JAMA | Special Communication | AI IN MEDICINE

Creation and Adoption of Large Language Models in Medicine

Nigam H. Shah, MBBS, PhD; David Entwistle, BS, MHSA; Michael A. Pfeffer, MD

IMPORTANCE There is increased interest in and potential benefits from using large language models (LLMs) in medicine. However, by simply wondering how the LLMs and the applications powered by them will reshape medicine instead of getting actively involved, the agency in shaping how these tools can be used in medicine is lost.

OBSERVATIONS Applications powered by LLMs are increasingly used to perform medical tasks without the underlying language model being trained on medical records and without verifying their purported benefit in performing those tasks.

CONCLUSIONS AND RELEVANCE The creation and use of LLMs in medicine need to be actively shaped by provisioning relevant training data, specifying the desired benefits, and evaluating the benefits via testing in real-world deployments.

JAMA. 2023;330(9):866-869. doi:10.1001/jama.2023.14217
Published online August 7, 2023.

Viewpoint page 801
Related article page 792
CME at jamacmelookup.com

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Nigam H. Shah, MBBS, PhD, Center for Biomedical Informatics Research, Stanford University, 3180 Porter Dr, Palo Alto, CA 94305 (nigam@stanford.edu).

Large language models (LLMs) and the applications built using them, such as ChatGPT, have become popular. Within 2 months of the November 2022 release, ChatGPT surpassed 100 million users. The medical community has been pursuing off-the-shelf LLMs provided by technology companies. New users have been asking how the LLMs and the chatbots powered by them will reshape medicine.[1] Perhaps the reverse question should be asked: How can the intended medical use shape the training of the LLMs and the chatbots or the other applications they power?

Language models learn the probabilities of occurrence for sequences of words from the corpus of text. For example, if the corpus had the 2 questions of "where are we going" and "where are we at," the probability is 0.5 for seeing the word *going* after seeing the 3 words *where are we*. An LLM is essentially learning such probabilities on a massive scale, such that the resulting model has billions of parameters (a glossary appears in the **Box**). In 2017, Vaswani et al[2] demonstrated that a certain kind of deep neural network, called a transformer, could learn LLMs that later performed amazingly well at language translation tasks. Their insight led to the creation of hundreds of language models that were reviewed by Zhao et al.[3]

Although language models are trained to predict the next word in a sentence (basically an advanced autocomplete), new capabilities (such the ability to summarize text and answer questions posed in natural language) become possible without explicitly training for them, which allow the model to perform tasks such as pass medical licensing examinations, simplify radiology reports, extract drug names from a physician's note, reply to patient questions, summarize medical dialogues, and write histories and physical assessments.[4] ChatGPT, perhaps the most popular application, uses an LLM called a generative pretrained transformer (GPT; version 3.5 or 4.0) underneath to ingest text and output text in response.

The creation of language models capable of such diverse tasks hinges on 2 things. First is the ability to learn generally useful patterns in large amounts of unlabeled data via self-supervision (training and interacting with an LLM in the **Figure**). For example, a commonly used form of self-supervision is to predict the next word in a sequence conditioned on prior words, which later identifies the words that go together in general. The GPT-3 model was trained on 45 terabytes of text data comprising roughly 500 billion tokens (1 token is approximately 4 characters or three-fourths of a word for English text) at a cost of approximately $4.6 million.[5]

Second is the subsequent tuning of the LLM to generate responses aligned with human expectations via instruction tuning. For example, in response to the request, "explain the moon landing to a 6-year-old in a few sentences," the GPT-3 model suggested possible completions as "explain the theory of gravity to a 6-year-old" and "explain the big bang theory to a 6-year-old" (instruction tuning an LLM in the Figure). Users helped train GPT-3 by providing the instructions (also called prompts) for which the labelers (hired by OpenAI, the company that built GPT-3) provided demonstrations of the desired output and ranked the outputs from the model. OpenAI used these pairs of instructions and their desired outputs to instruction tune GPT-3.[6]

Although general-purpose LLMs can perform many medically relevant tasks, they have not been exposed to medical records during self-supervised training and they are not specifically instruction tuned for any medical task. By not asking how the intended medical use can shape the training of LLMs and the chatbots or other applications they power, technology companies are deciding what is right for medicine. The medical profession has made a mistake in not shaping the creation, design, and adoption of most information technology systems in health care. Given the profound disruption that is possible for such diverse activities as clinical documentation, decision support, information technology operations, medical coding, and patient-physician communication with the use of LLMs (estimated in a McKinsey report to be as high as 1.8%-3.2% of total health care revenues[7]), the same mistake

Box. Glossary

Chatbot
A computer program designed to simulate conversation with human users, especially over the internet.

Deep neural network
A setup for machine learning inspired by biological neural networks in which computational units referred to as neurons are arranged in a network that is composed of multiple layers of interconnected neurons, allowing it to learn complex patterns in the data presented to it.

Large language model (LLM)
Learns the probabilities of occurrence of sequences of words from a corpus of text, whose probabilities are learned using textual corpora with trillions of words such that the resulting model has billions of parameters.

Self-supervision
A learning approach in which a machine learning model learns without relying on explicitly labeled data as examples. Instead, the model generates its own training objective from the input data without the need for human-annotated data, which can be time-consuming and expensive to produce. A common type of self-supervision is in the form of an autoregressive training objective, in which the model is trained to predict the next word or token in a sequence, given the previous words or tokens. The training objective is to maximize the likelihood of the correct word given the context. Training in this manner is often the first stage in training LLMs (generative pretrained transformer) and helps the model learn language structure, grammar, and semantics. Learning to predict the next medical code in a patient's longitudinal medical record does not require a human to label a code as the "next code"; that information is available in the data directly by looking at the sequence of appearance of the code in the medical record.

Transformer
A deep neural network architecture that is designed to be efficient at capturing relationships and dependencies between elements in a sequence, such as words in a sentence.

Instruction tuning
Refers to a kind of tuning in which an existing LLM is adapted (via tuning) to respond accurately and effectively to natural language instructions. This process involves continuing to train the model on a dataset containing pairs of instructions and corresponding desired outputs or responses. Doing so allows the model to be more useful in real-world applications, such as providing relevant information, answering questions, or following specific commands provided by users in a natural language.

Tuning
Refers to the process of adapting a pretrained LLM to perform well on a specific task or domain. This process involves training the model on a smaller labeled dataset that is specific to the target task, such as sentiment analysis, machine translation, or answering questions. For example, in a medical setting, a model could be tuned for tasks such as summarizing the available past medical records of a patient or the course of their current admission. During tuning, the model's weights and parameters are updated using pertinent examples to optimize its performance on the target task. This allows the model to build on the general language understanding it gained via self-supervised learning, while adapting to the nuances and specific requirements of the task at hand.

cannot be repeated. At a minimum, the medical profession should be asking the following questions.

Are the LLMs Being Trained With the Relevant Data and the Right Kind of Self-Supervision?

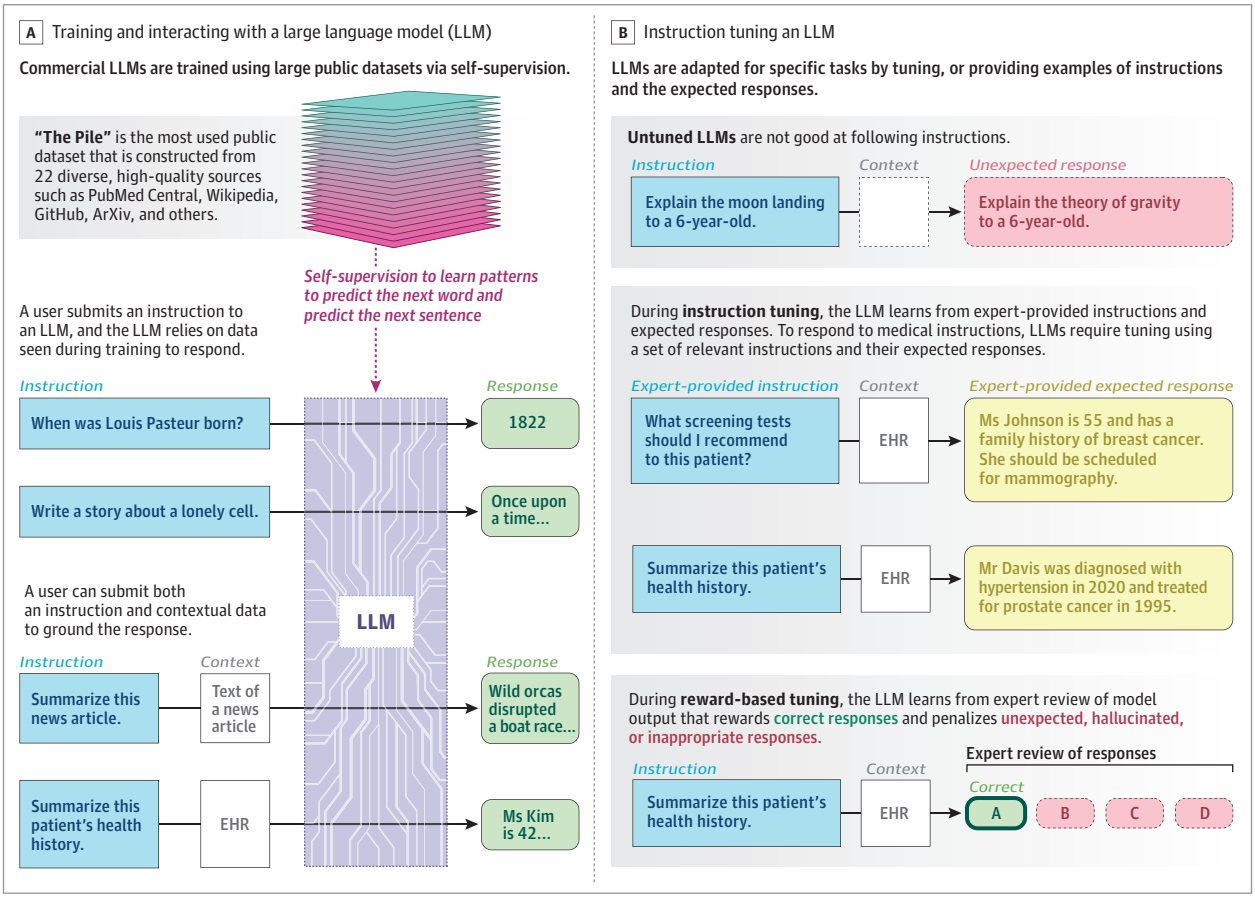
Medical records can be viewed as consisting of sequences of time-stamped clinical events represented by medical codes and textual documents, which can be the training data for a language model. Wornow et al[8] reviewed the training data and the kind of self-supervision used by more than 80 medical language models and found 2 categories.

First, there are medical LLMs that are trained on documents. The self-supervision is via learning to predict the next word in a textual document, such as a progress note or a PubMed abstract, and conditioned on prior words seen. Therefore, these models are similar in their anatomy to general purpose LLMs (eg, GPT-3), but are trained on clinical or biomedical text. These models can be used for language manipulation tasks such as summarization, translation, and answering questions. Given the increased training and use costs of LLMs, it is necessary to investigate whether smaller language models trained on relevant data may achieve the desired performance at a lower cost. For example, researchers at the Center for Research on Foundation Models at Stanford University created a model called Alpaca with 4% as many parameters as OpenAI's text-davinci-003, matching its performance at a cost of $600 to create.[9]

Second, there are medical LLMs that are trained on the sequence of medical codes in a patient's entire record that take time into account. Here, the self-supervision is in the form of learning the probability of the next day's set of codes, or learning how much time elapses until a certain code is seen. As a result, the sequence and timing of medical events in a patient's entire record is considered. As a concrete example, given the code for "hypertension," these models learn when a code for a stroke, myocardial infarction, or kidney failure is likely to occur. When provided with a patient's medical record as input, such models will not output text but instead a machine understandable "representation" of that patient, referred to as an "embedding," which is a fixed-length, high-dimensional vector representing the patient's medical record. Such embeddings can be used in building models for predicting 30-day readmissions, long hospital lengths of stay, and in-patient mortality using less training data (as few as 100 examples).[10]

The medical community needs to actively shape the creation of LLMs in medicine. For example, given the importance of instruction tuning, the medical community should be discussing how to create shared instruction tuning datasets with examples of prompts to be fulfilled, such as "summarize the past specialist visits of a patient" with its corresponding valid completion (Figure). Perhaps instead of using GPT-4 at the cost of $0.06 to $0.12 per 1000 tokens (about 75 words), health care systems

Figure. An Overview of the Key Issues in Shaping the Creation and Adoption of Large Language Models (LLMs) in Medicine



EHR indicates electronic health record. A, During training, LLMs learn generally useful patterns in large amounts of unlabeled data via self-supervision. For example, a commonly used form of self-supervision is to predict the next word in a sequence conditioned on prior words, as seen in large public datasets. Once an LLM is trained, users can interact with it via submitting an instruction (or prompt), to which the LLM responds with a sequence of words that are its valid completions. B, As is, LLMs are not good at following instructions. The LLMs are adapted for specific tasks by providing examples of instructions (blue) and the expected responses (yellow). This adaptation process is called tuning. In responding to medical instructions, such as "summarize the past specialist visits of a patient," LLMs require tuning using a set of relevant instructions and their expected responses.

should be training shared, open-source models using their own data. The technology companies should be asked whether the models being offered have seen any medical data during training and whether the nature of self-supervision used is relevant for the final use of the model.

Are the Purported Value Propositions of Using LLMs in Medicine Being Verified?

Current evaluations of LLMs also do not quantify the benefits of novel collaboration between humans and artificial intelligence that is at the core of using these models in clinical settings. The methods for evaluating LLMs in the real world remain unclear. Concerns with current evaluations range from training dataset contamination (such as when the evaluation data are included in the training dataset) to the inappropriateness of using standardized examinations designed for humans to evaluate the models. Consider the analogy of evaluating a person for a driver's license. The person takes a multiple-choice, knowledge-based test. The car, meanwhile, undergoes safety tests during manufacturing, some of which are regulated by the government. Then the person gets in the car for a road test to certify them for a license. The car does not take a multiple-choice test at the department of motor vehicles or get certified for driving, but that is the absurdity tolerated for LLMs when it is declared that they are certified to give medical advice because they passed the US medical licensing examination.

The purported benefits need to be defined and evaluations conducted to verify such benefits.[8] Only after these evaluations are completed should statements be allowed such as an LLM was used for a defined task in this specific workflow, it measured a metric, and observed an improvement (or deterioration) in a prespecified outcome. Such evaluations also are necessary to clarify the medicolegal risks that might occur with the use of LLMs to guide medical care,[11] and to identify mitigation strategies for the models' tendency to generate factually incorrect outputs that are probabilistically plausible (called hallucinations).

Conclusion

The building of relevant medical LLMs needs to be balanced with verifying the presumed value propositions via testing in real-world deployments akin to road driving tests. If the goal in using such models is to augment human judgment, and not replace it, adopting this driving test mindset is critically important. Otherwise, there is a risk of falling into the trap of automating tasks that individuals already know how to do, and failing to ask the question of what a person plus such models could do together that may yield better medical care.[12]

Given the highly disruptive potential of these technologies, clinicians cannot afford to be on the sidelines. The adoption of LLMs in medicine needs to be shaped by the medical profession that can identify the right training (and instruction tuning) data and perform the evaluations that verify the purported benefits of using LLMs in medicine.

ARTICLE INFORMATION

Accepted for Publication: July 11, 2023.

Published Online: August 7, 2023. doi:10.1001/jama.2023.14217

Author Affiliations: Stanford Health Care, Palo Alto, California (Shah, Entwistle, Pfeffer); Department of Medicine, School of Medicine, Stanford University, Stanford, California (Shah, Pfeffer); Clinical Excellence Research Center, School of Medicine, Stanford University, Stanford, California (Shah).

Author Contributions: Dr Shah had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: All authors.
Drafting of the manuscript: Shah, Pfeffer.
Critical review of the manuscript for important intellectual content: All authors.
Administrative, technical, or material support: Entwistle, Pfeffer.
Supervision: Pfeffer.

Conflict of Interest Disclosures: Dr Shah reported being a co-founder of Prealize Health (a predictive analytics company) and Atropos Health (an on-demand evidence generation company). No other disclosures were reported.

Additional Contributions: We acknowledge the members of the data science team at Stanford Health Care for helpful discussions to refine the arguments made in this article. We acknowledge Jason Fries, PhD, and Alison Callahan, PhD (both with Stanford University), for help in creating the first draft of the Figure; they were not compensated for their contributions.

REFERENCES

1. Li R, Kumar A, Chen JH. How chatbots and large language model artificial intelligence systems will reshape modern medicine: fountain of creativity or Pandora's box? *JAMA Intern Med*. 2023;183(6):596-597. doi:10.1001/jamainternmed.2023.1835

2. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems: NIPS '17*. Curran Associates Inc; 2017:6000-6010.

3. Zhao WX, Zhou K, Li J, et al. A survey of large language models. *arXiv*. Preprint posted online March 31, 2023. https://arxiv.org/abs/2303.18223v10

4. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388(13):1233-1239. doi:10.1056/NEJMsr2214184

5. Wikipedia contributors. GPT-3. Published May 8, 2023. Accessed July 25, 2023. https://en.wikipedia.org/w/index.php?title=GPT-3&oldid=1153892380

6. OpenAI. Aligning language models to follow instructions. Published January 27, 2022. Accessed May 22, 2023. https://openai.com/research/instruction-following

7. Chui M, Hazan E, Roberts R, et al. The economic potential of generative AI: the next productivity frontier. Published June 14, 2023. Accessed June 16, 2023. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

8. Wornow M, Xu Y, Thapa R, et al. The shaky foundations of large language models and foundation models for electronic health records. *NPJ Digital Med*. 2023;135(6). doi:10.1038/s41746-023-00879-8

9. Taori R, Gulrajani I, Zhang T, Dubois Y, Li X. Stanford Alpaca: code and documentation to train Stanford's Alpaca models, and generate the data. Accessed June 16, 2023. https://github.com/tatsu-lab/stanford_alpaca

10. Steinberg E, Jung K, Fries JA, Corbin CK, Pfohl SR, Shah NH. Language models are an effective representation learning technique for electronic health record data. *J Biomed Inform*. 2021;113:103637. doi:10.1016/j.jbi.2020.103637

11. Mello MM, Guha N. ChatGPT and physicians' malpractice risk. *JAMA Health Forum*. 2023;4(5):e231938. doi:10.1001/jamahealthforum.2023.1938

12. Brynjolfsson E. The turing trap: the promise and peril of human-like artificial intelligence. In: *Augmented Education in the Global Age*. Routledge; 2023:103-116.