How I GPT It: Development of Custom Artificial Intelligence (AI) Chatbots for Surgical Education

Tejas S. Sathe, MD,[*,†,††] Joshua Roshal, MD,[‡,§,††] Ariana Naaseh, MD,[‖,††] Joseph C. L'Huillier, MD,[¶,††] Sergio M. Navarro, MD, MBA,[#,**,††] and Caitlin Silvestri, MD[†,††]

[*]University of California San Francisco, San Francisco, California; [†]Columbia University Irving Medical Center, New York, New York; [‡]The University of Texas Medical Branch, Galveston, Texas; [§]Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; [‖]Washington University in St. Louis, St. Louis, Missouri; [¶]Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, New York; [#]University of Minnesota, Minneapolis, Minnesota; [**]Mayo Clinic, Rochester, Minnesota; and [††]The Collaboration of Surgical Education Fellows (CoSEF)

Artificial Intelligence (AI) chatbots provide a novel format for individuals to interact with large language models (LLMs). Recently released tools allow nontechnical users to develop chatbots using natural language. Surgical education is an exciting area in which chatbots developed in this manner may be rapidly deployed, though additional work will be required to ensure their accuracy and safety. In this paper, we outline our initial experience with AI chatbot creation in surgical education and offer considerations for future use of this technology. (J Surg Ed 81:772−775. © 2024 The Author(s). Published by Elsevier Inc. on behalf of Association of Program Directors in Surgery. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/))

KEYWORDS: artificial intelligence, chatbot, surgical education, education technology, innovation

COMPETENCIES: Practice-Based Learning and Improvement, Medical Knowledge

We are amidst a revolution in artificial intelligence (AI). In this paper, we aim to share our early experience with creating custom AI-powered chatbots in surgical education. While our versions are prototypes, we hypothesize that similar chatbots will one day play a significant role in the field.

AI is the ability of a computer to perform tasks that require human-like reasoning. Previously, AI was used to predict but not to create. For example, AI could analyze patterns in television show viewership and recommend new shows to watch but not make new shows on its own. This paradigm has shifted with the advent of new algorithms known as large language models (LLMs), which exhibit creative capacity. LLMs are trained on large collections of text data such as books, websites, and papers. Like prior algorithms, LLMs analyze a series of words (called a prompt) and predict the next best word to follow. What makes LLMs novel, however, is an ability to intuit not only the order of words, but their collective meaning too. As a result, LLMs can respond to prompts with human-like responses that do not exist explicitly within their training data - earning them the sobriquet of “generative AI”.[1]

The debut of ChatGPT by OpenAI made LLMs widely accessible to the public through an easy to use, web-based interface (https://chat.openai.com). In fact, ChatGPT was used by over 100 million people within two months of launch to plan vacations, draft emails, and even write code.[2] Following ChatGPT’s release, OpenAI introduced improved underlying models − such as GPT-4 − with greater capabilities, increased accuracy, and updated knowledge.

Thus far, surgeons have approached Generative AI with interest but skepticism. In a recent X (formerly Twitter) poll, a majority (52.5%) of surgeons felt uncomfortable with healthcare professionals or patients using ChatGPT to answer medical questions.[3] Despite this, one study found that GPT provided accurate responses to patient questions on colorectal cancer a majority of the time with high inter-rater reliability.[4] Moreover, another study found that a sample of adults from the United States had difficulty distinguishing between human and GPT-generated responses to questions.[5]

Correspondence: Inquires to Tejas S. Sathe, MD, +14407966769, 505 Parnassus Ave, San Francisco, CA 94143; e-mail: tejas.sathe@ucsf.edu

1931-7204/$30.00
https://doi.org/10.1016/j.jsurg.2024.03.004

TABLE 1. Examples of Custom GPTs

Initial Prompt	Training Knowledge	Link
"Make me a qualitative researcher who can help analyze focus group data with thematic analysis"	(Kiger and Vapiano, 2020)[13] (Stalmeije et. al. 2014)[14]	https://chat.openai.com/g/g-ANULT7gg3-qually-the-qualitative-researcher
"Make me an expert IRB reviewer that can help draft IRBs according to my institutions' format and help me determine if my study is exempt"	Federal Policy for the Protection of Human Subjects (the Common Rule) (2018 Revision)[15]	https://chat.openai.com/g/g-eDGmfjZb3-kirby
"Make me a patient support tool for patients who recently underwent bariatric surgery"	Post-operative pamphlets from the University of Texas Medical Branch	https://chat.openai.com/g/g-2Pyw7lbrh-barry
"Make me 10 high quality ABSITE-style multiple choice questions on thyroid-related topics."	Surgical Council on Resident Education modules, NBME Item-Writing Guide	https://chat.openai.com/g/g-IAbUj5I2h-high-yield-surgery-question-wizard

While the utility of LLMs in patient education requires further exploration, LLMs clearly exhibit a strong command of declarative knowledge for physicians. For example, GPT-4 was able to pass the United States Medical Licensing Exam and the written neurosurgery board certification exams.[6–8]

Within surgical education, one exciting application of LLMs is the ability to build custom chatbots that can create and disseminate educational content. Previously, making a surgery specific chatbot required choosing an LLM such as GPT-4, supplying an initialization prompt explaining what the bot was supposed to do, and building a chat-based user interface. We encountered two major obstacles with this approach. First, development of the chatbot required coding, making it inaccessible to nontechnical researchers. Moreover, while we were able to take advantage of GPT-4′s immense existing knowledge base, we were unable to supplement it with domain-specific knowledge unique to our use case. While methods exist to do so, they are currently beyond our technical abilities.[9]

On November 6, 2023, OpenAI unveiled new technology to develop custom chatbots (called **GPTs**) that may address both issues.[10] GPTs can be built and iteratively modified using natural language instead of code, and users can upload text documents to provide domain-specific knowledge. The user is first asked what type of chatbot they want to create and then is asked to further specify the chatbot's content, behavior, and tone. The user can upload domain-specific knowledge using a drag-and-drop interface. GPTs can even accept voice and image inputs. These chatbots can then be tested and modified in real time. For example, the user can type "make the chatbot's tone more formal" or "prompt the user by asking more questions."

The potential novel uses of this technology are expansive. Within the first week of its release, we designed chatbots to help users perform thematic analyses, design surveys, build curricula using Kern's Six-Step Framework, simulate American Board of Surgery In-Training Exam questions, and write research protocols according to institution-specific templates. We also developed a patient education chatbot to support patients after bariatric surgery (Table 1).

In our early experience, the GPTs we developed showed promise despite having significant limitations. For example, in our unpublished oral boards simulation GPT, the chatbot referenced pearls from the uploaded knowledge but often gave the user too much information up-front without being prompted. In other cases, the chatbot's performance was more readily usable. In our thematic analysis GPT, the chatbot intuited codes and themes similar to those created from manual review. Overall, GPTs performed better at offering general knowledge rather than adhering to a specific format. The ability to adjust additional parameters such as response-length and format will make these chatbots even more powerful in the future. This also highlights the importance of prompt engineering - the crafting of natural language prompts that the LLM can understand and implement.[11]

The ability for any user to make custom chatbots raises some concerns. First, how can the accuracy of these chatbots be verified? LLMs are known to hallucinate - that is, confidently say things that are made-up or false. However, each version of GPT claims to make fewer hallucinations than its predecessors, and the ability to integrate domain-specific knowledge may improve accuracy further still. Second, when we talk about improved accuracy, against what goalpost are we measuring? Currently, there are no rubrics with validity evidence to assess chatbot accuracy, and traditional methods such as Delphi consensus are unlikely to be fast enough to respond to the pace and breadth of chatbots developed. Novel frameworks are needed to determine chatbot accuracy at scale. Third, GPTs were found to leak information about their initialization prompts and training data, raising concerns about the use of private or proprietary information in their construction.[12] Finally, the OpenAI models used by GPTs are proprietary, meaning we do not know specifically how they work. Further development of open-source models will help alleviate this issue.

AI is a profound and rapidly evolving technology. It may seem reasonable to say that this technology is not accurate or safe enough today and that we should keep it at arm's length until it is. However, failure to use this resource risks the majority of surgeons, residents, and students being left out of the innovation process. Prior to ChatGPT, most AI-related work was performed by those with a technical background. The only way surgeons could participate was to provide their own data to train models. With the advent of generative AI tools, nontechnical users can participate in the AI revolution for the first time as partners. While there is significant discussion surrounding the regulation of AI, we must be wary of placing well-intentioned guardrails around its use that end up restricting access to those without prestige or position.

While remaining cognizant of their limitations, we should maintain proficiency in the use of these tools and iteratively find safe use cases. Education is likely to be an initial area of focus. As we have shown, educational chatbots can be deployed to simulate assessments, build curricula, analyze data, or write papers. By experimenting with them, we will have a skill base and infrastructure in place when the algorithms are accurate enough to be used in broader contexts such as patient care. This holds especially true for medical students and residents, who may be the most comfortable using these tools, but are the least likely to have a seat at the table when the parameters of their use are determined. Despite its risks, generative AI has the potential to be a democratizing technology in surgical education.

PREPRINT

A previous version of this article was preprinted at: https://www.ideasurg.pub/custom-surged-chatbots/.

GENERATIVE AI USE

Generative AI was used to make the chatbots described in this manuscript. We also used Generative AI to provide feedback on our revision, but did not use Generative AI written language in the revision itself.

ACKNOWLEDGMENT

The authors report no financial conflicts of interest relevant to this work. Of note, this manuscript references OpenAI technology. None of the authors have any financial relationship or association with OpenAI or Microsoft. Some authors purchase a monthly subscription to ChatGPT Plus. While we describe our experience with GPT-based chatbots, our paper does not constitute an endorsement of any one particular Generative AI product.

REFERENCES

1. Karpathy A. [1hr Talk] Intro to large language models. Published November 22, 2023. Accessed November 29, 2023. https://www.youtube.com/watch?v=zjkBMFhNj_g

2. ChatGPT sets record for fastest-growing user base - analyst note. *Reuters*. Published 2023: 2023. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/. Accessed November 29.

3. SURGERY Journal. I am comfortable for both healthcare professionals & patients to use #AI platforms such as #ChatGPT to answer their medical questions@SWexner @juliomayol @SyedAAhmad5 @dr_samehhany81 @mikifreund @nirhoresh @_TylerLoftus @hayfarani @Laparoscopes @luciacolorectal #SoMe4Surgery. Twitter. Published 2023. Accessed November 20, 2023. https://twitter.com/SurgJournal/status/1721280763998421415

4. Emile SH, Horesh N, Freund M, et al. How appropriate are answers of online chat-based artificial intelligence (ChatGPT) to common questions on colon cancer? *Surgery*. 2023;174(5):1273–1275. https://doi.org/10.1016/j.surg.2023.06.005.

5. Nov O, Singh N, Mann DM. Putting ChatGPT's medical advice to the (Turing) Test. *Biorxiv*. Published online 2023. https://doi.org/10.1101/2023.01.23.23284735.

6. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198.

7. 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(3):904–911. https://doi.org/10.3171/2023.2.JNS23419.

8. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board

examinations. *medRxiv*. Published online 2023: 2023.03.25.23287743. https://doi.org/10.1101/2023.03.25.23287743.

9. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP Tasks. *arXiv [csCL]*. Published online 2020. http://arxiv.org/abs/2005.11401.

10. Introducing GPTs. Accessed 2023. https://openai.com/blog/introducing-gpts.

11. Nori H, Lee YT, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv [csCL]*. Published online 2023. http://arxiv.org/abs/2311.16452.

12. Burgess M. OpenAI's custom chatbots are leaking their secrets. *Wired*. Published online 2023. Accessed December 3, 2023. https://www.wired.com/story/openai-custom-chatbots-gpts-prompt-injection-attacks/.

13. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE Guide No. 131. *Med Teach*. 2020;42(8):846–854. https://doi.org/10.1080/0142159X.2020.1755030.

14. Stalmeijer RE, Mcnaughton N, Van Mook WNKA. Using focus groups in medical education research: AMEE Guide No. 91. *Med Teach*. 2014;36(11):923–939. https://doi.org/10.3109/0142159X.2014.917165.

15. Office for Human Research Protections (OHRP). Revised common rule. HHS.gov. Published 2017. Accessed December 5, 2023. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/finalized-revisions-common-rule/index.html