OPEN

PRS GLOBAL OPEN

ORIGINAL ARTICLE

Education

Class in Session: Analysis of GPT-4-created Plastic Surgery In-service Examination Questions

Daniel Najafali, BS*
Logan G. Galbraith, BA†
Justin M. Camacho, MBA‡
Victoria Stoffel, MS‡
Isabel Herzog, BA§
Civanni Moss, BSN, RN¶
Stephanie L. Taiberg, BS||
Leonard Knoedler, MD**††

Background: The Plastic Surgery In-Service Training Examination (PSITE) remains a critical milestone in residency training. Successful preparation requires extensive studying during an individual's residency. This study focuses on the capacity of Generative Pre-trained Transformer 4 (GPT-4) to generate PSITE practice questions.

Methods: GPT-4 was prompted to generate multiple choice questions for each PSITE section and provide answer choices with detailed rationale. Question composition via readability metrics were analyzed, along with quality. Descriptive statistics compared GPT-4 and the 2022 PSITE.

Results: The overall median Flesch–Kincaid reading ease for GPT-4-generated questions was 43.90 (versus 50.35 PSITE, $P = 0.036$). GPT-4 provided questions that contained significantly fewer mean sentences (1 versus 4), words (16 versus 56), and percentage of complex words (3 versus 13) than 2022 PSITE questions ($P < 0.001$). When evaluating GPT-4 generated questions for each examination section, the highest median Flesch–Kincaid reading ease was on the core surgical principles section (median: 63.30, interquartile range [54.45–68.28]) and the lowest was on the craniomaxillofacial section (median: 36.25, interquartile range [12.57–58.40]). Most readability metrics were higher for the 2022 PSITE compared with GPT-4 generated questions. Overall question quality was poor for the chatbot.

Conclusions: Our study found that GPT-4 can be adapted to generate practice questions for the 2022 PSITE, but its questions are of poor quality. The program can offer general explanations for both the correct and incorrect answer options but was observed to generate false information and poor-quality explanations. Although trainees should navigate with caution as the technology develops, GPT-4 has the potential to serve as an effective educational adjunct under the supervision of trained plastic surgeons. *(Plast Reconstr Surg Glob Open 2024; 12:e6185; doi: 10.1097/GOX.0000000000006185; Published online 19 September 2024.)*

*From the *Carle Illinois College of Medicine, University of Illinois Urbana-Champaign, Urbana, Ill.; †Northeast Ohio Medical University College of Medicine, Rootstown, Ohio; ‡Drexel University College of Medicine, Philadelphia, Pa.; §Rutgers New Jersey School of Medicine, Newark, N.J.; ¶Temple University Lewis Katz School of Medicine, Philadelphia, Pa.; ||Chicago Medical School at Rosalind Franklin University of Medicine and Science, Chicago, Ill.; **Department of Plastic, Hand and Reconstructive Surgery, University Hospital Regensburg, Regensburg, Germany; and ††Division of Plastic and Reconstructive Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, Mass.*

Received for publication March 28, 2024; accepted July 24, 2024.

Najafali and Galbraith contributed equally to this work.

DOI: 10.1097/GOX.0000000000006185

INTRODUCTION

The expansion of the capabilities of Chat Generative Pre-trained Transformer (ChatGPT) into the domain of medical education has been demonstrated by Kung et al, who were among the first to show that ChatGPT can pass the United States Medical Licensing Examination at all three levels.[1] In plastic surgery, ChatGPT has demonstrated that it can generate novel ideas for plastic surgery applications in conjunction with proficiency with respect to plastic surgery knowledge.[2–4] It achieved 60.1% on the 2021 Plastic Surgery In-Service Training Examination (PSITE). On the 2022 PSITE, it was in the 49th percentile compared with first-year integrated plastic surgery residents, but it performed in the 0th percentile compared with PGY-5 and PGY-6 residents.[5] The chatbot's ease of use has made some apply it to medical education with implications on possible clinical decision-making in the future if substantially improved.[1,6]

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

The PSITE evaluates residents over five subsections, which include comprehensive, hand and lower extremity surgery, craniomaxillofacial surgery, breast and cosmetic, and core surgical principles. With the plethora of resources available, mastery becomes not only challenging, but daunting. The material contained within these resources are vast, and parsing through this dense material remains challenging even for the progressing resident. Therefore, chatbots may become a resident's newest ally by allowing residents to develop individualized preparation material for their PSITE. However, the viability and efficacy of chatbots need to be assessed to determine their quality. Utilization of chatbots could serve as a double-edged sword. On the one hand, utilization has the potential to save residents time but may also introduce false or partially false information into their knowledge set. This is not a worry with higher quality and vetted resources. Still, the software can directly benefit residents, as well as save the test writers time by creating examination questions that use the current literature and resource landscape while further improving the quality of the questions. This study explores the ability of Generative Pre-trained Transformer 4 (GPT-4) to generate practice questions for the PSITE by assessing its readability in comparison to a recent set of questions written by the examination committee. Efforts to screen the examination for syntax may assist in overall literacy and question quality.

Takeaways

Question: What is the quality of Generative Pre-trained Transformer–generated in-service examination questions?

Findings: Generative Pre-trained Transformer can generate questions and answer explanations that provide rationale but they are of poor quality.

Meaning: Trainees should navigate chatbots with caution, which may be more effective in answering questions rather than in generating high-quality standardized multiple choice questions that are evidence-based in nature for the in-service examination.

METHODS

Chatbot

The chatbot used in this study was GPT-4 (Open AI, San Francisco, Calif.). The chatbot was prompted with the following: "Create 10 multiple choice questions for the Plastic Surgery In-Service Examination for the [relevant section] section." All five sections of the examination were used for the prompting. Answers with explanations and references were generated using the following prompt: "Generate the answers with explanations and references." These were evaluated for quality and usefulness from the lens of trainees.

Plastic Surgery In-Service Training Examination

The American Society of Plastic Surgeons PSITE from 2022 was used because it was the most recent examination. To compare the generated questions from the chatbot to questions written by test writers, the first 10 questions were used in the readability platform from the five sections: comprehensive (section 1), hand and lower extremity surgery (section 2), craniomaxillofacial surgery (section 3), breast and cosmetic (section 4), and core surgical principles (section 5). This study received approval from the American Society of Plastic Surgeons education leadership.

Assessing Quality of Generated Chatbot Content

An open access readability assessment tool was used via WebFX that has also been used in other studies for the same purpose.[7–9] The enter text feature was selected, and either the generated GPT-4 questions or the 2022 PSITE questions were copied verbatim into the platform. Readability results consisted of the Flesch–Kincaid reading ease, Flesch–Kincaid reading grade level, Gunning Fog score, Simple Measure of Gobbledygook (SMOG) index, Coleman–Liau index, and automated readability index. These readability metrics and others have been used in other plastic surgery studies.[10–19] Text statistics included sentences, words, complex words, percentage of complex words, average words per sentence, and average syllables per word.

Statistical Analysis

A standardized Microsoft Excel spreadsheet (Microsoft Corporation, Redmond, Wash.) was used to collect all data with relevant fields. Descriptive statistics were used to report mean (± SD) or median (interquartile range [IQR]). Continuous variables were compared for significance using a Student *t* test or a Mann–Whitney U test as appropriate.

All statistical analysis was performed in the R (version 4.1.0) software in the RStudio (version 1.4.1717) environment. Statistical significance was set at a threshold of a two-sided P value of less than 0.05.

RESULTS

Readability and Test Metrics: GPT-4 PSITE Questions and 2022 PSITE Questions

Table 1 summarizes the readability and text metrics for the 50 questions that were generated by GPT-4 and the 50 questions that were used from the 2022 PSITE examination. GPT-4 had a significantly higher median Flesch–Kincaid reading ease in comparison with the 2022 PSITE (median: 50.35, IQR [36.25–63.10] GPT-4 versus 43.90 [34.50–50.23] 2022 PSITE, $P = 0.0360$). The Flesch–Kincaid grade level corresponded to a lower education level for GPT-4 when compared with the 2022 PSITE (median: 9.95, IQR [7.73–12.23] GPT-4 versus median: 10.75, IQR [9.80–12.60] 2022 PSITE, $P = 0.08$). The other readability metrics were statistically similar ($P > 0.05$) between GPT-4 and the 2022 PSITE. When examining the difference between text-based metrics for the questions, GPT-4 had a significantly lower mean sentence total ($P < 0.001$), word count ($P < 0.001$),

Table 1. GPT-4 and 2022 PSITE Readability and Text Metrics Comparison

Metrics	GPT-4	2022 PSITE	*P**
Readability, median [IQR]			
Flesch–Kincaid reading ease	50.35 [36.25–63.10]	43.90 [34.50–50.23]	**0.036**
Flesch–Kincaid grade level	9.95 [7.73–12.23]	10.75 [9.80–12.60]	0.08
Gunning Fog score	13.90 [11.40–16.85]	14.80 [13.15–16.62]	0.16
SMOG index	10.10 [8.30–11.60]	10.35 [9.40–11.60]	0.13
Coleman–Liau index	13.75 [10.95–15.80]	14.05 [12.90–17.10]	0.15
Automated readability index	8.95 [7.23–11.75]	10.55 [8.40–12.40]	0.12
Text, mean (SD)			
Sentences	1.08 (0.27)	3.84 (1.89)	**<0.001**
Words	15.66 (5.15)	55.58 (25.47)	**<0.001**
Complex words	3.22 (1.67)	12.96 (7.50)	**<0.001**
Percent of complex words	21.19 (10.85)	22.82 (6.81)	0.37
Average words per sentence	14.63 (3.86)	15.09 (3.12)	0.52
Average syllables per word	1.71 (0.27)	1.78 (0.17)	0.16

*Bold cells indicate statistically significant findings ($P < 0.05$).

and complex word usage ($P < 0.001$) compared with the 2022 PSITE.

Readability and Test Metrics Stratified by Examination Section for GPT-4 PSITE Questions and 2022 PSITE Questions

The five sections were stratified for GPT-4 and the 2022 PSITE reported in Table 2 and Table 3, respectively. GPT-4 was found to have its highest median Flesch–Kincaid reading ease for the core surgical principles section (median: 63.30, IQR [54.45–68.28]) and its lowest for the craniomaxillofacial selection (median: 36.25, IQR [12.57–58.40]). Most words were used for the comprehensive section (mean = 19.60), and the highest percentage of complex words were used in the comprehensive section (mean = 4.40). The 2022 PSITE highest median Flesch–Kincaid reading ease was for the breast and cosmetic section (49.60, IQR [40.73–55.27]) and its lowest was the core surgical principles section (median: 38.60, IQR [32.25–48.23]). The most words were used for the core surgical principles section (mean = 68.60) and the highest percentage of complex words used were in the comprehensive section (mean = 16.40). The consistency of the metrics for readability was greater amongst the five examination sections for the 2022 PSITE, whereas outliers existed for GPT-4 generated questions.

DISCUSSION

Plastic and reconstructive surgery is known to be an innovative field always open to adapting to new technological advancements. GPT-4 is rapidly gaining attention, and the authors believe it has exciting potential applications in plastic and reconstructive surgery if used appropriately with relevant upgrades. Although chatbots are relatively new, researchers have started to determine how this technology will affect the classical educational learning model. Kahf et al[20] conducted a study that elucidated that students who used chatbot technology had a higher success rate compared with those who did not use chatbots in their studies. Furthermore, not only did the study highlight effective learning, but students also demonstrated a positive response to this teaching tool, asking for more pedagogical comments and effective ways to integrate it into their study routines. Similar trends have been observed across various disciplines of education, including accounting and engineering, further amplifying the true diversity this rapidly evolving technology has.[21,22] Within the field of medicine, it has been shown to assist in research as a tool to provide quick access to medical information, generate case scenarios for practice, and facilitate with language translation. Classically, it has been shown to help with documentation, decision support, and even patient communication such as scheduling appointments and managing medications.[23]

Multiple studies so far investigate the use of chatbot tools in helping students across all levels of education prepare for examinations; however, there has been a paucity of studies investigating the use of GPT-4 in developing study materials such as practice questions for higher level examinations such as the PSITE. This study sought to analyze the ability of GPT-4 to generate practice questions for the PSITE and compared its readability and text metrics to the test writers' questions of the 2022 PSITE across all five examination sections. During the analysis, several findings emerged regarding the questions generated by GPT-4. First, it was observed that these questions did not align with the question stem style or complexity typically observed in the PSITE. In addition, a significant proportion of the generated questions heavily relied on specific knowledge extracted solely from the prompting. Also, these questions lacked the required level of detail to foster exploration of broader concepts and the integration of knowledge from diverse domains within the field of medicine. Furthermore, the answers and explanations generated in both instances do not provide the reader with the same level of detail. In contrast, the PSITE uses multiple references, covering a much broader breadth of knowledge. Overall, most readability metrics were shown to be higher for the 2022 PSITE compared with GPT-4 generated questions. Our findings suggest that GPT-4 is capable of generating questions that have comparable readability metrics to the 2022 PSITE, but the quality of the questions

Table 2. GPT-4 Stratified by Examination Section Readability and Text Metrics

Metrics	Breast and Cosmetic	Comprehensive	Core Surgical Principles	Craniomaxillofacial Surgery	Hand and Lower Extremity	*P*
Readability, median [IQR]						
Flesch–Kincaid reading ease	40.00 [36.25–56.28]	46.45 [33.25–53.30]	63.30 [54.45–68.28]	36.25 [12.57–58.40]	54.15 [43.52–64.00]	0.11
Flesch–Kincaid grade level	11.30 [9.50–11.88]	10.80 [10.27–13.40]	8.45 [6.62–9.38]	11.35 [9.10–17.53]	8.75 [6.98–11.05]	0.08
Gunning Fog score	14.15 [13.90–16.00]	16.65 [14.48–18.08]	11.35 [8.95–12.22]	15.65 [11.40–21.93]	12.20 [8.85–14.58]	**0.028**
SMOG index	10.10 [10.10–10.10]	11.60 [10.10–12.90]	8.30 [6.58–8.30]	11.50 [8.30–12.90]	8.75 [6.58–11.22]	**0.025**
Coleman–Liau index	14.35 [12.70–16.85]	13.90 [11.75–16.78]	11.60 [9.40–14.05]	14.40 [11.12–18.60]	13.85 [10.60–14.97]	0.49
Automated readability index	9.80 [7.72–12.35]	11.20 [9.38–12.65]	7.45 [6.00–9.33]	9.20 [6.72–14.35]	8.45 [6.45–10.10]	0.13
Text, mean (SD)						
Sentences	1.00 (0.00)	1.20 (0.42)	1.00 (0.00)	1.00 (0.00)	1.20 (0.42)	0.17
Words	14.80 (3.61)	19.60 (5.66)	14.10 (3.11)	14.10 (4.84)	15.70 (6.52)	0.09
Complex words	3.00 (1.41)	4.40 (1.17)	2.00 (0.94)	3.90 (2.23)	2.80 (1.40)	**0.008**
Percent of complex words	20.82 (9.42)	24.38 (9.80)	14.59 (7.08)	27.90 (14.60)	18.28 (8.54)	0.05
Average words per sentence	14.80 (3.61)	17.05 (4.26)	14.10 (3.11)	14.10 (4.84)	13.10 (2.64)	0.21
Average syllables per word	1.72 (0.22)	1.72 (0.20)	1.57 (0.26)	1.91 (0.39)	1.65 (0.16)	0.06

Values in boldface indicate statistically significant findings ($P < 0.05$).

Table 3. 2022 PSITE stratified by Examination Section Readability and Text Metrics

Metrics	Breast and Cosmetic	Comprehensive	Core Surgical Principles	Craniomaxillofacial Surgery	Hand and Lower Extremity	*P*
Readability, median [IQR]						
Flesch–Kincaid reading ease	49.60 [40.73–55.27]	42.45 [26.75–55.80]	38.60 [32.25–48.23]	38.65 [32.47–44.52]	47.05 [43.05–57.95]	0.27
Flesch–Kincaid grade level	10.45 [9.18–11.90]	11.20 [8.95–13.20]	10.65 [9.85–12.73]	11.95 [11.10–13.00]	10.55 [9.05–11.47]	0.41
Gunning Fog score	14.05 [12.65–17.78]	15.50 [13.17–16.95]	14.00 [13.33–16.72]	15.30 [14.70–16.10]	14.20 [12.03–15.75]	0.59
SMOG index	10.10 [9.10–12.45]	10.80 [9.50–12.25]	9.75 [9.48–10.85]	10.70 [10.60–11.52]	10.10 [8.72–11.40]	0.69
Coleman–Liau index	13.20 [11.55–15.12]	14.95 [13.20–17.95]	13.40 [12.90–17.42]	15.70 [13.70–16.90]	14.15 [11.10–15.97]	0.36
Automated readability index	9.80 [7.35–12.43]	10.50 [8.45–12.40]	9.30 [8.33–12.85]	11.70 [10.75–13.08]	10.45 [8.88–11.13]	0.45
Text, mean (SD)						
Sentences	3.60 (1.58)	4.40 (1.84)	5.00 (2.11)	2.80 (1.32)	3.40 (2.01)	0.07
Words	56.10 (25.05)	62.30 (27.56)	68.60 (29.40)	43.90 (16.69)	47.00 (23.03)	0.16
Complex words	12.90 (7.87)	16.40 (9.72)	16.00 (7.53)	10.90 (4.95)	8.60 (4.38)	0.09
Percent of complex words	23.46 (9.33)	25.11 (6.09)	23.72 (6.31)	23.20 (4.92)	18.63 (6.10)	0.27
Average words per sentence	15.78 (2.78)	13.97 (3.28)	14.02 (3.17)	16.46 (2.94)	15.21 (3.26)	0.31
Average syllables per word	1.77 (0.24)	1.80 (0.18)	1.84 (0.16)	1.80 (0.08)	1.68 (0.16)	0.29

is heavily dependent on prompting style. Simply asking the chatbot to generate questions and explanations does not result in the quality of questions that residents may desire and can be supported by fictitious references. It is yet to be determined if updates such as the beta version of GPT-4 being able to access the internet will correct this. Another contributor to the observed quality of GPT-4 questions may be attributed to the plastic and reconstructive surgery data that was derived from the internet during the OpenAI training period for the chatbot. A considerable amount of high-quality plastic and reconstructive surgery literature are available as subscriptions and are not open access. Additionally, textbooks also have this potential limitation, as they are not freely available online. Pairing the fact that such valuable literature is not easily accessible and that prompting heavily influences the chatbot output, it may be inappropriate to conclude that the chatbot is incapable of producing questions at or exceeding the level of what is currently provided on the PSITE in terms of quality. The fundamentals of algorithms should apply broadly to plastic and reconstructive surgery and other fields; however, the availability of training data might be a challenge.

Despite the findings of this study, this tool should not be completely disregarded. Although the software is still in its early stages of development and there are existing flaws and inaccuracies, there is potential for improvement as more data is collected and analyzed. The chatbot's algorithm can be further refined to better meet the needs of the plastic and reconstructive surgery community, and plugins can be used to tailor its content. In the future one of the many benefits of using GPT-4 in medical education is possible cost and resources saved. Rather than purchasing multiple study materials, students can access a wealth of knowledge and resources through a single source, potentially minimizing traditional costs associated with textbooks. GPT-4 thus works toward equity and accessibility, which are emerging concerns within the plastic surgery community. The added benefit is that students can also individualize their learning to be more personalized. This makes it an even more valuable tool for medical education, specifically allowing for improvement of existing question banks such as those found on The American Society of Plastic Surgeons EdNet. Although our study demonstrated that PSITE questions rely on multiple references, covering a much broader scope of knowledge compared with the ones generated by the software, there is a potential for future collaboration between GPT-4 and the existing database used for question banks. This collaboration could further facilitate the incorporation of more up-to-date information and perspectives.

Ultimately, the use of chatbots has the potential to democratize medical education, but certainly has its drawbacks such as inaccuracies, ethical considerations, plagiarism, and direction of trainees toward nonevidence-based materials.[24] Although this tool shows significant promise and demonstrates adaptability to various tasks, it is important to recognize that it is still in its nascent stages. Therefore, it is possible that certain limitations of the technology may result in errors and inaccuracies that may compromise the quality of the results. Additionally, it is important to ensure that the use of chatbots does not completely replace other valuable learning resources. Rather, there should be a balance between using chatbots as an aid to education and incorporating other materials. Ultimately, the goal should be to optimize the use of chatbots as an adjunctive tool, enhancing medical education and improving student performance, rather than relying solely on them.

Limitations

This study is limited to the prompting schemes chosen to generate the questions used in the readability and text-based comparison with the 2022 PSITE. Moreover, there are limitations to the readability metrics as they were not directly intended for this use case and pangrams have been shown to generate perfect reading scores despite being obscure. Future studies should examine the answers and explanations to a further extent with subsequent updates. Potentially, the generated GPT-4 practice examination could be distributed to residents to determine their thoughts on using chatbots as well as the overall quality of the question. Additionally, cohorts could be directed to these technologies and compared with traditional learning resources to determine their true utility for plastic surgery learning. The data sources available and provided during training also influence the chatbot's outputs, an area that warrants increased attention and research.

CONCLUSIONS

Our study found that GPT-4 can be tuned to generate practice questions for the 2022 PSITE, but its quality is poor. GPT-4 also generates unreliable references and authors in its explanations. The program can offer general explanations for both the correct and incorrect answer options. With specific prompting and the expertise of trained plastic surgeons, GPT-4 may have the potential to serve as an effective educational adjunct, but trainees should navigate with caution as the technology develops.

Leonard Knoedler, MD
Department of Plastic, Hand and Reconstructive Surgery
University Hospital Regensburg
Regensburg, Germany
E-mail: lknoedler@mgh.harvard.edu

DISCLOSURE

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

ACKNOWLEDGMENT

We would like to acknowledge OpenAI GPT-4's contributions.

REFERENCES

1. 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.
2. Gupta R, Pande P, Herzog I, et al. Application of ChatGPT in cosmetic plastic surgery: ally or antagonist? *Aesthet Surg J.* 2023;43:NP587–NP590.
3. 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.
4. 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.
5. 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.
6. 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.
7. Mason AM, Compton J, Bhati S. Disabilities and the digital divide: assessing web accessibility, readability, and mobility of popular health websites. *J Health Commun.* 2021;26:667–674.
8. Dowdle TS, Nguyen JM, Steadman JN, et al. Online readability analysis: Mohs micrographic surgery postsurgical wound care. *Adv Skin Wound Care.* 2022;35:213–218.
9. WebFX Readability Test. Available at https://www.webfx.com/tools/read-able/. Accessed March 27, 2024.
10. Seth AK, Vargas CR, Chuang DJ, et al. Readability assessment of patient information about lymphedema and its treatment. *Plast Reconstr Surg.* 2016;137:287e–295e.

11. Tiourin E, Barton N, Janis JE. Health literacy in plastic surgery: a scoping review. *Plast Reconstr Surg Glob Open*. 2022;10:e4247.
12. Tran BNN, Singh M, Singhal D, et al. Readability, complexity, and suitability of online resources for mastectomy and lumpectomy. *J Surg Res*. 2017;212:214–221.
13. Vargas CR, Ricci JA, Lee M, et al. The accessibility, readability, and quality of online resources for gender affirming surgery. *J Surg Res*. 2017;217:198–206.
14. Chen AD, Ruan QZ, Bucknor A, et al. Social media: Is the message reaching the plastic surgery audience? *Plast Reconstr Surg*. 2019;144:773–781.
15. Vargas CR, Kantak NA, Chuang DJ, et al. Assessment of online patient materials for breast reconstruction. *J Surg Res*. 2015;199:280–286.
16. Fanning JE, Okamoto LA, Levine EC, et al. Content and readability of online recommendations for breast implant size selection. *Plast Reconstr Surg Glob Open*. 2023;11:e4787.
17. Ricci JA, Vargas CR, Chuang DJ, et al. Readability assessment of online patient resources for breast augmentation surgery. *Plast Reconstr Surg*. 2015;135:1573–1579.
18. Barton N, Janis JE. Missing the mark: the state of health care literacy in plastic surgery. *Plast Reconstr Surg Glob Open*. 2020;8:e2856.
19. Patel AA, Joshi C, Varghese J, et al. Do websites serve our patients well? A comparative analysis of online information on cosmetic injectables. *Plast Reconstr Surg*. 2022;149:655e–668e.
20. Al Kahf S, Roux B, Clerc S, et al. Chatbot-based serious games: a useful tool for training medical students? A randomized controlled trial. *PLoS One*. 2023;18:e0278673.
21. Naser M, Ross B, Ogle J, et al. Can AI chatbots pass the fundamentals of engineering (FE) and principles and practice of engineering (PE) structural exams? arXiv preprint arXiv:230318149. 2023.
22. Wood DA, Achhpilia MP, Adams MT, et al. The ChatGPT artificial intelligence chatbot: how well does it answer accounting assessment questions? *Issues Accounting Educ*. 2023;38:81–108.
23. Khan RA, Jawaid M, Khan AR, et al. ChatGPT—reshaping medical education and clinical management. *Pak J Med Sci*. 2023;39:605–607.
24. Flanagin A, Bibbins-Domingo K, Berkwits M, et al. Nonhuman "Authors" and implications for the integrity of scientific publication and medical knowledge. *JAMA*. 2023;329:637–639.