Article published online: 2023-08-28

Leveraging Large Language Models (LLM) for the Plastic Surgery Resident Training: Do They Have a Role?

Devi Prasad Mohapatra[1] Friji Meethale Thiruvoth[1] Satyaswarup Tripathy[2] Sheeja Rajan[3] Madhubari Vathulya[4] Palukuri Lakshmi[5] Veena K. Singh[6] Ansar Ul Haq[4]

[1] Department of Plastic Surgery, Jawaharlal Institute of Postgraduate Medical Education and Research (JIPMER), Pondicherry, India
[2] Department of Plastic Surgery, Post Graduate Institute of Medical Education and Research, Chandigarh, India
[3] Department of Plastic Surgery, Government Medical College, Thrissur, Kerala, India
[4] Department of Burns and Plastic Surgery, All India Institute of Medical Sciences (AIIMS), Rishikesh, Uttarakhand, India
[5] Department of Plastic Surgery, Osmania General Hospital, Hyderabad, Telangana, India
[6] Department of Burns and Plastic Surgery, All India Institute of Medical Sciences (AIIMS), Patna, Bihar, India

Address for correspondence Devi Prasad Mohapatra, MCh, Department of Plastic Surgery, Superspeciality Block, Jawaharlal Institute of Postgraduate Medical Education and Research (JIPMER), Pondicherry 605006, India (e-mail: devimohapatra1@gmail.com).

Indian J Plast Surg 2023;56:413–420.

Abstract

Introduction Large language models (LLMs) are designed for recognizing, summarizing, translating, predicting, and generating text-based content from knowledge gained from extensive data sets. ChatGPT4 (Generative Pre-trained Transformer 4) (OpenAI, San Francisco, California, United States) is a transformer-based LLM model pretrained on public data as well as data obtained from third-party sources using deep learning techniques of fine tuning and reinforcement learning from human feedback to predict the next text. We wanted to explore the role of LLM as a teaching assistant (TA) in plastic surgery.

Material and Methods TA roles were first identified in available literature, and based on the roles, a list of suitable tasks was created where LLM could be used to perform the task. Prompts designed to be fed in to the LLM (specifically ChatGPT) to generate appropriate output, were then created and fed to the ChatGPT model. The outputs generated were scored by evaluators and compared for interobserver agreement.

Results A final set of eight TA roles were identified where a LLM could be utilized to generate content. These contents were scored for usefulness and accuracy. These were scored independently by the eight study authors in a scoring sheet created for the study. Interobserver agreements for content accuracy, usefulness, and clarity were 100% for content generated for the following: interactive case studies (generation), simulation of preoperative consultations, and generation of ethical considerations.

Discussion LLMs in general and ChatGPT (on which this study is based) in specific, can generate answers to questions and prompts based on huge amount of text fed into the model for training the underlying language model. The answers generated have been found to be accurate, readable, and even indistinguishable from human-generated text. This capability of automated content synthesis can be exploited to generate summaries to text, answer short and long answers, and generate case scenarios. We

Keywords
- large language models (LLM)
- plastic surgical education
- educational technology
- future of surgical training
- ChatGPT in education

article published online
August 28, 2023

DOI https://doi.org/10.1055/s-0043-1772704.
ISSN 0970-0358.

Thieme Medical and Scientific Publishers Pvt. Ltd., A-12, 2nd Floor, Sector 2, Noida-201301 UP, India

could identify a few such scenarios where the LLM could in general be utilized to play the role of a TA and aid plastic surgery residents in particular. In addition, these models could also be used by students to obtain feedback and gain reflection which itself stimulates critical thinking.

Conclusion Incorporating LLMs into the educational arsenal of plastic surgery residency programs can provide a dynamic, interactive, and individualized learning experience for residents and prove to be worthy TAs of future.

Introduction

Large language models (LLMs) are designed for recognizing, summarizing, translating, predicting, and generating text-based content from knowledge gained from extensive data sets. These are deep learning models, trained on massive volumes of text data using an unsupervised learning artificial intelligence (AI) algorithm to predict and generate text based on user input[1] (►**Fig. 1**).

ChatGPT4 (Generative Pre-trained Transformer 4) (OpenAI, San Francisco, California, United States) is a transformer-based LLM model pretrained on public data as well as data obtained from third-party sources using deep learning techniques of fine tuning and reinforcement learning from human feedback to predict the next text.[2,3] Strengths of the model lie on analyzing, predicting, and generating text (known as natural language processing) in a conversational manner based on user input.[4,5] This makes it suitable for wide variety of time-consuming text-based outputs like (but not limited to) generating lesson plans, designing curricula, creating assessment materials, summarizing textbooks or articles, and providing personalized learning content. This process of "automated content synthesis" (ACS) allows users to focus on core competencies and drive innovations. Studies have been done on use of LLM in health care delivery and medical education.[6] The study objective was to evaluate the role of LLMs like ChatGPT as a teaching assistant (TA) to plastic surgery residents. A TA is someone who assists teachers by providing teaching and learning support in whatever way possible. A traditional TA has two main roles to play, support the students in their learning process and support the instructors in the effective delivery of the course content.[7]

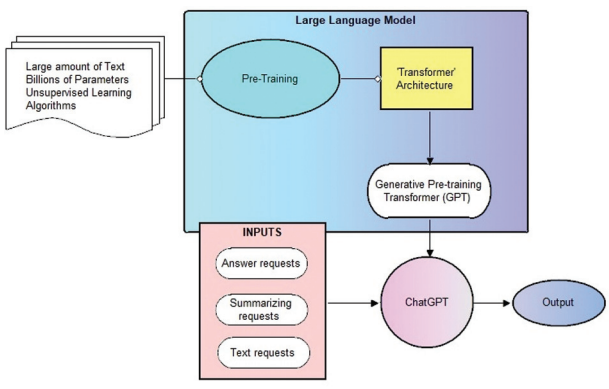

Fig. 1 Schematic representation of a large language model.

Material and Methods

The roles of a TA were first identified in available literature, and based on the roles, a list of suitable tasks was created where LLM could be used to perform the task. Prompts designed to be fed in to the LLM (specifically ChatGPT) to generate appropriate output, were then created. Each prompt was created to request a specific TA task from the LLM. The prompts were fed to ChatGPT and outputs generated were analyzed for content accuracy and usefulness. The outputs were evaluated by a two set of evaluators. The first set composed of the evaluators who were aware of the use of LLM to generate the outputs and the second set of evaluators blinded from the source of content. The scores of all observers were compared for interobserver variability and agreement using kappa statistics.[8]

Results

As a first step of this project a set of TA roles were identified and from among these, roles suitable for LLM were shortlisted[7] (►**Table 1**). The authors could identify only three tasks suitable for LLM, that is, assisting faculty with classroom instruction, records, and assignments, grading assignments or papers, and providing feedback on assignments. As can be seen, all the identified roles were oriented toward supporting the course instructor and none were directed toward assisting the residents.

To identify if there were any roles a LLM model could play in supporting residents as well as any other teaching-related tasks in plastic surgery, a prompt was fed into the ChatGPT dialog box "*suggest use case scenarios for LLM in teaching of Plastic Surgery*." The roles listed in the response were then analyzed by the authors to see whether it was relevant to the objective in question, appropriate for using a LLM model, and feasible by the LLM model (►**Table 2**). Relevance and appropriateness of a particular AI-generated task was resolved by discussion among authors and by checking LLM responses in certain cases. Items which were either irrelevant or inappropriate were excluded from the final list of tasks to be evaluated for feasibility. Of the 10 roles identified by LLM, role related to research, like research assistance, was excluded as our aim was to identify use case scenarios for academic

Table 1 Identification of LLM suitable roles

Tasks traditionally associated with TAs	Whether LLM can carry out this task (score from 0 to 4, 0 = not at all, 1 = to some extent, 2 = to a moderate extent, 3 = to a large extent, and 4 = to full extent)
Assisting faculty with classroom instruction, records, and assignments	4
Leading discussion sections	0
Meeting with students during office hours	0
Conferencing with students individually or in small groups	0
Delivering lectures or guest lectures	0
Leading group projects or discussions	0
Grading assignments or papers	3
Managing course communications	0
Preparing laboratory materials	0
Recording and calculating grades	0
Providing feedback on assignments	3
Enforcing laboratory rules and procedures	0
Proctoring examinations	0
Taking attendance or monitoring participation	0
Obtaining and distributing course materials	0
Ordering course textbooks and monitoring supply	0

Abbreviations: LLM, large language models; TA, teaching assistant.

Table 2 Evaluation of LLM-generated roles as a teaching assistant for plastic surgery residents

Sl.no.	LLM-generated roles for assisting in plastic surgery education	Score (0 = no, 1 = yes) Whether the LLM identified roles is		
		Relevant (for the student)	Appropriate (for the present study)	Feasible (for the authors)
1	Virtual tutor	1	1	1
2	Interactive case studies (generation)	1	1	1
3	Simulation of preoperative consultations	1	1	1
4	Procedure step-by-step guidance	1	1	1
5	Research assistance	1	0	1
6	Decision-making support	1	1	1
7	Collaborative learning platform	1	1	0
8	Ethical considerations (generation)	1	1	1
9	Enhancing surgical skills	1	1	1
10	Postoperative care and complication management	1	1	1

Abbreviation: LLM, large language models.

plastic surgery. Similarly, a role "collaborative learning platform" which referred to the integration of LLM into the learning management systems was also excluded as it did not appear feasible at the moment. A suitably designed prompt relevant to the role to the LLM model was submitted to the model, so as to generate an output. The output was scored for accuracy, usefulness, and clarity by two sets of evaluators on a Likert scale. The first set consisted of five evaluators who were aware of LLM as the source of the content. The second set consisted of three evaluators who were blinded to the source of the content.

Roles Identified for LLM

A final set of eight TA roles were identified, where a LLM could be utilized to generate content. These contents were

scored for usefulness, clarity, and accuracy. These were scored independently by the eight evaluators in a scoring sheet created for the study. The study authors were the evaluators divided into two sets. Initially, five authors including the corresponding author designed the prompts and analyzed the responses of the LLM model. Subsequently, the remaining three authors were independently enlisted to review these same responses. At this stage, they were not made aware of the responses' origin to ensure unbiased evaluation. Once their evaluations were complete, these authors were incorporated as contributors to the study, providing valuable insights for the discussion section. All authors are teaching faculty in medical colleges with a minimum of 4 years' experience in teaching plastic surgery. In addition, all authors are active participants in the medical education unit of their respective institutions.

Virtual tutor: This refers to the use of LLM to generate explanation of critical concepts in plastic surgery to aid the resident in improving their understanding of the subject. For the study, the LLM generated an explanation of mechanism of claw hand deformity in ulnar nerve palsy.

Prompt used: "*I am not able to understand why claw hand happens in an ulnar nerve injury. Can you explain me in an easy way the mechanism of claw hand deformity?*"

Percent overall agreement = 68.57%.

While most of the evaluators agreed with the perceived usefulness of the role, there was disagreement among authors regarding the accuracy of the content generated. There was agreement among authors on the clarity, comprehensiveness, and length of the passage generated.

Interactive case studies (generation): The LLM was utilized to generate hypothetical patient cases with varying levels of complexity for students to analyze, diagnose, and plan treatment, targeted toward residents in the 1st, 2nd, and 3rd year of their training.

Prompts used: "*generate hypothetical patient cases with varying levels of complexity for students to analyze, diagnose, and plan treatment in Plastic surgery.*"

Here, different short clinical scenarios were generated along with a treatment plan.

Percent overall agreement = 43.33%.

Prompts used: "*generate a case scenario for a beginner on mandible fracture in a 20 year old male. Give one subjective question based on this scenario and two multiple-choice questions (MCQs) based on the same scenario.*"

"*generate a case scenario for an advanced trainee on mandible fracture in a 20 year old male. Do Not display the treatment plan. Give one subjective question based on this scenario and two MCQs based on the same scenario.*"

Here, different short clinical scenarios were generated along with a few questions to test recall and understanding.

Percent overall agreement = 100.00%.

The case studies generated consisted of hypothetical patient details, examination findings, a possible diagnosis, and treatment plan. All evaluators agreed with the usefulness of the role and accuracy of the contents. With suitable prompts, the LLM was able to generate multiple-choice objective questions based on the scenarios generated and subjective questions, for stimulating critical thinking and recall among residents.

Simulation of preoperative consultations: The LLM was used to create preoperative consultation scenarios in plastic surgery to generate realistic patient concerns and expectations and present it to students. This would help students practice their communication skills and learn how to address patient concerns effectively.

Prompts used: "*simulate a preoperative consultation to help Plastic surgery residents students develop their communication skills and learn how to address patient concerns effectively by generating realistic patient concerns and expectations for a patient requiring free Latissimus dorsi flap cover for a leg defect.*"

"*simulate a preoperative consultation to help Plastic surgery residents students their communication skills and learn how to address patient concerns effectively by generating realistic patient concerns and expectations for a patient requiring breast implant and augmentation for hypoplastic breast due to Poland's syndrome. Create two suitable questions for residents based on the above scenario.*"

The LLM generated two scenarios one on use of free latissimus dorsi flap for leg defect and second on use of breast implants for breast augmentation in Poland syndrome.

Percent overall agreement = 100 and 66.67%, respectively, for the two scenarios.

There was absolute agreement among all evaluators on the usefulness of the role as well as clarity of the content generated. However, evaluators felt that content should be reviewed by instructors prior to administration to residents.

Procedure step-by-step guidance: The aim was to use the LLM to provide detailed, step-by-step guidance on various plastic surgery procedures, enabling students to learn at their own pace and reinforcing their understanding of surgical techniques.

Prompts used: "*list the operative steps of pollicization for thumb hypoplasia*"

"*list the steps for microsurgical arterial anastomosis for a Plastic Surgery resident*"

The evaluators were asked to score on accuracy, usefulness, and completeness of the content. While there were agreements on the usefulness of the content and role, there were disagreements among authors on the accuracy, clarity, and completeness of the generated content.

Percent overall agreement = 37.50 and 30%, respectively, for the two procedures.

Here, it is important to note that while the model gave fairly accurate procedural steps it did have inaccurate statements for the pollicization prompt like "*Neurovascular anastomosis: Reapproximate the radial digital neurovascular bundles to provide sensation and blood supply to the new thumb*." Neurovascular approximation is not done in pollicization rather the index finger is rotated in to thumb position over an intact neurovascular pedicle.

Evaluators felt that some critical steps were missed out on in the content and were likely to cause confusion among residents.

Decision-making support: The aim is to utilize LLM to create content to help plastic surgery residents weigh the pros and cons of different surgical approaches or techniques, promoting informed decision-making. This decision-making process would allow the resident to develop a comprehensive understanding of various surgical techniques and their applications in different clinical scenarios.

Prompt used: "*give an example for a plastic surgery resident on use of LLM for Decision-making Support: LLMs can be used as a tool to help students weigh the pros and cons of different surgical approaches or techniques, promoting informed decision-making.*"

The model created a content with a patient scenario having a leg defect in the lower third and gave a comparison between the two options, microvascular free latissimus dorsi flap and free anterolateral thigh flap.

Percent overall agreement = 36.67%.

While the evaluators were in good agreement on the accuracy and usefulness of the content, there was disagreement on clarity of the content. Moreover, the evaluators felt that responses were too generic and additional factors (for example, defect factors, patient factors, etc.) were not taken into consideration by the LLM when generating responses.

Ethical considerations (generation): The LLM was used to generate content with the aim to stimulate discussions on ethical dilemmas that may arise in plastic surgery practice, helping students develop a strong ethical foundation.

Prompt used: "*generate discussions on ethical dilemmas on a scenario that may arise in plastic surgery practice, to help residents develop a strong ethical foundation.*"

The LLM generated a case scenario of a female requesting rhinoplasty with unrealistic expectations, thus creating a dilemma for the operating surgeon to whether to operate or not in such a situation. The model also identified discussion points which may be taken into consideration to arrive at a decision, namely, patients' expectations, psychological factors, informed consent, patient autonomy versus surgeon expertise, and financial incentives.

Percent overall agreement = 100.00%.

All the evaluators were in absolute agreement on the usefulness, accuracy, and clarity of the content generated.

Enhancing surgical skills: The LLM was prompted to generate content with the aim to provide tips and tricks to residents for refining surgical skills, including suturing techniques, tissue handling, and instrument use.

Prompt used: "*Give an example of use of LLM in Enhancing Surgical Skills: LLMs can provide tips and tricks for refining surgical skills, including suturing techniques, tissue handling, and instrument use.*"

The LLM generated content directed toward a resident wanting to refine their suturing techniques and tissue handling skills during a cleft lip repair procedure.

Percent overall agreement = 36.67%.

While there was absolute agreement between the evaluators on the content clarity, there were disagreements regarding the accuracy and usefulness of the content. The evaluators felt that importance was not given to individual steps in the generated content. They also felt that this being more of a psychomotor role, LLM would not be able to assist residents as a standalone module.

Postoperative care and complication management: The LLM model was used to create content with the aim to educate students on postoperative care protocols and how to manage potential complications that may arise after plastic surgery procedures.

Prompt used: "*Suggest a postoperative care protocol for a 10-month-old child undergone Bardach palatoplasty for complete cleft palate. Suggest potential complications and way to manage them.*"

Percent overall agreement = 53.33%.

There was good agreement among the evaluators on the content clarity and usefulness, though some errors were identified in the content generated. For example, for the given prompt one statement said "*4. Wound care: Clean the surgical site gently with a cotton swab dipped in sterile saline or water, ensuring that no food particles or debris accumulate at the site. Apply antibiotic ointment as recommended by the surgeon*," which was not accurate for palatoplasty. Evaluators, though, felt that resident learning could be reinforced on specified postoperative protocols through the generated content.

Discussion

LLMs in general and ChatGPT (on which this study is based) in specific, are capable of generating answers to questions and prompts based on huge amount of text fed into the model for training the underlying language model.[9] The answers generated have been found to be accurate, readable, and even indistinguishable from human-generated text.[10] This capability of ACS can be exploited to generate summaries to text, answer short and long answers, and generate case scenarios. Inherent to the underlying model is the inability to reason or create original knowledge. Hence, this makes it unsuitable as a primary source of factual information, a disclaimer also given by the creators of ChatGPT on its chat interface. While there have been concerns on the use of ChatGPT in education and few universities and countries even going to the extent of blocking them, there seems to be available opportunities to integrate these in learning activities to aid students and teachers.[11]

Plastic surgery as a field often encompasses a broad variety of procedures that require a unique blend of art, creativity, and precision. Well-trained LLMs can offer theoretical knowledge and detailed technical insights, contributing to a cognitive foundation. However, the actual execution in the operative room, and the dexterity it demands, cannot be taught by an LLM.

We could identify a few scenarios where the LLM could in general be utilized to play the role of a TA and aid plastic surgery residents in particular. In addition, these models could also be used by students to obtain feedback and gain reflection which itself stimulates critical thinking. Each domain of Bloom's Taxonomy, namely, the cognitive, affective, and psychomotor, can be addressed to some extent by the LLM (►**Table 3**). For the cognitive domain, LLMs like

ChatGPT4 (as in our study) can provide conceptual knowledge and facilitate analytical thinking by engaging in discussions or scenario simulations. However, as we found in our study, the factual accuracy of the present models cannot be relied upon. LLM development being a rapidly changing field, this limitation may soon be overcome if the models are trained upon subject-specific technical content and have access to latest information and guidelines. The affective domain can be approached through discussing case scenarios that focus on empathy and professional values. For the psychomotor domain, while the LLM cannot perform physical demonstrations, it can provide step-by-step guidance, explanations of techniques, and respond to queries on practical aspects.

While the LLM was able to carry out the different roles of a TA in this study, it is important to identify roles which it could carry out adequately without a secondary check, that is, interactive case studies (generation), simulation of preoperative consultations, and generation of ethical considerations scenarios. The model was found to need secondary check of content generated in other roles. This was because the output in the former was more of a language output in a clinical scenario with low emphasis on content accuracy and the model has been adequately designed to carry out such activities. Activities needing presentation of factual information where there is a high emphasis on content accuracy will not be a suitable independent role for an LLM. Such generated content has to be validated by a subject expert.

It has to be remembered that the quality of output in LLM depends on the preciseness of the input prompt. As the model has been designed to identify words and present output in response to these words, a precise, well-described input prompt would generate an appropriate output. Nonetheless, it is preferable to check the accuracy of output responses.

TAs, albeit human, have been supporting teachers and students in an educational system. The idea of a virtual TA seems innovative and useful, where such assistants can be available at the fingertips of residents undergoing plastic surgical training.[12,13]

The Indian National Medical Commission's current Competency-Based Medical Education model can be incorporated into the LLM's virtual tutor task. Using the "Must Know, Need to Know, Good to Know" model, the LLM can be designed to prioritize information in response to queries, based on the categorization of these domains. For instance, for a "Must Know" topic, the LLM could provide comprehensive details while for a "Good to Know" topic, it might offer a brief overview unless more details are requested. In addition, the LLMs' capacity in objective structured clinical examination-based teaching and evaluation can be an interesting avenue for future research.

In the present era of evidence-based medicine, LLMs like ChatGPT4 can be trained to generate text based on the level of evidence available. However, it is important to note that at present ChatGPT4 cannot independently evaluate the quality of the evidence since its knowledge is based on preexisting data up to its last training cutoff, which for ChatGPT4 is September 2021. In future, these capabilities can be augmented by model training and allowing access to latest information.

With rapid strides in the development of AI in general and LLMs in particular, the possibility of having virtual TAs seems real. These language models have been trained on humongous amounts of textual data and their strength lies in rapidly generating text-based outputs to user inputs. Although all these models are in experimental phase and have not been converted into commercial applications, there lies an opportunity where the foundational models of ChatGPT and other LLMs can be fine-tuned to subject-specific activities and be made available as a personal chatbot.[14] Studies have demonstrated the use of these models as virtual surgical assistants.[15]

The advantages of LLMs also lie in the fact that most of the present generational learners are comfortable with the use of technology in their daily academic lives and so also the ubiquitousness of mobile devices in education and health care.[16] LLMs can thus behave as personal TAs to them aiding in cognitive and to some extent affective domains of learning. The conversational nature of output in ChatGPT is likely to retain learner attention and this makes it more attractive as a guide compared to routine search engines. They can assist in the cognitive domain of learning by aiding in knowledge acquisition, comprehension, and application.

Table 3 List of identified roles for LLM as a teaching assistant and the domains of the roles according to Bloom's Taxonomy

Sl. no.	LLM generated roles for assisting in plastic surgery education	Domains according to Bloom's Taxonomy
1	Virtual tutor	Cognitive
2	Interactive case studies (generation)	Cognitive
3	Simulation of preoperative consultations	Cognitive
4	Procedure step-by-step guidance	Psychomotor
5	Decision-making support	Cognitive
6	Ethical considerations (generation)	Affective
7	Enhancing surgical skills	Psychomotor
8	Postoperative care and complication management	Affective

Abbreviation: LLM, large language models.

Table 4 List of advantages and disadvantages of using large language models in plastic surgical training

Advantages	Drawbacks
1. Access to a vast amount of information 2. Automated content synthesis 3. Self-paced learning 4. Interactive learning experience	1. Lack of direct instructor interaction 2. Technical limitations (availability of Internet, mobile devices, etc.) 3. Potential information overload 4. Limited opportunity for practical application 5. Environmental concerns due to use of large computational resources 6. Limited validation of experimental models 7. Potential for overreliance on technology 8. Challenges in assessing practical skills

LLMs can also contribute to the affective domain of learning by encouraging learners to develop attitudes and responses as a reaction to simulated patient conversations, ethical dilemmas, and values conducive to professional growth (►**Table 4**).

The known concerns of LLMs include the ability to hallucinate, which means creation of imaginary content which sound real in response to a text input, lack of privacy, as all data entered into the model are used for training the underlying model further. Hence, it is strictly recommended not to enter personal information into the system. Other known concerns include environmental concerns due to the use of massive amounts of computational resources which consume significant energy and contribute toward carbon emissions.[17,18]

This study has tried to evaluate the role of LLM in plastic surgery residency programs and is the first such study to use ChatGPT4 model. The generated responses have been objectively and independently scored by the evaluators and inter-observer agreements have been calculated for the responses. The study has a few limitations, including a smaller number of evaluators and use of only one LLM although multiple such models are available. In addition, scores of a few responses and clinical scenarios of plastic surgery could only be evaluated. It is likely the study outcomes will change with a higher number of evaluators and case scenarios. Further larger studies with randomization, response evaluation by larger number of evaluators, or response evaluation even by residents could throw in more insights.

The future of education is an evolving landscape, where models like ChatGPT and others will play significant roles. With gradual acceptance of LLM in higher education, it is a matter of time when these models will prove to be useful as TA in the domain of plastic surgery education.[19,20]

Conclusion

This study has attempted to identify the use of LLM for resident training in plastic surgery. Incorporating LLMs into the educational arsenal of plastic surgery residency programs can provide a dynamic, interactive, and individualized learning experience for residents and prove to be worthy TAs of future.

Conflict of Interest
None declared.

References

1 Rouse M. Large Language Model (LLM). April 28, 2023https://www.techopedia.com/. Accessed on May 5, 2023 at: https://www.techopedia.com/definition/34948/large-language-model-llm#:~:text=level%20of%20accuracy.-,How%20Do%20Large%20Language%20Models%20Work%3F,implementation%20of%20a%20transformer%20architecture
2 Open AI. ChatGPT. Accessed February 8, 2023 at: https://openai.com/blog/chatgpt
3 Open AI. GPT-4 technical report. 2023. arXiv. Accessed August 7, 2023 at: https://arxiv. org/abs/2303.08774
4 Wolfram S. What is ChatGPT doing ... and why does it work? Stephen Wolfram Writings. February 14, 2023. Accessed August 7, 2023 at: https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/
5 Riedl M. A Very Gentle Introduction to Large Language Models Without the Hype. Medium. April 14, 2023. Accessed August 7, 2023 at: https://mark-riedl.medium.com/a-very-gentle-introduction-to-large-language-models-without-the-hype-5f67941fa59e
6 Li J, Dada A, Kleesiek J, Egger J. ChatGPT in Healthcare: A Taxonomy and Systematic Review. MedRxiv. 2023. Accessed April 4, 2023 at: https://doi.org/10.1101/2023.03.30.23287899
7 University of Wisconsin-Milwaukee. (n.d.). Roles and Responsibilities of Teaching Assistants. Graduate Assistants. Accessed April 20, 2023 at: https://uwm.edu/graduate-assistants/handbook/teaching-assistants/roles-and-responsibilities-of-teaching-assistants/
8 Randolph JJ. Online Kappa Calculator [Computer software]. 2008. Accessed August 7, 2023 at: http://justusrandolph.net/kappa/
9 Korinek A. Exploring the impact of language models on cognitive automation with David Autor, ChatGPT, and Claude. Brookings. March 6, 2023. Accessed March 22, 2023 at: https://www.brookings.edu/research/exploring-the-impact-of-language-models/#:~:text=One%20of%20the%20key%20advantages,more%20accurate%20and%20appropriate%20responses
10 Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Square 2023. rs.3.rs-2566942. https://doi.org/10.21203/rs.3.rs-2566942/v1 PubMed
11 Milano S, McGrane JA, Leonelli S. Large language models challenge the future of higher education. Nat Mach Intell 2023;5:333–334
12 Bagadood NH, Saigh BH. Teaching assistants as a prerequisite for best practice in special education settings in Saudi Arabia. Int J Comput Sci Network Security 2022;22(03):101–1C06
13 Lachman N, Christensen KN, Pawlina W. Anatomy teaching assistants: facilitating teaching skills for medical students through apprenticeship and mentoring. Med Teach 2013;35(01):e919e925
14 Wiggers K. The emerging types of language models and why they matter. TechCrunch. April 28, 2022. Accessed August 7, 2023 at: https://techcrunch.com/2022/04/28/the-emerging-types-of-language-models-and-why-they-matter/
15 Cheng K, Li Z, Li C, et al. The potential of GPT-4 as an AI-powered virtual assistant for surgeons specialized in joint arthroplasty. Ann Biomed Eng 2023;51(07):1366–1370

16 Mohapatra D, Mohapatra M, Chittoria R, Friji M, Kumar S. The scope of mobile devices in health care and medical education. Int J Adv Med Health Res 2015;2(01):3–8
17 Poda M. Large language models: the basics and their applications. Moveworks. February 9, 2023. Accessed August 7, 2023 at: https://www.moveworks.com/insights/large-language-models-strengths-and-weaknesses
18 Maastricht University. (n.d.). Large Language Models and Education. Accessed May 2, 2023 at: https://www.maastrichtuniversity.nl/large-language-models-and-education#risk
19 ChatGPT and Artificial Intelligence in Higher Education: Quick Start Guide [Internet]. United Nations Educational, Scientific and Cultural Organization; 2023. Accessed August 7, 2023 at: https://www.iesalc.unesco.org/wp-content/uploads/2023/04/ChatGPT-and-Artificial-Intelligence-in-higher-education-Quick-Start-guide_EN_FINAL.pdf
20 Tajik E, Tajik F. A comprehensive examination of the potential application of Chat GPT in higher education institutions. (Version 1). TechRxiv. 2023. Accessed August 7, 2023 at: https://doi.org/10.2196/45312

