastic Surgeons (ASPS). As most AI chatbots are limited to processing text-based inputs, questions requiring an image or table to reach an appropriate answer were excluded unless the correct answer could be reasonably derived with solely text-based information. This determination was made depending on whether the answer key's explanation referred to the image as part of its reasoning and whether the image was described in appropriate detail in the question stem.

AI Testing

We investigated the PSITE performance of seven AI chatbots: ChatGPT-3.5, ChatGPT-4.0, Google Bard, Google PaLM, Microsoft Bing AI, Claude, and My AI by Snapchat. Before entering questions into the platforms, we provided a prompt that asked the chatbots to "answer the following multiple-choice question and provide an explanation." The prompt was input before each question for consistency.

Data Analysis

We recorded the number of questions answered correctly by each AI platform per exam using the ASPS PSITE answer key. These data were then analyzed using analysis of variance and covariance tests with Stata 15.1 (Stata Corp, 2017, Stata Statistical Software: Release 15; Stata Corp LLC, College Station, Tex.). Analyses were also run using ChatGPT-4.0 to gauge its computational ability. A P value of less than 0.05 was considered statistically significant.

Additionally, we conducted an in-depth analysis of responses produced by ChatGPT-4.0 on the 2023 PSITE. Each question was classified into one of the following categories: anatomy, pathophysiology, clinical recall, or clinical reasoning. For questions answered incorrectly, the responses were evaluated to identify the root cause of the inaccuracy and categorized as follows:

1. Factual inaccuracy: the answer justification contained false scientific information.
2. Use of outdated information: the answer was based off of outdated guidelines, publications, or scientific beliefs.
3. Improper consideration of the clinical vignette: the answer justification failed to incorporate a crucial piece of information from the question stem, leading to the wrong answer choice; however, the information provided was factually sound.
4. Logical fallacy: the correct answer was explained in the reasoning but the chatbot failed to select the correct multiple-choice option.

RESULTS

After excluding questions reliant on tables or images, the number of usable PSITE questions per exam varied by year: 238 (2019), 239 (2020), 228 (2021), 234 (2022), and 236 (2023). The percentage of correctly answered questions was calculated for each chatbot and examination year (Table 1). Across all examination years, the highest accuracy was 79% on the 2020 PSITE by ChatGPT-4.0, whereas the lowest score, 45%, was obtained by Google PaLM on the 2019 examination. Average AI performance

Table 1. Scores of AI Chatbots on the 2019–2023 PSITE ($P < 0.01$)

	ChatGPT 3.5, %	Google Bard, %	Microsoft Bing AI, %	My AI by Snapchat, %	ChatGPT 4.0, %	Google PaLM, %	Claude+, %
2019	46	57	51	47	68	45	50
2020	57	62	56	64	79	49	54
2021	52	55	55	58	73	50	60
2022	49	51	54	55	68	48	54
2023	59	60	63	57	71	51	57

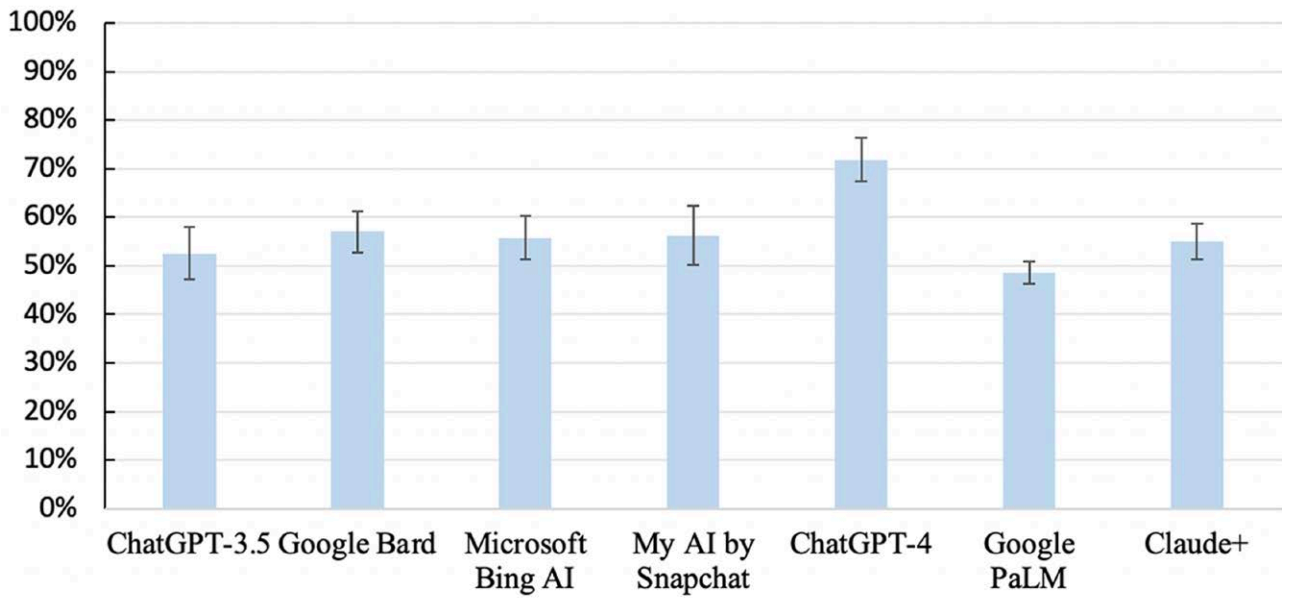

Fig. 1. Average AI chatbot performance on the 2019–2023 PSITE.

on the PSITE was calculated using individual scores from the 2019–2023 examinations. Of the AI platforms investigated, ChatGPT-4.0 performed the strongest, obtaining an average score of 71.8 ± 4.5%. Average PSITE performance was comparable between the other platforms: ChatGPT-3.5 (52.6 ± 5.4%), Google Bard (57.0 ± 4.3%), Microsoft Bing AI (55.8 ± 4.4%), My AI by Snapchat (56.2 ± 6.1%), Google PaLM (48.6 ± 2.3%), and Claude (55.0 ± 3.7%) (Fig. 1). ANOVA multifactor analysis using overall raw scores for all five examination years revealed that performance differences among the AI platforms was statistically significant ($P < 0.01$).

Two versions of the ChatGPT platform from OpenAI were utilized in this study. The newer model, ChatGPT-4.0, performed better than its older counterpart to a statistically significant degree ($P < 0.01$). The same is true when comparing the two chatbots created by Google; Google Bard outperformed Google PaLM ($P = 0.01$).

Average scores for the five subsections of the PSITE were calculated for each chatbot. All platforms reached their highest average on the core surgical principles section (ChatGPT-3.5 58.8%, Google Bard 61.2%, Microsoft Bing AI 66.2%, My AI by Snapchat 59%, ChatGPT-4.0 81.2%, Google PaLM 53.6%, and Claude 63.2%). Statistical significance between sections was found for all platforms except Google Bard and My AI by Snapchat ($P = 0.20$ and $P = 0.08$, respectively). For LLMs whose training database ended in 2021 (ChatGPT-3.5, ChatGPT-4.0, and Claude), no statistical significance was noted when comparing examination scores before and after the training cutoff ($P = 0.70$, $P = 0.43$, and $P = 0.85$, respectively).

Further investigation into ChatGPT-4.0's performance on the 2023 PSITE demonstrated that the platform performed substantially worse on anatomy questions compared with other question types; it correctly answered 77% of clinical reasoning questions, 72% of clinical recall questions, 72% of pathophysiology questions, and 42% of anatomy questions. Analysis of answer justifications revealed that 79% of incorrect answers can be attributed to factual inaccuracy, 9% to logical fallacy, 6% to improper consideration of the clinical vignette, and 6% to the use of outdated information. Due to copyright restrictions, we are unable to provide specific answer examples.

Statistical analyses were also input into ChatGPT-4.0 to examine its computational ability. Upon entering the data and asking the chatbot to conduct an ANOVA analysis, ChatGPT-4.0 informed the user that it is unable to perform complex mathematical equations and instructed the user to use Python. However, with various prompt manipulation, we were able to elicit results from ChatGPT-4.0. Of the 13 ANOVA tests conducted, ChatGPT-4.0's determination of statistical significance aligned with that of Stata 15.1 69% of the time, although the chatbot could not provide a numerical P value (Table 2).

DISCUSSION

Proliferation of AI Chatbots

AI has captured the attention of the medical community for its potential to drastically transform the future of healthcare. Technology that once seemed like a distant possibility is now a reality, and if utilized properly, presents endless opportunities to benefit physicians and patients alike. Chatbots, or LLMs, are a form of AI designed to imitate human conversation; these platforms can process substantial amounts of data and use pattern recognition to answer questions or respond to prompts, rapidly accomplishing these tasks with little to no cost.[1,2] As the first chatbot of its kind, ChatGPT-3.5 has dominated much of the medical literature. However, since its release in 2022, various other AI chatbots have been developed and are rising in popularity, their potential remaining undiscovered.

Table 2. Statistical Analyses Conducted on Stata 15.1 and ChatGPT-4.0

	Statistically Significant Result? ($P < 0.05$)	
ANOVA Analysis	**Stata 15.1**	**ChatGPT 4.0**
PSITE performance between the different AI platforms	Yes	Yes
PSITE performance between ChatGPT 3.5 and ChatGPT 4.0	Yes	Yes
PSITE performance between Google PaLM and Google Bard	Yes	Yes
PSITE performance between subsections by ChatGPT 3.5	Yes	No
PSITE performance between subsections by Google Bard	No	Yes
PSITE performance between subsections by Microsoft Bing AI	Yes	Yes
PSITE performance between subsections by My AI by Snapchat	No	Yes
PSITE performance between subsections by ChatGPT 4.0	Yes	No
PSITE performance between subsections by Google PaLM	Yes	Yes
PSITE performance between subsections by Claude+	Yes	Yes
PSITE performance between pre-2021 and post-2021 examinations by ChatGPT 3.5	No	No
PSITE performance between pre-2021 and post-2021 examinations by ChatGPT 4.0	No	No
PSITE performance between pre-2021 and post-2021 examinations by Claude+	No	No

Of the 13 ANOVA tests conducted, the platforms yielded the same determination of statistical significance 69% of the time.

Table 3. AI Performance on the 2023 PSITE Compared with Residents in Integrated Programs

		Percentile Ranking Compared with Residents					
	Exam Score (% Correct)	**First Year**	**Second Year**	**Third Year**	**Fourth Year**	**Fifth Year**	**Sixth Year**
ChatGPT-4	71	95th	82nd	57th	40th	34th	12th
Microsoft Bing AI	63	63rd	33rd	12th	11th	2nd	1st
Google Bard	60	40th	23rd	6th	4th	1st	1st
ChatGPT-3.5	59	32nd	17th	6th	2nd	1st	1st
My AI by Snapchat	57	25th	10th	3rd	1st	1st	0th
Claude+	57	25th	10th	3rd	1st	1st	0th
Google PaLM	51	10th	3rd	0th	0th	0th	0th

Adapted from American Society of Plastic Surgeons.[16]

Among the new AI chatbots are ChatGPT-4.0, Microsoft Bing AI, My AI by Snapchat, Claude, Google PaLM, and Google Bard. ChatGPT-4.0 is the newer model of ChatGPT-3.5 that is operated by OpenAI (San Francisco, Calif.). According to the company, the new model is slated to be "more reliable, creative, and able to handle much more nuanced instructions than GPT-3.5."[3] Interestingly, OpenAI's technology was utilized in the creation of Microsoft Bing AI and My AI by Snapchat, with additional modifications made by the companies.[12,13] Claude is an independent AI model created and operated by Anthropic (San Francisco, Calif.) and both Google chatbots, PaLM and Bard, are run by Google (Mountain View, Calif.).[4,5,14,15]

Performance of Chatbots on the PSITE

Results of our study indicate that ChatGPT-4.0 has potential as an educational tool in the field of plastic surgery; however, the use of other chatbots in plastic surgery education should be cautioned against unless they are improved. Of the chatbots tested, ChatGPT-4.0 outperformed its competitors, reaching scores up to 79% on the PSITE. The performance among other chatbots was comparable, with average PSITE score (2019–2023) ranging from 48.6% to 57.0% (Fig. 1).

Although there is no passing score for the PSITE, chatbot performance can be compared with human performance using ASPS resident norms (Table 3). On the 2023 PSITE, ChatGPT-4.0 performed remarkably well compared with first-year residents, ranking in the 95th percentile. However, the chatbot's percentile ranking subsequently declined when compared with residents further along in their education, scoring in the 82nd percentile of second year, 57th percentile of third year, 40th percentile of fourth year, 34th percentile of fifth year, and 12th percentile of sixth year.[16] The performance of the other chatbots was unimpressive. Google PaLM fared the worst compared with residents, ranking in only the 10th percentile of first year, the third percentile of second year, and the 0th percentile for third through sixth years.[16] Microsoft Bing and Google Bard performed moderately on the 2023 PSITE compared with first-year residents (63rd and 40th percentile, respectively), whereas ChatGPT-3.5, My AI by Snapchat, and Claude had poorer outcomes (32nd, 25th, 25th, respectively); none of these platforms surpassed the second percentile when compared with fifth- and sixth-year residents.[16] Percentile ranking for all platforms worsened when compared with upper-level residents, suggesting that the use of AI chatbots as an educational resource is most beneficial to physicians early in their career. Of the chatbots analyzed, ChatGPT-4.0 would be the most suitable platform for this purpose, as it is the only chatbot whose performance ranked in the top quartile of residents.

Further analysis into ChatGPT-4.0's performance on the 2023 PSITE revealed that the platform struggled most with anatomy questions (42% correct) compared with clinical recall (72%), pathophysiology (72%), and clinical reasoning (77%) questions. It is possible that this finding is related to the chatbot's inability to work with image-based data, as anatomy is largely a visual-dependent field. Although plastic surgery residents should consult more traditional resources for anatomical questions, ChatGPT-4.0's performance on the clinical and pathophysiologic questions is promising.

Analysis of ChatGPT-4.0's justifications for incorrect answers on the 2023 PSITE revealed that the majority of errors were caused by factual inaccuracy (79%) rather than logical fallacy (9%), the use of outdated information (6%), or improper consideration of the clinical vignette (6%). This indicates that the program possesses the ability to think through complex problems and answer correctly, given it has access to the proper knowledge base. The low percentage of questions missed due to outdated information is particularly interesting, given that ChatGPT-4.0 only contains a knowledge base through 2021, whereas other platforms can harness up-to-date information through web searches.[3,17] The lack of post-2021 training does not seem to inhibit the platform's potential in plastic surgery education, especially considering that changes in the field often take years to develop.

Finally, we wanted to assess whether ChatGPT-4.0 outperformed its predecessor, ChatGPT-3.5; OpenAI grants free access to ChatGPT-3.5, but charges $20 per month for ChatGPT-4.0. Analysis revealed that ChatGPT-4.0 performed drastically better than ChatGPT-3.5 on the PSITE, which is consistent with previous studies comparing their performance on neurosurgery and general surgery board examinations.[18,19]

Utilization of AI Chatbots in Plastic Surgery Education

Although ChatGPT-4.0 has demonstrated its potential as an educational resource in plastic surgery based on its PSITE performance, there remains room for improvement, especially for the other chatbots investigated in this study. Considering many of these chatbots were only released within the last few years and newer models have already been developed, there is significant hope that the technology will progress at an accelerated pace in upcoming years. Although widespread implementation has not yet occurred, there are numerous ways in which AI chatbots can be incorporated into plastic surgery education:

1. Study resource for plastic surgery residents: Many have suggested that chatbots could assume the role of a "personal tutor," answering questions, providing feedback, and creating novel study resources for physicians in training (practice questions, clinical vignettes, etc).[1,20,21] Furthermore, chatbots can provide these services at little to no cost; the platforms analyzed in this study are available free-of-charge, except ChatGPT-4.0 ($20/month).
2. Research tool: Chatbots will likely become a valuable research tool for plastic surgery residents because their ability to process extensive amounts of data allows for unique pattern recognition and generation of innovative research questions.[22–24] Additionally, they have potential to assist in the literature review and rough draft writing process which, some have suggested, could afford physicians more time to focus on the clinical significance of their research projects, rather than on the more arduous tasks.[25–27]
3. Patient medical resource: Chatbots can serve as an additional outlet for patient questions. Multiple studies have illustrated ChatGPT's ability to answer common patient questions, ranging from drug–drug interactions to inquiries regarding the risks, benefits, and expectations of plastic surgery procedures.[28–34] This has potential to improve patient education, shorten consult times, and increase patient satisfaction with their physician.
4. Clinical applications: Although still in the early stages, some researchers have attempted harnessing AI in the clinical realm, using the technology to assist with diagnostics, risk stratification, and surgical planning.[35–37] Additionally, there is hope that AI could alleviate the time-consuming administrative work often tasked to residents (operative notes, data collection, and discharge summaries), allowing them more time to focus on direct patient care.[38,39] Finally, AI has been tested as an objective assessment tool for surgical outcomes and technique, which may have potential to evaluate resident performance in a less biased manner.[40–42]

Limitations of AI Chatbot Use in Plastic Surgery Education

Although chatbots have significant potential within plastic surgery, it is paramount that their limitations are recognized to ensure responsible usage:

1. Restricted usefulness for statistical analysis: Despite their ability to assist in research question formulation and study design, chatbots may be confined in their ability to conduct quantitative analysis. In this study, we ran statistical analyses through ChatGPT-4.0 after conducting the tests using Stata 15.1. ChatGPT-4.0 incorrectly determined statistical significance 31% of the time and was unable to provide specific *P* values. This limitation is recognized by the chatbot itself, which informs users that it is not adequately trained to perform complex mathematical equations and instead provides instructions on how to utilize other programming platforms, such as Python. Although proper manipulation of prompts input into ChatGPT-4.0 can ultimately yield statistical analyses, it is not able to replace an experienced statistician, and plastic surgery residents would likely obtain better results through mathematical programs like Python or Stata.
2. Fabricated references: Perhaps the most concerning drawback of AI is its well-documented tendency of fabricating references for the information it generates, providing references unrelated to the topic of

discussion, and incorrectly citing references that do exist.[43–46] This is a significant limitation for those wanting to use AI in research or trace back information provided by chatbots to their original source.

3. Information bias: Chatbots' ability to recognize patterns and make predictions is based on their data set training; if these data were biased, AI responses will be as well.[25–27,47] If AI chatbots are implemented in diagnostics or treatment planning, there is the possibility of patient endangerment if the chatbot is reliant on biased or out-of-date information.
4. Information inaccuracies: As witnessed by the moderate to poor performance of chatbots in this study, AI chatbots are not infallible. Although able to generate responses and explain their reasoning, the information provided may be inaccurate.[47,48] Even the strongest performing platform, ChatGPT-4.0, had an average PSITE score of 71.8 ± 4.5% and of incorrect responses on the 2023 PSITE, 79% were attributed to factual inaccuracy. Even though results of the present study seem to suggest newer chatbots are improving in their information accuracy, those wishing to utilize AI in plastic surgery should proceed cautiously, utilizing more traditional resources to fact-check and serve as ultimate authority. Revision to AI chatbots is needed to avoid the distribution of false information that can mislead trainees and patients alike.
5. Risk of plagiarism: Finally, AI chatbots present the risk of plagiarism. If utilized in the composition of research articles, practice questions, or literature reviews, the material must be subject to further human review to avoid infringing on others' intellectual property.[48] Furthermore, ChatGPT is unable to fulfill authorship criteria, making its use in drafting materials for publication ethically ambiguous and a topic of debate that will likely escalate as its use becomes more widespread.[49]

CONCLUSIONS

AI chatbots have the potential to revolutionize plastic surgery education. Although most chatbots lack proficiency in plastic surgery, the performance of ChatGPT-4.0 was encouraging for those wishing to harness AI as an educational resource in the field; the chatbot ranked in the 95th percentile of first-year residents on the 2023 PSITE and had accuracy rates up to 79%. Analysis revealed that the utility of AI as an educational resource seems greatest at the onset of residency training, and despite its higher price, ChatGPT-4.0 should be utilized over ChatGPT-3.5 due to its significantly enhanced proficiency in plastic surgery. Although ChatGPT-4.0 currently demonstrates promise as an educational tool, further refinements must be made before its use becomes widely implemented.

Nicole DiDonna, BA
University of North Carolina School of Medicine
321 S. Columbia Street
Chapel Hill, NC 27599
E-mail: nicole_didonna@med.unc.edu

DISCLOSURE

The authors have no financial interest to declare in relation to the content of this article.

REFERENCES

1. Bassiri-Tehrani B, Cress PE. Unleashing the power of ChatGPT: revolutionizing plastic surgery and beyond. *Aesthet Surg J.* 2023;43:1395–1399.
2. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med Educ.* 2023;9:e48291.
3. OpenAI. GPT-4. Available at https://openai.com/research/gpt-4. Accessed July 16, 2023.
4. Narang S, Chowdhery A. Pathways Language Model (PaLM): scaling to 540 billion parameters for breakthrough performance. Available at https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html. Accessed July 16, 2023.
5. Google AI. AI across Google: PaLM 2. Available at https://ai.google/discover/palm2/. Accessed July 16, 2023.
6. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ.* 2023;9:e45312.
7. 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.
8. 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.
9. 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 examination. *Aesthet Surg J.* 2023;43:NP1085–NP1089.
10. American Society of Plastic Surgeons. Administrative information. Available at https://www.plasticsurgery.org/for-medical-professionals/education/events/in-service-exam-for-residents/administrative-information. Accessed July 16, 2023.
11. 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.
12. Mehdi Y. Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. Available at https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/. Published February 7, 2023. Accessed July 16, 2023.
13. Snapchat. What is my AI on Snapchat and how do I use it?. Available at https://help.snapchat.com/hc/en-us/articles/13266788358932-What-is-My-AI-on-Snapchat-and-how-do-I-use-it-. Accessed July 16, 2023.
14. Pichai S. An important next step on our AI journey. Available at https://blog.google/technology/ai/bard-google-ai-search-updates/. Accessed July 16, 2023.
15. Anthropic. Meet Claude. Available at https://www.anthropic.com/product. Accessed July 16, 2023.
16. American Society of Plastic Surgeons. ASPS in-service self-assessment examination for residents computation and interpretation of test scores. ASPS: 2023.
17. MLQ. Claude. Available at https://www.mlq.ai/tools/claude/. Accessed July 16, 2023.
18. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral board preparation question bank. *Neurosurgery.* 2023;93:1090–1098.
19. Oh N, Choi G, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical

education and training in the era of large language models. *Ann Surg Treat Res.* 2023;104:269–273.

20. Lee H. The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ.* 2023 [E-pub ahead of print].
21. Koljonen V. What could we make of AI in plastic surgery education. *J Plast Reconstr Aesthet Surg.* 2023;81:94–96.
22. Liang X, Yang X, Yin S, et al. Artificial intelligence in plastic surgery: applications and challenges. *Aesthetic Plast Surg.* 2021;45:784–790.
23. Gupta R, Herzog I, Weisberger J, et al. Utilization of ChatGPT for plastic surgery research: friend of foe? *J Plast Reconstr Aesthet Surg.* 2023;80:145–147.
24. Gupta R, Park JB, Bisht C, et al. Expanding cosmetic plastic surgery research with ChatGPT. *Aesthet Surg J.* 2023;43:930–937.
25. ElHawary H, Gorgy A, Janis JE. Large language models in academic plastic surgery: the way forward. *Plast Reconstr Surg Glob Open.* 2023;11:e4949.
26. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell.* 2023;6:1169595.
27. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel).* 2023;11:887.
28. Seth I, Cox A, Xie Y, et al. Evaluating chatbot efficacy for answering frequently asked questions in plastic surgery: a ChatGPT case study focused on breast augmentation. *Aesthet Surg J.* 2023;43:1126–1135.
29. Juhi A, Pipil N, Santra S, et al. The capability of ChatGPT in predicting and explaining common drug-drug interactions. *Cureus.* 2023;15:e36272.
30. Xie Y, Seth I, Hunter-Smith DJ, et al. Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. *Aesthetic Plast Surg.* 2023;47:1985–1993.
31. Boczar D, Sisti A, Oliver JD, et al. Artificial intelligent virtual assistant for plastic surgery patient's frequently asked questions: a pilot study. *Ann Plast Surg.* 2020;84:e16–e21.
32. Avila FR, Boczar D, Spaulding AC, et al. High satisfaction with virtual assistant for plastic surgery frequently asked questions. *Aesthet Surg J.* 2023;43:494–503.
33. Eldaly AS, Avila FR, Torres-Guzman RA, et al. Simulation and artificial intelligence in rhinoplasty: a systematic review. *Aesthetic Plast Surg.* 2022;46:2368–2377.
34. Knoedler L, Odenthan J, Prantl L, et al. Artificial intelligence-enabled simulation of gluteal augmentation: a helpful tool in preoperative outcome simulation? *J Plast Reconstr Aesthet Surg.* 2023;80:94–101.
35. Sayadi LR, Hamdan US, Zhangli Q, et al. Harnessing the power of artificial intelligence to teach cleft lip surgery. *Plast Reconstr Surg Glob Open.* 2022;10:e4451.
36. Moura FSE, Amin K, Ekwobi C. Artificial intelligence in the management and treatment of burns: a systematic review. *Burns Trauma.* 2021;9:tkab022.
37. Mirnezami R, Ahmed A. Surgery 3.0, artificial intelligence and the next-generation surgeon. *Br J Surg.* 2018;105:463–465.
38. Singh S, Djalilian A, Ali MJ. ChatGPT and ophthalmology: exploring its potential with discharge summaries and operative notes. *Semin Ophthalmol.* 2023;38:503–507.
39. Taritsa IC, Sandepudi K, Williams T, et al. Visualizations in plastic surgery: open-source artificial intelligence can accelerate reconstructive operative techniques and reports. *Plast Reconstr Surg.* 2024;153:225e–226e.
40. Zhang BH, Chen K, Lu SM, et al. Turning back the clock: artificial intelligence recognition of age reduction after face-lift surgery correlates with patient satisfaction. *Plast Reconstr Surg.* 2021;148:45–54.
41. Boonipat T, Hebel N, Zhu A, et al. Using artificial intelligence to analyze emotion and facial action units following facial rejuvenation. *J Plast Reconstr Aesthet Surg.* 2022;75:3628–3651.
42. Elliott ZT, Bheemreddy A, Fiorella M, et al. Artificial intelligence for objectively measuring years regained after facial rejuvenation surgery. *Am J Otolaryngol.* 2023;44:103775.
43. Bhattacharyya M, Miller VM, Bhattacharyya D, et al. High rates of fabricated and inaccurate referenes in ChatGPT-generated medical content. *Cureus.* 2023;15:e39238.
44. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Med Educ.* 2023;9:e46885.
45. Wagner MW, Ertl-Wagner BB. Accuracy of information and references using chatgpt-3 for retrieval of clinical radiological information. *Can Assoc Radiol J.* 2023; 75:69–73.
46. Athaluri SA, Manthena SV, Kesapragada VSRKM, et al. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus.* 2023;15:e37432.
47. Weidman AA, Valentine L, Chung KC, et al. OpenAI's ChatGPT and its role in plastic surgery research. *Plast Reconstr Surg.* 2023;151:1111–1113.
48. Van de Ridder JMM, Shoja MM, Rajput V. Finding the place of ChatGPT in medical education. *Acad Med.* 2023;98:867.
49. Najafali D, Reiche E, Camacho JM, et al. Let's chat about chatbots: additional thoughts on ChatGPT and its role in plastic surgery along with its ability to perform systematic reviews. *Aesthet Surg J.* 2023;43:NP591–NP592.