

# Large Language Models in Medicine: The Potentials and Pitfalls

## A Narrative Review

Jesutofunmi A. Omiye, MD, MS<sup>\*</sup>; Haiwen Gui, BS<sup>\*</sup>; Shawheen J. Rezaei, MPhil; James Zou, PhD; and Roxana Daneshjou, MD, PhD

Large language models (LLMs) are artificial intelligence models trained on vast text data to generate humanlike outputs. They have been applied to various tasks in health care, ranging from answering medical examination questions to generating clinical reports. With increasing institutional partnerships between companies producing LLMs and health systems, the real-world clinical application of these models is nearing realization. As these models gain traction, health care practitioners must understand what LLMs are, their development, their current and potential applications, and the associated pitfalls in a medical

setting. This review, coupled with a tutorial, provides a comprehensive yet accessible overview of these areas with the aim of familiarizing health care professionals with the rapidly changing landscape of LLMs in medicine. Furthermore, the authors highlight active research areas in the field that promise to improve LLMs' usability in health care contexts.

*Ann Intern Med.* 2024;177:210-220. doi:10.7326/M23-2772 **Annals.org**

For author, article, and disclosure information, see end of text.

This article was published at Annals.org on 30 January 2024.

\* Dr. Omiye and Ms. Gui contributed equally as co-first authors.

### Key Summary Points

Large language models (LLMs) represent a rapidly advancing field with transformative potential across health care, requiring elucidation of their clinical promise and risks.

Rigorous benchmarks must be developed to evaluate LLMs on standardized medical data sets and tasks, enabling reproducibility and progress tracking.

Innovative techniques to mitigate algorithmic biases are needed to ensure that LLMs promote health equity across diverse populations.

Novel experiments integrating LLMs as clinical decision support tools can show effects on outcomes, productivity, and patient satisfaction.

Multidisciplinary perspectives crossing computer science, medicine, ethics, and health policy will provide vital insights into optimizing the societal benefits of LLMs in health care.

Active research questions on general-purpose LLMs versus medical LLMs, explainability, and bias mitigation are introduced.

Tutorial-styled use cases with one of the popular LLMs for medical tasks are provided in the **Supplement** (available at Annals.org).

### See also:

Web-Only  
Supplement

Large language models (LLMs) have become mainstream since the launch of OpenAI's publicly available ChatGPT in November 2022 (1). This milestone was quickly followed by the unveiling of similar models like Google's Bard (2) and Anthropic's Claude (3), alongside open-source variants, such as Meta's LLaMA (4). Large language models are a type of foundation model (5) (Glossary) trained on massive amounts of text data and can have billions of parameters (6). They receive text queries and generate text in return (7). They can be adapted to a wide range of language-related tasks beyond their primary training objective. In medicine, they have been applied to various tasks like note taking (8), answering medical examination questions (9), answering patient questions (10), and generating clinical summaries (8). Despite their versatility, LLMs' behaviors are poorly understood (7), and they have the potential to produce medically inaccurate outputs (11) and amplify existing biases (12, 13).

Evidence suggests that interest in LLMs is growing among physicians (8, 14), and institutional partnerships are on the rise. Examples include their use in a training module for medical residents at Beth Israel Deaconess Medical Center (Boston, Massachusetts) (15) and a partnership with Epic, a major provider of electronic health records, to integrate generative pre-trained transformer (GPT)-4 into their services (16). As these models gain traction, it is imperative for physicians and other health care professionals to understand what LLMs are, their development, their current and potential applications, and the associated pitfalls when used in medicine. In this review, we offer a comprehensive overview of the training processes behind LLMs, providing background to understand their benefits and shortcomings. We delve into the historical and current applications of LLMs in medicine, discussing both their potential and limitations. We provide

descriptive overviews of popular LLMs and explore various prompting techniques, which define how users can interact with these models. We spotlight pressing research questions in the domain and introduce platforms that allow health professionals to remain updated on LLM developments in medicine. Also, we present tutorial-style use cases (**Supplement**, available at [Annals.org](https://annals.org)), enabling health care practitioners to experiment with capabilities—such as drafting prior authorizations and creating patient handouts—using the freely available ChatGPT model powered by GPT-3.5.

## LLM ARCHITECTURE

Large language models rely on the “Transformer” architecture (17, 18), which leverages an “attention” mechanism that uses multilayered neural networks to help LLMs comprehend context and learn meaning within sentences and long paragraphs (6). Akin to how a physician identifies important details of a patient's case while ignoring extraneous information, this mechanism enables LLMs to “learn” important relationships between words while ignoring irrelevant information.

The training of these complex models involves billions of parameters and vast amounts of data and has been made possible by recent advancements in computational power and model architecture (5) (**Figure 1**). For example, GPT-3 reportedly has about 175 billion parameters (19), whereas the open-source LLaMA family of models have 7 billion to 70 billion parameters (4, 20). The first step of LLM training, known as pretraining, is a self-supervised approach that involves training on a large corpus of unlabeled data, such as internet text, Wikipedia, Github code, social media posts, and

BookCorpus (4–6). The result of this step is a base model that is a general language-generating model but can lack the capacity for nuanced tasks. Models may then undergo reinforcement learning, where outputs are further refined with human feedback. Base models can then be fine-tuned, which is less computationally expensive, on narrow data sets like medical transcripts for a health care application, or legal briefs for a legal assistant bot (21).

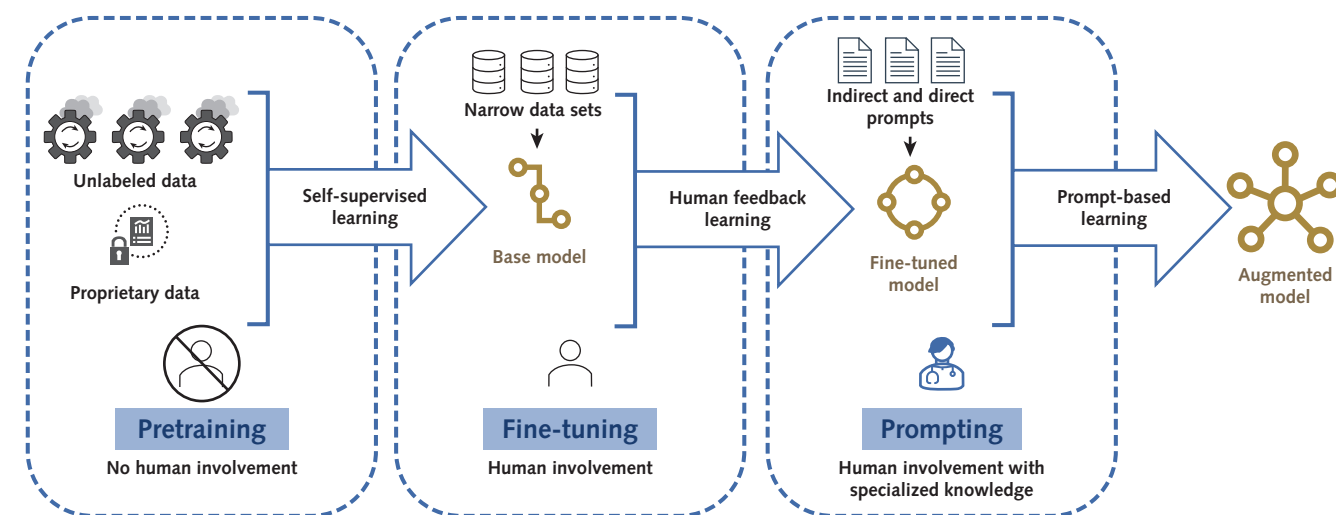
The adaptability of LLMs to unfamiliar tasks and their apparent reasoning abilities (22) are captivating. However, unlocking their full potential in specialized fields like medicine is believed to require specific training strategies, including prompting techniques, where humans interact with the model to provide specialized instructions. Overall, these methods augment the core training processes of fine-tuned models and can enhance their alignment with medical tasks, as recently shown with the Fine-tuned LLaLanguage Net Pathways Language Model (Flan-PaLM) (11). As these models continue to evolve, understanding their training methodologies will serve as a good foundation for discussing their current capabilities and future applications.

Section 1 of the **Supplement** provides more details on LLM architecture.

## OVERVIEW OF CURRENT MEDICAL LLMs

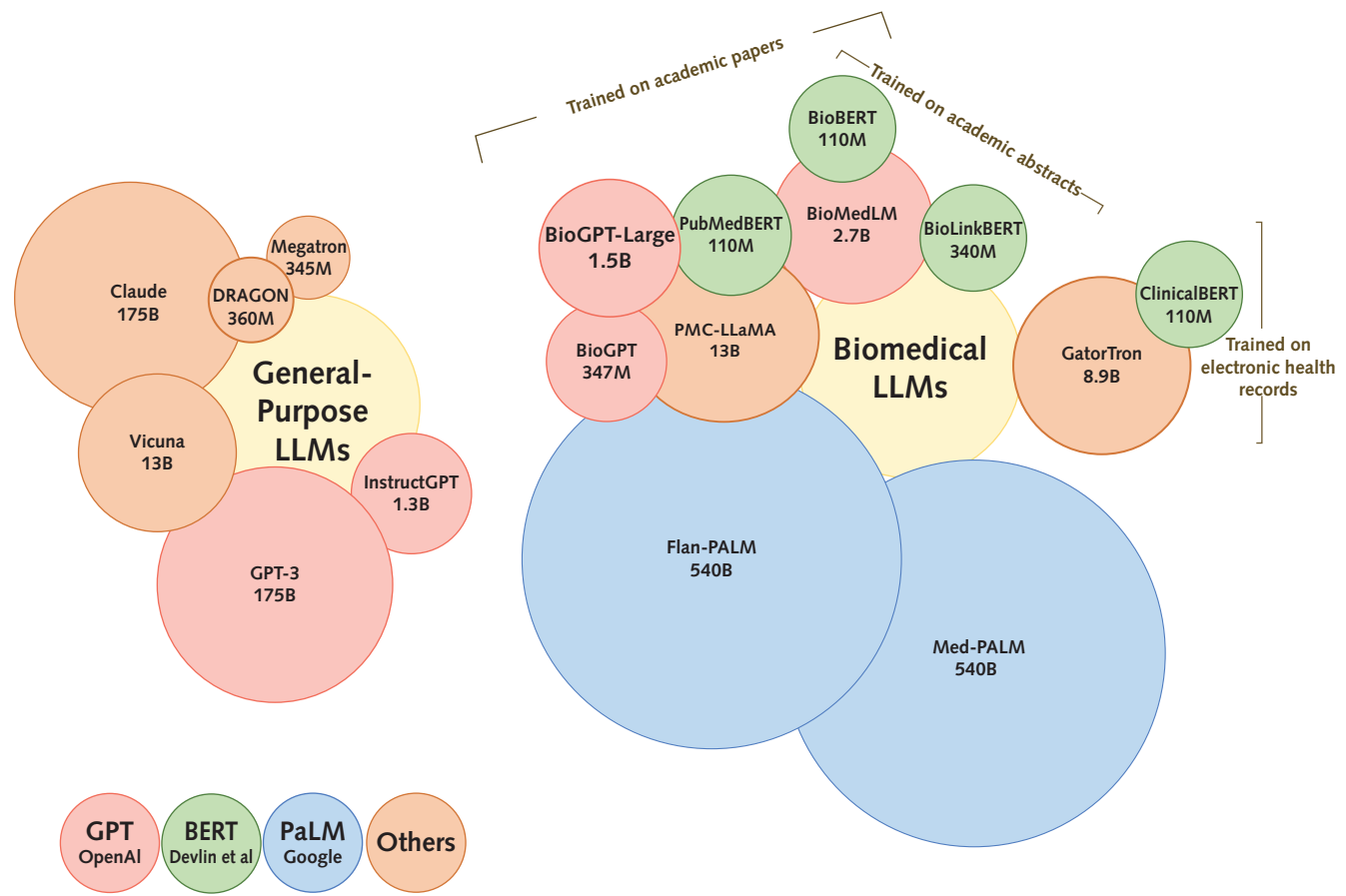
Before the emergence of LLMs, natural language processing challenges were tackled by more rudimentary models like statistical language models and neural language models (6), which had significantly fewer parameters and trained on relatively small data sets. The advent of the Transformer architecture heralded the age of LLMs we see today. The performance

**Figure 1.** Overview of LLM training process.



LLMs “learn” from more focused inputs at each stage of the training process. The first phase of this learning is pretraining, where the LLM can be trained on a mix of unlabeled data and proprietary data without any human supervision. The second phase is fine-tuning, where narrower data sets and human feedback are introduced as inputs to the base model. The fine-tuned model can then enter an additional phase, where humans with specialized knowledge implement prompting techniques that can transform the LLM into a model that is augmented to perform specialized tasks. LLM = large language model.

Figure 2. Current LLMs in medicine.



General-purpose and biomedical LLMs are currently used for medical tasks. Although GPT by OpenAI, BERT by Devlin and colleagues (26), and PaLM by Google have led the development of LLMs with applications in medicine, other proprietary and open-source LLMs also exist in this space. Circle sizes reflect the model size and the number of parameters used to build the models. LLMs with applications in medicine vary widely in how they were trained. BioMedLM 2.7B, based on GPT, was trained on the corpus of PubMed articles and abstracts, for example, whereas ClinicalBERT was trained specifically on electronic health records. These differences in training and development can have important implications for how LLMs perform in certain medical scenarios. BERT = Bidirectional Encoder Representations from Transformers; DRAGON = Deep Bidirectional Language-Knowledge Graph Pretraining; Flan-PaLM = Fine-tuned LAnguage Net PaLM; GPT = generative pretrained transformer; LLM = large language model; PaLM = Pathways Language Model.

of LLMs on medical tasks has largely focused on question answering (QA) benchmarks like MedMCQA, PubMedQA (11, 23) (Glossary), and MultiMedBench (24). However, to better reflect actual medical tasks, recent studies have evaluated models on generation of medical content, such as radiology reports (25).

In this section, we will provide an overview of general-purpose LLMs, with a specific emphasis on those that have been applied to tasks within the medical field. In addition, we will delve into domain-specific LLMs, referring to models that have been either pre-trained or fine-tuned using medical literature (Figure 2).

GPTs

Arguably the most popular of the general LLMs are those that belong to the GPT lineage, which generate language. Developed by OpenAI in 2018 (27),

the GPT series has significantly scaled in recent years. The latest version, GPT-4, represents a leap forward in its ability to handle multimodal input, such as images, text, and audio—an attribute that aligns seamlessly with the multifaceted nature of medical practice. Novel prompting techniques were introduced with the GPT models, paving the way for the popular ChatGPT product, which is based on GPT-3.5 and GPT-4. ChatGPT has demonstrated its utility in various medical scenarios discussed in the section Overview of LLMs Applied to Medicine (8, 28–30). Certain studies have concentrated on evaluating the health care utility of ChatGPT and InstructGPT, whereas others have focused on fine-tuning them for specific medical tasks. Luo and colleagues (28) introduced BioGPT, a model that used the GPT-2 framework pretrained on 15 million PubMed abstracts for tasks including QA, relation extraction, and

document classification. Their model outperformed the state-of-the-art models across all evaluated tasks. Similarly, BioMedLM 2.7B (formerly known as PubMedGPT), pre-trained on both PubMed abstracts and full texts (31), demonstrates the continued advancements in this field. Some researchers have even leveraged GPT-4 to create multimodal medical LLMs, reporting promising results (32, 33).

### BERT

First introduced by Devlin and colleagues (26), Bidirectional Encoder Representations from Transformers (BERT) uniquely focuses on understanding sentences through bidirectional training of the model (for example, the model learns to predict words on the basis of words that come before and after), compared with models that used 1-sided context. As an encoder-based model, BERT learns context and language relationships but does not generate language, making it adept for classifying language. For medical tasks, researchers have developed domain-specific versions of BERT tailored to scientific and clinical text. BioBERT incorporates biomedical corpus data from PubMed abstracts and PubMed Central articles during pretraining (34). PubMedBERT follows a similar methodology using just PubMed abstracts (35). ClinicalBERT adapts BERT for clinical notes, trained on the large Medical Information Mart for Intensive Care III data set of electronic health records (36). More recent work has focused on enhancing BERT for specific applications; for example, BioLinkBERT adds entity linking to connect biomedical concepts in text to ontologies (37). These extensions showcase how baseline BERT architectures can be customized for medicine.

### PaLM

Developed by Google, PaLM is one of the largest LLMs to date. Researchers first fine-tuned PaLM for medical QA, creating Flan-PaLM (38), which achieved state-of-the-art results on QA benchmarks. Building on this, Med-PaLM was produced via instruction tuning, demonstrating strong capabilities in clinical knowledge, scientific consensus, and medical reasoning (11). This has recently been extended to create a multimodal medical LLM (24). Models based on PaLM underscore the utility of large foundation models fine-tuned for medicine.

### Other LLMs

Beyond big tech companies, other proprietary and open-source medical LLMs have emerged. Models trained from scratch on clinical corpora, such as GatorTron (39), have shown improved performance on certain tasks compared with general-domain LLMs. Claude, developed by Anthropic, has been evaluated on medical biases and other safety issues for clinical applications (12). Active open-source projects are also contributing to the medical LLM field. For example, PMC-LLaMA leverages the LLaMA model and incorporates biomedical

articles in its pretraining (40). Other popular base models like Deep Bidirectional Language-Knowledge Graph Pretraining (DRAGON) (41), Megatron (42), and Vicuna (43) have enabled development of multimodal LLMs incorporating visual data (32).

Overall, domain-specific pretraining on medical corpora produces models that excel on biomedical tasks compared with generalist LLMs (with some exceptions). However, fine-tuning approaches for adapting general models like BERT and GPT-3 have achieved strong results on medical tasks in a more computationally efficient manner. This is promising given the challenges of limited medical data for training.

## APPLYING LLMs TO MEDICAL TASKS

### Overview of LLMs Applied to Medicine

In a short time, research has explored the usage of LLMs in medicine, ranging from answering patient questions in cardiology (9), to serving as a support tool in tumor boards (44), to aiding researchers in academia (45). A brief search in PubMed for *ChatGPT* in December 2023 returned more than 1800 results, showing the rapid investigation of this technology.

Before ChatGPT, many patients used the internet to learn more about their health conditions (46, 47). As ChatGPT surfaced, one proposed use was answering patient questions. In cardiology, researchers have found that ChatGPT can adequately respond to prevention questions, suggesting that LLMs could help augment patient education and patient-clinician communication (9). Similarly, researchers have explored ChatGPT's responses to common patient questions about hip replacements (10) and radiology report findings (48). In addition, there has been interest in using ChatGPT to translate medical texts and clinical encounters to improve patient communication and satisfaction (49). These findings suggest a potential for bridging gaps in patient education; however, additional testing is needed to ensure fairness and accuracy.

In addition to augmenting patient education, researchers are exploring the use of LLMs as clinical workflow support tools. One study evaluated ChatGPT's recommendations for next-step management in breast cancer tumor boards, which frequently comprise the most complex clinical cases (44). Others explored the use of ChatGPT in responding to patient portal messages (29), creating discharge summaries (30), and generating structured templates for radiology (48). Although these studies suggest opportunities to mitigate the documentation burden facing physicians, rigorous real-world evaluation should be completed before any clinical use.

Aside from uses in clinical medicine, LLMs are being used in medical education and academia. Researchers have explored LLMs' ability to conduct radiation oncology physics calculations (50), answer medical board questions in the style of the United States Medical Licensing



**Table.** Analysis of Possible LLM Tasks in Medicine

Task	Potential Pitfalls	Mitigation Strategies
Administrative: Write insurance authorization letters Summarize medical notes Aid medical record documentation Create patient communication (e-mail/letter/text)	Lack of HIPAA adherence: No publicly available model is currently HIPAA-compliant, and thus PHI cannot be shared with the models.	Integrate LLMs within electronic health record systems.
Augmenting knowledge: Answer diagnostic questions Answer questions about medical management Create and translate patient education material	Inherent bias: Pretrained data models used for diagnostic analyses will introduce inherent bias.	Create domain-specific models that are trained on carefully curated data sets. Always include a human in the loop.
Medical education: Write recommendation letters Create new examination questions and case-based scenarios Generate summaries of medical text at a student level	Lack of personalization: LLMs are generated from prior work already published, resulting in repetitive and unoriginal work.	Educate clinicians and users in using LLM tools to augment their work rather than replace them. Encourage understanding how the technology works to mitigate unrealistic expectations of output.
Medical research: Generate research ideas and novel directions Write academic papers Write grants	Ethics: A large amount of discussion has occurred among the scientific community on the ethics of using ChatGPT to generate scientific publications. This also raises the question of accessibility and the potential difficulties of future access to this technology.	Engage in conversation to increase accessibility of this technology to prevent widening gaps in research disparities.

HIPAA = Health Insurance Portability and Accountability Act; LLM = large language model; PHI = protected health information.

Examination (51), and respond to clinical vignettes (52). The ability of this technology to adequately achieve passing scores on these medical examinations raises questions on the need to revise medical curricula and practices (53). Other programs have started exploring using LLMs’ generative ability to create multiple-choice questions for student examinations (54). In academia, LLMs’ ability to aid researchers has been increasingly explored, ranging from topic brainstorming (55) to writing journal articles (45) and resulting in a rising debate on the ethics and usage of LLMs in academic writing.

**Proposed Tasks of LLM in Medicine**

On the basis of the current body of literature, this technology has many potential applications, from administrative tasks to gathering and enhancing medical knowledge (Table).

**LIMITATIONS AND MITIGATION STRATEGIES**

Although researchers have demonstrated the potential for LLMs in medicine, these preliminary studies have many limitations, emphasizing the need for future research. As discussed briefly in the Table, there are many potential pitfalls that clinicians should consider. Key challenges posed by LLMs include issues related to accuracy, bias, model inputs and outputs, and privacy and ethical concerns. By understanding and addressing these limitations, researchers can develop and use these models to create a more equitable and trustworthy ecosystem.

**Accuracy Issues and Data Set Bias**

Models are only as accurate as the data sets used to train them, resulting in a reliance on the accuracy and completeness of the training data. Large language

models are trained on large data sets that surpass the ability of human teams to manually check quality. This results in a model that is trained on a nebulous data set that may further decrease user trust in these algorithms. Because of the inability to check the data set’s quality, the training and testing data sets often overlap, resulting in overprediction of model accuracies (19, 21). In addition, factual information used in model training can become outdated, and retraining the model on updated information is nontrivial.

Of note, ChatGPT and many other LLMs are not trained on curated medical data sets but rather on a broad range of inputs, from news articles to literary works, that allow models to capture linguistic patterns and features. Moreover, the models do not “understand” the actual content and thus can generate completely fabricated responses. This can result in poor performance in domain-specific questions, including medical applications (56).

Models also frequently enhance and reinforce structural biases that are found in the training data sets. Groups have shown that models are promoting race-based medicine practices that have long been scientifically refuted. When answering questions about calculations of estimated glomerular filtration rate, several LLMs tried to justify race-based medicine with false assertions that Black people have different muscle mass and thus higher creatinine levels (12). Other researchers have found that LLMs associate phrases referencing people with disabilities with words expressing more negative sentiment, and that gun violence, homelessness, and drug addiction are overrepresented in texts discussing mental illness (57). In another scenario, LLMs were asked to provide analgesia choices for chest pain for White and Black patients, resulting in weaker analgesic recommendations for Black patients (58).

These accuracy and bias issues could be mitigated by training these models on domain-specific data sets (59). Aside from fine-tuning previously trained LLMs, work has been done to create models from scratch using electronic health record data, called clinical foundation models (60). These models were shown to have better predictive performance, require fewer labeled data, more effectively handle multimodal data, and offer novel interfaces for human-artificial intelligence (AI) interaction (60). Clinicians can also aid developers in decreasing data set bias by working to gather more diverse data sets for these models to train on and providing human feedback to improve model responses.

Innovations to the basic LLM architecture, including retrieval-augmented generation models (61), can also help increase accuracy and reduce “hallucinations,” where the model produces nonsensical and factually incorrect responses (62). Retrieval-augmented generation is a framework that allows LLMs to retrieve facts from an external knowledge base to allow increased accuracy and explainability analyses. It combines 2 models: A pretrained retriever model finds information relevant to the query in the external database, which then feeds the output to the pretrained generator model, such as ChatGPT. This framework creates the ability to update information in the database without retraining the billion-parameter models. Of greater note, it also creates the ability to point to sources in the database that aided in the text generation, increasing trust in these LLMs. By enabling augmentation of LLMs with external databases, health organizations can build systems that are compliant with the Health Insurance Portability and Accountability Act using proprietary hospital data (63).

### Weak Input, Poor Output, and Change Over Time

Large language models can be fickle; small changes in the input prompting can result in dramatic changes in the output. These variations in prompt syntax often occur in ways that are unintuitive to the users (21). This causes difficulties with ensuring consistency when using LLMs in a health care setting.

Models frequently generate hallucinations; this is exacerbated when insufficient information is provided in the prompt, a scenario that is frequently seen in health care. Researchers prompted LLMs to summarize documents and showed that the models will insert grossly inaccurate information not found in the original document inputs (64). In addition, these language models frequently use confident language in the output, which could lead users to blindly trust LLM outputs despite incorrect information (65).

Because many LLMs take inputs as truthful, they attempt to generate an output that fits the user's assumption rather than offering factual corrections or asking to clarify prompts (58). This inherently raises challenges for use cases in medicine, where researchers and clinicians may exacerbate misconceptions, worsening confirmation biases. To help mitigate some of these

limitations, clinicians and researchers should be well versed in prompt engineering to encourage accurate and sensible use of this new technology (66).

Aside from the fickle nature of inputs and unpredictable nature of outputs, LLMs are also evolving rapidly and unpredictably over time. This makes it challenging to incorporate these models into larger workflows. Also, as LLMs become widespread, automation bias could emerge, where users overrely on their outputs and potentially overlook their limitations or biases. Given the novelty of this technology, much work still remains to observe long-term trends and analyses. Use of this technology in health care should not be without careful oversight.

### Privacy and Ethical Concerns

Personally identifiable information has been found frequently in pretraining data sets in earlier LLMs, including telephone numbers and e-mail addresses (67). Even if the data sets are completely devoid of such information, privacy violations can occur due to inference. Large language models may make correct assumptions about a patient based on correlational data about other people without access to private data about that particular individual. In other words, LLMs may attempt to predict the patient's gender, race, income, or religion on the basis of user input, ultimately violating the individual's privacy (65).

Aside from these privacy issues, there are ethical concerns with LLM use in medicine. Even when we assume nonmalicious users of these models, opportunities unfortunately exist for the models to generate harmful content. For example, when clinicians disclose difficult diagnoses in medicine, steps are in place that can provide support and help patients cope. With the rise of LLMs in medicine, patients may inadvertently be exposed to difficult topics that can cause severe emotional harm. Although this problem is not unique to language models because patients have other means of accessing information (such as Google), LLMs produce a greater risk given their conversation-like structure and public availability. Many of these models are human enough, but they frequently lack the ability to provide additional personalized emotional support. Also, LLMs are prone to adversarial attacks that could make them generate objectionable outputs, which can be catastrophic for patient management (68).

Another ethical concern is the increasing difficulty of distinguishing between LLM-generated and human-written text, which could result in the spread of misinformation, plagiarism, and impersonation. For example, although the use of LLMs to aid clinicians in administrative tasks can help decrease documentation burden, it could also result in malicious use by others to generate false documents. Mostly big technology companies currently oversee the race to develop AI and LLMs, which raises ethical questions about yielding control to corporate interests. There is also the unresolved question of who bears the liability when these models make

## GLOSSARY

**Neural networks:** Systems inspired by the neuronal connections in the brain that are capable of learning, recognizing patterns, and making predictions on tasks without explicit programming. They are the building blocks of many modern machine-learning (deep-learning) algorithms.

**Foundation model:** A large-scale neural network model trained on vast data to develop broad learning capabilities, which can be fine-tuned for specific tasks. A foundation model can be fine-tuned to generate reports or answer medical questions.

**Generative artificial intelligence (AI):** Models trained on large data sets that can produce seemingly novel realistic content. This can be audio, visual, or text.

**Large language models (LLMs):** AI models trained on an enormous amount of text data. LLMs are capable of generating humanlike text and learning relationships between words.

**Transformer architecture:** A deep-learning model architecture that relies on self-attention mechanisms by differentially weighting the importance of each part of the model's input. This makes it particularly useful for language tasks.

**Attention:** A mechanism within the Transformer architecture that enables the differential weighting mentioned above.

**Parameters:** Values that are learned during the training process of a model.

**Self-supervised learning:** A form of training a model where it learns from unlabeled data but uses the input data as its own supervision. A popular example is predicting the next word in a sentence.

**Tokenization:** A pretraining process in which text is converted into smaller units, like a character or a word, before being fed into the model. For example, hypertension can be tokenized into the following: "hy," "per," "tension."

**Pretraining:** The initial phase of training a model on a large data set before fine-tuning it on a task-specific data set. The parameters are updated in the training process.

**Fine-tuning:** Further training a pretrained model on a specific task and adjusting the preexisting parameters to achieve better performance for a particular task.

**Zero-shot prompting:** A technique in which language prompts are used to get a model to perform specific

tasks without having seen explicit examples of those tasks.

**Few-shot prompting:** A technique in which the model is provided with some examples of the task, hence the name "few."

**Chain-of-thought prompting:** A technique in which a sequential series of intermediate logical steps toward arriving at the problem is provided to the language model. This has been shown to be useful in numerical problems. An example would be, "Let us think step-by-step like a physician/cardiologist/rheumatologist."

**Instruction tuning:** Refining a pretrained model by providing explanations (instructions) on how to perform a task, alongside labeled examples that demonstrate the training objective or desired behavior.

**Multimodal LLMs:** Models capable of processing and generating different types of data, such as text, images, and audio. They are an emerging form of LLM with a wide range of potential applications in medicine.

**In-context learning:** The ability of a model to understand and generate appropriate responses based on the context of a given input.

**Bias (in AI):** Systematic errors in the output of a model due to flawed assumptions in the machine-learning process. This is usually from the data the model are trained on and can also be accentuated in the fine-tuning process.

**MedMCQA:** An open-source data set of question-answer pairs that contains a collection of high-quality multiple-choice questions (over 194 000) from 2 medical entrance examinations, All India Institute Of Medical Science and National Eligibility Entrance Test (Postgraduate). It is one of the data sets used to develop and evaluate medical-related LLMs.

**PubMedQA:** Like the MedMCQA, but focused on biomedical research question-answer pairs. It is also useful for developing and evaluating LLMs.

mistakes. Because of these concerns, health professionals should be aware and take active roles in the development of LLMs for medicine.

Regular auditing and evaluation of LLMs can help identify, address, and mitigate these privacy and ethical concerns. There have been calls recently in the AI research community to develop regulations for LLM use in medicine (69). In addition, care must be taken in the selection of training data sets, especially when using medical domain-specific data sets, to ensure adequate handling of sensitive data.

## FUTURE OF LLMs IN MEDICINE AND RESEARCH DIRECTIONS

Large language models are currently at the forefront of AI innovation in medicine, with a surge of new developments being introduced regularly. Their potential to improve care delivery and alter the practice of medicine is notable. However, some research questions need to be answered for safe deployment in clinical settings. Here, we discuss what the future developments for LLMs in medicine could look like, drawing on currently emerging trends and future conjectures.

### Technological Advancements

The integration of multiple data types into LLMs, called multimethod, is an emerging trend with significant implications for health care (70). Initially introduced by GPT-4, this property has been further developed for medicine with a proof-of-concept generalist medical AI called Med-PaLM Multimodal (Med-PaLM M) (24). Recent studies, such as Large Language and Vision Assistant for BioMedicine (LLaVA-Med) (71), SkinGPT-4 (33), and MiniGPT-4 (32), provide compelling evidence for the effectiveness of multimodal LLMs, which are poised to gain prevalence in health care due to the multifaceted nature of medical data that span text, images, audio, and genetics.

Simultaneously, progress in minimizing resource requirements for LLMs is likely to democratize access. For example, methods like low-rank adaptation of LLMs (72) can significantly reduce training time and computational needs. This could benefit health professionals in resource-limited settings, enabling them to train LLMs for their own clinical and research tasks. By extension, this could reduce racial and gender bias (13) in model outputs as more robust models are developed. With this decrease in computational resources, the development of institution-specific LLMs becomes feasible. These would be models trained on data unique to a particular health institution, thereby mirroring its standard protocols, patient population, and distinct challenges. Such models have the potential to boost productivity, mitigate burnout, and improve patient care. Of note, current general-purpose LLMs like GPT-4 often outperform even specialized medical LLMs in medical tasks. This observation raises a pivotal question: Should the focus be on refining medical LLMs, or is the pursuit of even larger general-purpose LLMs more beneficial for health care?

### Accessibility and Equity

The creation of synthetic medical data by leveraging the generative capabilities of LLMs also offers a promising approach to address the challenges associated with the scarcity of medical research data. More diverse medical data could be available for training AI models, leading to more inclusive and equitable medical research. Further studies in the medical domain are essential to elucidate the potential of LLMs in advancing

health equity. Although synthetic generation of structured data has recently been reported (73), few studies have explored this technique in medical applications. This leaves a notable gap in the literature and underscores the need for benchmarks in evaluating synthetic medical data generation by LLMs.

### Regulatory Considerations

From a regulatory standpoint, it is imperative to establish standard frameworks for validating LLMs across clinical tasks while ensuring fairness. This is particularly crucial in medicine, where inaccurate model outputs can have severe consequences and lead to patient harm. Governance structures for LLMs need to evolve to protect patient privacy and address issues like model transparency, fairness, and accuracy (69).

### Future Research Directions

Future research directions for LLMs in health care include enhancing model explainability, standardizing holistic evaluation metrics, and addressing medical bias. The **Supplement** provides more details on future directions. To make significant progress in LLM-medicine research, cross-disciplinary and institutional collaborations will be crucial.

For physicians and other health professionals following the latest updates on LLMs in medicine, several platforms offer valuable insights. Before the peer review process, many LLM papers are first uploaded to preprint servers, including arXiv (<https://arxiv.org>), a preprint server for computer science, quantitative fields, and AI that hosts papers on LLMs applied to medicine; bioRxiv ([www.biorxiv.org](http://www.biorxiv.org)), a preprint server for biology; and medRxiv ([www.medrxiv.org](http://www.medrxiv.org)), a preprint server for medicine and health sciences. The Hugging Face model repository (<https://huggingface.co/models>) offers hands-on experience with the models and hosts a leaderboard of popular LLMs and data sets.

In summary, the future of LLMs in medicine will likely feature advancements that enhance their utility as supportive tools for health care workers, not replacement. These developments could play a crucial role in addressing challenges related to health care shortages and inefficiencies.

### Conclusion

Large language models have risen in popularity as models become more widely available for public use, and potential opportunities exist for the application of LLMs in the medical field. Tech companies have already developed models trained with the intention of performing medical tasks. There are several areas of medicine where LLMs could be used, such as administrative tasks (for example, summarizing medical notes), augmentation of clinician knowledge (for example, translating patient materials), medical education (for example, creating new examination questions), and medical research (for example, generating novel research ideas). Despite these opportunities, many notable challenges with LLMs



remain unresolved, limiting the implementation of these models in medicine. Issues affecting adoption include underlying biases in data sets, data quality and unpredictability of outputs, patient privacy, and ethical concerns. Physicians and other health care professionals must weigh potential opportunities with these existing limitations as they seek to incorporate LLMs into their practice of medicine.

From Department of Dermatology and Department of Biomedical Data Science, Stanford University, Stanford, California (J.A.O., R.D.); Department of Dermatology, Stanford University, Stanford, California (H.G., S.J.R.); and Department of Biomedical Data Science, Stanford University, Stanford, California (J.Z.).

**Disclosures:** Disclosures can be viewed at [www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M23-2772](http://www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M23-2772).

**Corresponding Author:** Roxana Daneshjou, MD, PhD, 1265 Welch Road, MSOB West Wing, Third Floor, Stanford, CA 94305, e-mail, [roxanad@stanford.edu](mailto:roxanad@stanford.edu).

**Correction:** This article was amended on 19 March 2024 to correct numerical input errors in Figure 2. These changes do not affect the conclusions of the article. A correction has been published (doi:10.7326/L24-0048).

Author contributions are available at [Annals.org](http://Annals.org).

## References

1. Introducing ChatGPT. OpenAI blog. 30 November 2022. Accessed at <https://openai.com/blog/chatgpt> on 4 August 2023.
2. Pichai S. An important next step on our AI journey. Google. 6 February 2023. Accessed at <https://blog.google/technology/ai/bard-google-ai-search-updates> on 4 August 2023.
3. Introducing Claude. Anthropic. 14 March 2023. Accessed at [www.anthropic.com/index/introducing-claude](http://www.anthropic.com/index/introducing-claude) on 4 August 2023.
4. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. arXiv. Preprint posted online 27 February 2023. doi:10.48550/arXiv.2302.13971
5. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. arXiv. Preprint posted online 16 August 2021. Revised 12 July 2022. doi:10.48550/arXiv.2108.07258
6. Zhao WX, Zhou K, Li J, et al. A survey of large language models. arXiv. Preprint posted online 31 March 2023. Revised 29 June 2023. doi:10.48550/arXiv.2303.18223
7. Liang P, Bommasani R, Lee T, et al. Holistic evaluation of language models. arXiv. Preprint posted online 16 November 2022. Revised 1 October 2023. doi:10.48550/arXiv.2211.09110
8. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388:1233-1239. [PMID: 36988602] doi:10.1056/NEJMSr2214184
9. Sarraju A, Bruemmer D, Van Iterson E, et al. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA*. 2023;329:842-844. [PMID: 36735264] doi:10.1001/jama.2023.1044
10. Mika AP, Martin JR, Engstrom SM, et al. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. *J Bone Joint Surg Am*. 2023;105:1519-1526. [PMID: 37459402] doi:10.2106/JBJS.23.00209

## Large Language Models in Medicine: Potentials and Pitfalls

11. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620:172-180. [PMID: 37438534] doi:10.1038/s41586-023-06291-2
12. Omiye JA, Lester JC, Spichak S, et al. Large language models propagate race-based medicine. *NPJ Digit Med*. 2023;6:195. [PMID: 37864012] doi:10.1038/s41746-023-00939-z
13. Bender EM, Gebru T, McMillan-Major A, et al. On the dangers of stochastic parrots: can language models be too big? In: *FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, virtual, 3-10 March 2021. Association for Computing Machinery; 2021:610-623. doi:10.1145/3442188.3445922
14. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. arXiv. Preprint posted online 30 December 2022. doi:10.48550/arXiv.2212.14882
15. Trang B, Palmer K. Preparation over panic: how a Boston hospital is priming medical residents for an era of AI medicine. *STAT*. 20 July 2023. Accessed at [www.statnews.com/2023/07/20/chatgpt-gpt4-health-care-medical-education](http://www.statnews.com/2023/07/20/chatgpt-gpt4-health-care-medical-education) on 4 August 2023.
16. Eddy N. Epic, Microsoft partner to use generative AI for better EHRs. *Healthcare IT News*. 18 April 2023. Accessed at [www.healthcareitnews.com/news/epic-microsoft-partner-use-generative-ai-better-ehrs](http://www.healthcareitnews.com/news/epic-microsoft-partner-use-generative-ai-better-ehrs) on 4 August 2023.
17. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Guyon I, Von Luxburg U, Bengio S, et al, eds. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, California, 4-9 December 2017. Accessed at <https://proceedings.neurips.cc/paper/7181-attention-is-all> on 4 August 2023.
18. Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, virtual, 16-20 November 2020. Association for Computational Linguistics; 2020:38-45. Accessed at <https://aclanthology.org/2020.emnlp-demos.6> on 4 August 2023.
19. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, et al, eds. *Advances in Neural Information Processing Systems 34 (NeurIPS 2020)*, virtual, 6-12 December 2020. Accessed at [https://proceedings.neurips.cc/paper\\_files/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html) on 4 August 2023.
20. Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models. arXiv. Preprint posted online 18 July 2023. Revised 19 July 2023. doi:10.48550/arXiv.2307.09288
21. Kaddour J, Harris J, Mozes M, et al. Challenges and applications of large language models. arXiv. Preprint posted online 19 July 2023. doi:10.48550/arXiv.2307.10169
22. Lewkowycz A, Andreassen A, Dohan D, et al. Solving quantitative reasoning problems with language models. arXiv. Preprint posted online 29 June 2022. Revised 1 July 2022. doi:10.48550/arXiv.2206.14858
23. Liévin V, Hother CE, Winther O. Can large language models reason about medical questions? arXiv. Preprint posted online 17 July 2022. Revised 24 January 2023. doi:10.48550/arXiv.2207.08143
24. Tu T, Azizi S, Driess D, et al. Towards generalist biomedical AI. arXiv. Preprint posted online 26 July 2023. doi:10.48550/arXiv.2307.14334
25. Liu Z, Li Y, Shu P, et al. Radiology-Llama2: best-in-class large language model for radiology. arXiv. Preprint posted online 29 August 2023. doi:10.48550/arXiv.2309.06419
26. Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online 11 October 2018. Revised 24 May 2019. doi:10.48550/arXiv.1810.04805
27. Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. Accessed at [www.mikecaptain.com/resources/pdf/GPT-1.pdf](http://www.mikecaptain.com/resources/pdf/GPT-1.pdf) on 4 August 2023.


28. Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform.* 2022;23:bbac409. doi:10.1093/bib/bbac409 [36156661]
29. Liu S, McCoy AB, Wright AP, et al. Leveraging large language models for generating responses to patient messages. *medRxiv.* Preprint posted online 16 July 2023. doi:10.1101/2023.07.14.23292669
30. Singh S, Djalilian A, Ali MJ. ChatGPT and ophthalmology: exploring its potential with discharge summaries and operative notes. *Semin Ophthalmol.* 2023;38:503-507. [PMID: 37133418] doi:10.1080/08820538.2023.2209166
31. Stanford Center for Research on Foundation Models; MosaicML. BioMedLM. Hugging Face. Updated 22 December 2022. Accessed at <https://huggingface.co/stanford-crfm/BioMedLM> on 4 August 2023.
32. Zhu D, Chen J, Shen X, et al. MiniGPT-4: enhancing vision-language understanding with advanced large language models. *arXiv.* Preprint posted online 20 April 2023. Revised 2 October 2023. doi:10.48550/arXiv.2304.10592
33. Zhou J, He X, Sun L, et al. SkinGPT-4: an interactive dermatology diagnostic system with visual large language model. *medRxiv.* Preprint posted online 13 June 2023. doi:10.1101/2023.06.10.23291127
34. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics.* 2020;36:1234-1240. [PMID: 31501885] doi:10.1093/bioinformatics/btz682
35. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc.* 2021;3:1-23. [10.1145/3458754]
36. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. *arXiv.* Preprint posted online 10 April 2019. Revised 29 November 2020. doi:10.48550/arXiv.1904.05342
37. Yasunaga M, Leskovec J, Liang P. LinkBERT: pretraining language models with document links. *arXiv.* Preprint posted online 29 March 2022. doi:10.48550/arXiv.2203.15827
38. Chung HW, Hou L, Longpre S, et al. Scaling instruction-finetuned language models. *arXiv.* Preprint posted online 20 October 2022. Revised 6 December 2022. doi:10.48550/arXiv.2210.11416
39. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med.* 2022;5:194. [PMID: 36572766] doi:10.1038/s41746-022-00742-2
40. Wu C, Lin W, Zhang X, et al. PMC-LLaMA: towards building open-source language models for medicine. *arXiv.* Preprint posted online 27 April 2023. Revised 25 August 2023. doi:10.48550/arXiv.2304.14454
41. Yasunaga M, Bosselut A, Ren H, et al. Deep bidirectional language-knowledge graph pretraining. *arXiv.* Preprint posted online 17 October 2022. Revised 19 October 2022. doi:10.48550/arXiv.2210.09338
42. Shoyebi M, Patwary M, Puri R, et al. Megatron-LM: training multi-billion parameter language models using model parallelism. *arXiv.* Preprint posted online 17 September 2019. Revised 13 March 2020. doi:10.48550/arXiv.1909.08053
43. Chiang WL, Li Z, Lin Z, et al. Vicuna: an open-source chatbot impressing GPT-4 with 90%\* ChatGPT quality. 30 March 2023. Accessed at <https://lmsys.org/blog/2023-03-30-vicuna> on 4 August 2023.
44. Sorin V, Klang E, Sklair-Levy M, et al. Large language model (ChatGPT) as a support tool for breast tumor board. *NPJ Breast Cancer.* 2023;9:44. [PMID: 37253791] doi:10.1038/s41523-023-00557-8
45. Brameier DT, Alnasser AA, Carnino JM, et al. Artificial intelligence in orthopaedic surgery: can a large language model "write" a believable orthopaedic journal article? *J Bone Joint Surg Am.* 2023;105:1388-1392. [PMID: 37437021] doi:10.2106/JBJS.23.00473
46. Cocco AM, Zordan R, Taylor DM, et al. Dr Google in the ED: searching for online health information by adult emergency department patients. *Med J Aust.* 2018;209:342-347. [PMID: 30107763] doi:10.5694/mja17.00889
47. Van Riel N, Auwerx K, Debbaut P, et al. The effect of Dr Google on doctor-patient encounters in primary care: a quantitative, observational, cross-sectional study. *BJGP Open.* 2017;1:bjgpopen17X100833. [PMID: 30564661] doi:10.3399/bjgpopen17X100833
48. Grewal H, Dhillon G, Monga V, et al. Radiology gets chatty: the ChatGPT saga unfolds. *Cureus.* 2023;15:e40135. [PMID: 37425598] doi:10.7759/cureus.40135
49. Ali SR, Dobbs TD, Hutchings HA, et al. Using ChatGPT to write patient clinic letters. *Lancet Digit Health.* 2023;5:e179-e181. [PMID: 36894409] doi:10.1016/S2589-7500(23)00048-1
50. Holmes J, Liu Z, Zhang L, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *arXiv.* Preprint posted online 1 April 2023. doi:10.48550/arXiv.2304.01938
51. 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. [PMID: 36812645] doi:10.1371/journal.pdig.0000198
52. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. *medRxiv.* Preprint posted online 26 February 2023. doi:10.1101/2023.02.21.23285886
53. Strong E, DiGiammarino A, Weng Y, et al. Performance of ChatGPT on free-response, clinical reasoning exams. *medRxiv.* Preprint posted online 29 March 2023. doi:10.1101/2023.03.24.23287731
54. Cross J, Robinson R, Devaraju S, et al. Transforming medical education: assessing the integration of ChatGPT into faculty workflows at a Caribbean medical school. *Cureus.* 2023;15:e41399. [PMID: 37426402] doi:10.7759/cureus.41399
55. Gupta R, Herzog I, Weisberger J, et al. Utilization of ChatGPT for plastic surgery research: friend or foe? *J Plast Reconstr Aesthet Surg.* 2023;80:145-147. [PMID: 37023599] doi:10.1016/j.bjps.2023.03.004
56. Gururangan S, Marasović A, Swayamdipta S, et al. Don't stop pretraining: adapt language models to domains and tasks. In: Jurafsky D, Chai J, Schluter N, et al, eds. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, virtual, 5-10 July 2020. Association for Computational Linguistics; 2020:8342-8360. Accessed at <https://aclanthology.org/2020.acl-main.740> on 4 August 2023.
57. Hutchinson B, Prabhakaran V, Denton E, et al. Social biases in NLP models as barriers for persons with disabilities. In: Jurafsky D, Chai J, Schluter N, et al, eds. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, virtual, 5-10 July 2020. Association for Computational Linguistics; 2020:5491-5501. Accessed at <https://aclanthology.org/2020.acl-main.487> on 4 August 2023.
58. Au Yeung J, Kraljevic Z, Luintel A, et al. AI chatbots not yet ready for clinical use. *Front Digit Health.* 2023;5:1161098. [PMID: 37122812] doi:10.3389/fdgth.2023.1161098
59. Karn SK, Ghosh R, Kusuma P, et al. shs-nlp at RadSum23: domain-adaptive pre-training of instruction-tuned LLMs for radiology report impression generation. *arXiv.* Preprint posted online 5 June 2023. doi:10.48550/arXiv.2306.03264
60. Wornow M, Xu Y, Thapa R, et al. The shaky foundations of large language models and foundation models for electronic health records. *NPJ Digit Med.* 2023;6:135. [PMID: 37516790] doi:10.1038/s41746-023-00879-8
61. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv.* Preprint posted online 22 May 2020. Revised 12 April 2021. doi:10.48550/arXiv.2005.11401
62. Goddard J. Hallucinations in ChatGPT: a cautionary tale for biomedical researchers. *Am J Med.* 2023;136:1059-1060. [PMID: 37369274] doi:10.1016/j.amjmed.2023.06.012
63. Hiesinger W, Zakka C, Chaurasia A, et al. Almanac: retrieval-augmented language models for clinical medicine. *Research Square.* Preprint posted online 2 May 2023. doi:10.21203/rs.3.rs-2883198/v1
64. Maynez J, Narayan S, Bohnet B, et al. On faithfulness and factuality in abstractive summarization. *arXiv.* Preprint posted online 2 May 2020. doi:10.48550/arXiv.2005.00661

65. Weidinger L, Mellor J, Rauh M, et al. Ethical and social risks of harm from language models. arXiv. Preprint posted online 8 December 2021. doi:10.48550/arXiv.2112.04359
66. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. Ann Biomed Eng. 2023;51:2629-2633. [PMID: 37284994] doi:10.1007/s10439-023-03272-4
67. Lukas N, Salem A, Sim R, et al. Analyzing leakage of personally identifiable information in language models. arXiv. Preprint posted online 1 February 2023. Revised 23 April 2023. doi:10.48550/arXiv.2302.00539
68. Zou A, Wang Z, Zico Kolter J, et al. Universal and transferable adversarial attacks on aligned language models. arXiv. Preprint posted online 27 July 2023. doi:10.48550/arXiv.2307.15043
69. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023;6:120. [PMID: 37414860] doi:10.1038/s41746-023-00873-0
70. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616:259-265. [PMID: 37045921] doi:10.1038/s41586-023-05881-4
71. Li C, Wong C, Zhang S, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. arXiv. Preprint posted online 1 June 2023. doi:10.48550/arXiv.2306.00890
72. Hu EJ, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. arXiv. Preprint posted online 17 June 2021. Revised 16 October 2021. doi:10.48550/arXiv.2106.09685
73. Tang X, Zong Y, Phang J, et al. Struc-Bench: are large language models really good at generating complex structured data? arXiv. Preprint posted online 16 September 2023. Revised 19 September 2023. doi:10.48550/arXiv.2309.08963

ANNALS INFORMATION


ANNALS AWARDS

*Personae Photography Prize:* Annals awards a \$500 prize for the best photograph submitted each year. Personae photographs are pictures that catch people in the context of their lives and that capture personality.




SCAN THE QR CODE WITH YOUR PHONE

*Ad Libitum Poetry Prize:* All poems published within 1 calendar year are automatically entered into our poetry prize contest. The winning poem is selected by a panel led by Dr. Michael Lacombe and 2 or 3 external judges. The prize for the winning poem is \$500.



SCAN THE QR CODE WITH YOUR PHONE

*Early Career Investigator Awards:* Annals and the American College of Physicians recognize excellence among internal medicine trainees and early career investigators with annual awards for original research and scholarly review articles published in Annals.



SCAN THE QR CODE WITH YOUR PHONE

**Author Contributions:** Conception and design: R. Daneshjou, H. Gui, J.A. Omiye, J. Zou.  
Analysis and interpretation of the data: H. Gui.  
Drafting of the article: R. Daneshjou, H. Gui, J.A. Omiye, S.J. Rezaei.  
Critical revision for important intellectual content: R. Daneshjou, H. Gui, J.A. Omiye, S.J. Rezaei, J. Zou.  
Final approval of the article: R. Daneshjou, H. Gui, J.A. Omiye, S.J. Rezaei, J. Zou.  
Obtaining of funding: R. Daneshjou.  
Administrative, technical, or logistic support: R. Daneshjou, H. Gui, J.A. Omiye.  
Collection and assembly of data: H. Gui.