MAYO CLINIC PROCEEDINGS: DIGITAL HEALTH

SPECIAL ARTICLE

Fine-Tuning Large Language Models for Specialized Use Cases

D.M. Anisuzzaman, PhD; Jeffrey G. Malins, PhD; Paul A. Friedman, MD; and Zachi I. Attia, PhD

Abstract

Large language models (LLMs) are a type of artificial intelligence, which operate by predicting and assembling sequences of words that are statistically likely to follow from a given text input. With this basic ability, LLMs are able to answer complex questions and follow extremely complex instructions. Products created using LLMs such as ChatGPT by OpenAI and Claude by Anthropic have created a huge amount of traction and user engagements and revolutionized the way we interact with technology, bringing a new dimension to human-computer interaction. Fine-tuning is a process in which a pretrained model, such as an LLM, is further trained on a custom data set to adapt it for specialized tasks or domains. In this review, we outline some of the major methodologic approaches and techniques that can be used to fine-tune LLMs for specialized use cases and enumerate the general steps required for carrying out LLM fine-tuning. We then illustrate a few of these methodologic approaches by describing several specific use cases of fine-tuning LLMs across medical subspecialties. Finally, we close with a consideration of some of the benefits and limitations associated with fine-tuning LLMs for specialized use cases, with an emphasis on specific concerns in the field of medicine.

 ■ Mayo Clin Proc Digital Health 2025;3(1):100184

Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN.

Large language models (LLMs), a specialized subset of artificial intelligence (AI), are designed to generate text through a process known as autoregression (often leading them to be termed autoregressive LLMs). These models operate by predicting and assembling sequences of words that are statistically likely to follow from a given text input, thereby enabling them to produce coherent and relevant sentences. The models can accept conversational input as text or via speech (using language recognition) and can generate outputs at various levels ranging from technical/professional to that of a high school education and more. They can summarize vast quantities of data, have access to unimaginably large volumes of information, and stand to make this available, easily, to the user. The public release of ChatGPT has opened the public's imagination and given a glimpse into an information-rich future.

These capabilities allow LLMs to perform a variety of general purpose tasks such as answering questions, completing sentences, and even generating entire articles. One of the breakthroughs that led to the creation of LLMs is the use of foundational models that process and comprehend natural language using deep learning methods. The 2 primary ideas of foundation models are self-supervised learning and scale. In self-supervision, instead of training a model to perform a task that requires explicit annotations, the model learns from the vast amounts of unlabeled data available, extracting patterns and understanding context without human intervention. In addition to being more scalable, self-supervised tasks can allow a model to anticipate a portion of the inputs, which makes the model richer and potentially more valuable than models trained on a more constrained label space. Once the model learns the foundational patterns of language, the same model can then be applied using transfer learning followed by fine-tuning, which enables the model to learn to perform more specific tasks using a smaller set of labeled samples. For scale, the era of the internet provides a nearly limitless amount of data[1] and, coupled with advances in computing power,

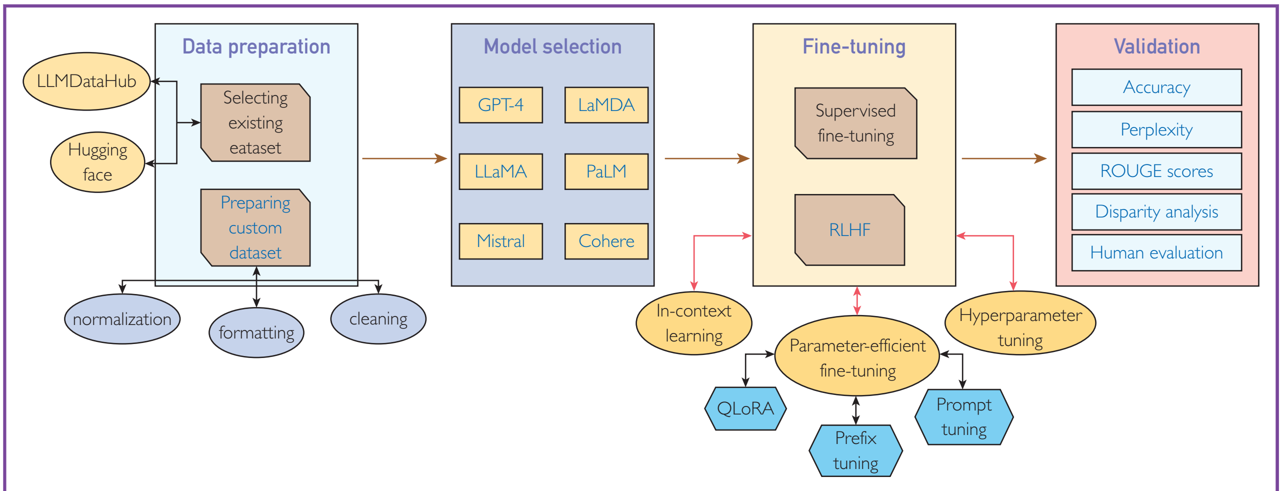


FIGURE 1. A general workflow of LLM fine-tuning for specialized use cases. LLM, large language model; QLoRA, quantized low-rank adaptation; RLHF, reinforcement learning from human feedback.

enables the training of models on an unprecedented scale using graphical processing units (GPUs). Together, these developments—enhanced by innovations such as the transformer model architecture[2]—have significantly propelled the capabilities and applications of LLMs. A general workflow of LLM fine-tuning for specialized use cases is shown in Figure 1.

Some existing LLMs to date are Alpaca,[3] BERT,[4] BLOOM,[5] Claude,[6] Cohere,[7] Ernie,[8] Falcon,[9] Flan,[10] Gemini,[11] Gemma,[12] GPT-3.5,[13] GPT-4,[14] LaMDA,[15] LLaMA,[16] Mistral,[17] MPT,[18] Orca,[19] PaLM 2,[20] Phi-1,[21] StableLM,[22] T5,[23] Vicuna,[24] and Zephyr.[25] All these models were developed to handle language-related tasks by different for-profit and nonprofit organizations such as Google, Meta, and Stanford. Although most of these models were created as general task models, some were developed for specialized tasks such as language translation, human-like chat, and code generation.

In addition to anticipating subsequent text, because models are trained with billions of tokens, many words map to multiple tokens (ie, they are represented by word vectors), enabling mathematical connections between multiple meanings of a term. For example, Paris will have connections to France, city, capital, and so on, so that the relationships between Paris and France and London and United kingdom may be used by an LLM. A limitation of LLMs is that after training is completed, a model no longer learns or acquires new information, and the information it was trained on may be general (such as Wikipedia), but not well-suited to a specific task. These limitations can be mitigated with fine-tuning to better sculpt an LLM to address a specific field (such as medicine or law), and retrieval augmented generation, which provides additional information that a model may use to address questions and which is particularly useful if that additional information was not included in the model's training.

In the domain of health care, a number of LLMs have been fine-tuned to perform tasks associated with preconsultation, diagnosis, management, and prediction of future medical outcomes, as well as medical education and medical writing.[26-28] LLMs specific to the medical domain include BioBERT,[29] BioGPT,[30] BioMistral,[31] ChatDoctor,[32] Clinical Camel,[33] DoctorGLM,[34] Med-Alpaca,[35] Med-PaLM,[36] Med-PaLM 2,[37] Med42-v2,[38] Meditron-70b,[39] OpenBioLLM-70B,[40] and PMC-LLaMA.[41] One particularly powerful use of models such as these is obtaining answers to questions rather than links to articles, with the caveat that using systems not designed to address medical questions may be

TABLE. Uses of LLMs in Medicine

Description of LLM medical uses	Strengths	Limitations	Example usage
Medical research assistance	Can quickly synthesize and summarize existing medical literature, helping researchers stay up-to-date with recent developments	May not have access to the most recent studies owing to training data cutoffs; could miss context or nuance in highly specialized areas	Documentation for clinical trials[43]
Clinical decision support	Provides support in diagnosing complex cases by suggesting possible diagnoses on the basis of symptoms and medical history	Relies on the data it was trained on, which may not include rare diseases or latest treatment modalities	Differentiation between abdominal pathologies[44]
Patient interaction automation	Handles routine inquiries from patients, such as explaining medical procedures and advising on medication schedules	May lack the empathetic nuances that human interaction provides; risk of miscommunication in complex scenarios	Answering cataract operation-related questions[45]
Medical education and training	Assists in the education of medical students and professionals by providing explanations, generating quizzes, and simulating patient cases	Might not perfectly mimic the unpredictability of real-life medical cases; information may become outdated	Interactive practice cases to evaluate medical reasoning[46]
Documentation and reporting	Helps in generating and organizing medical reports, thereby reducing the administrative burden on health care providers	Possible issues with accuracy and privacy concerns; needs constant verification	Generation of radiology reports on the basis of chest X-rays[47]
Treatment plan management	Suggests treatment plans on the basis of clinical guidelines and individual patient data	May not incorporate experiential learning or adapt to unconventional cases as effectively as a human would	Assistance with complex decision making for breast cancer care[48]
Support for remote areas	Provides medical information and support in remote areas where medical expertise is limited	Dependence on internet connectivity; may not handle local medical practices or nonstandard treatments well	Providing community health workers with contextually appropriate medical knowledge[49]

The first 3 columns of this table were generated by ChatGPT 4.0 on May 6, 2024. The final column with example usages was added by the authors.

inaccurate.[42] LLMs can potentially be used for any task that requires reading text, and summarizing it, or extracting pertinent information. Examples of data extraction uses could include reviewing of medical records to create a discharge summary, identifying and summarizing all risks for stroke in a patient with atrial fibrillation, or determining preoperative surgical risk using standardized scoring criteria. A list of potential uses of LLMs in medicine along with specific examples is provided in Table.[43-49]

In the following sections, we first outline some of the major approaches and techniques for fine-tuning LLMs in the medical domain and touch on retrieval augmented generation. Then, we describe specialized use cases in which LLMs have been fine-tuned for medical applications across various medical subspecialties. After this, we close with a consideration of some of the benefits and key limitations associated with fine-tuning LLMs in the medical domain.

FINE-TUNING METHODOLOGY

Fine-tuning is a process in which a pretrained model is adapted for particular tasks or domains by continuing to train the model using only a domain-specific data set that is different than the original data set used to the train the base model. Various fine-tuning strategies and approaches are used to adjust the model parameters to a specific need. Some fine-tuning approaches are briefly described in this article.

Supervised Fine-Tuning

With this approach, every input data point is linked to a label, and the model is trained on a task-specific labeled data set. The model learns to modify its parameters to anticipate these labels as precisely as possible. Some supervised fine-tuning techniques are as follows:

1. Transfer learning: In this approach, a model is first initialized with saved weights from a model pretrained on a large, general data set and then is subsequently trained with limited task-specific data. Weights refer to the learned parameters of a model that has been trained on a large data set for a specific task, which represent the knowledge the model has gained during its training process, encapsulating features and patterns relevant to the task it was originally trained on.
2. Multitask learning: Here, models are fine-tuned on numerous related tasks, taking advantage of their similarities and differences, in order to maximize performance. For example, with a CNN model trained on a generic large data set (eg, KINETICS400), one can perform some specific tasks (eg, estimating left ventricular ejection fraction, patient age, and patient sex from an echocardiogram) with a much smaller data set by leveraging the generic features the model learned from the large data set.
3. Instruction-tuning: Instruction-tuning involves fine-tuning a pretrained LLM to follow specific task instructions, such as translation, summarization, or question answering. For example, in translation, the model is trained on examples in which each input includes an instruction like "Translate the following sentence from English to French," followed by an English sentence and its French translation. After fine-tuning, the model learns to follow translation instructions and can generalize to translate new sentences.

Reinforcement Learning From Human Feedback

This method uses the knowledge of human evaluators; in addition, it also allows the model to adjust and develop in response to real-world input, resulting in enhanced and more efficient applications. Some standard reinforcement learning from human feedback (RLHF) techniques are as follows:

1. Reward modeling: In this method, the model generates multiple potential outputs or actions, which are subsequently assessed by human evaluators who assign a ranking or rating on the basis of their quality. The model uses these human-provided assessments to generate predictions and adapt its behavior to optimize the anticipated rewards.
2. Proximal policy optimization: Proximal policy optimization (PPO) modifies the language model's policy to maximize the expected reward. A policy refers to the strategy or set of rules that a reinforcement learning agent uses to make decisions in an environment. For example, in PPO, the policy determines how a robotic arm should move to pick up objects on the basis of visual inputs from a camera. PPO's primary goal is to make policy improvements whereas ensuring the modifications do not deviate too much from the previous policy. To achieve this balance, the policy update process introduces a constraint that prohibits detrimental large updates although permitting advantageous minor updates. Compared with other reinforcement learning techniques, PPO is more reliable and effective.
3. Comparative ranking: In this method, the model produces several outputs or actions, which human investigators then rank according to compatibility or quality. The model then modifies its behavior to generate higher-ranked outputs. This method provides relative and better feedback to the model by ranking multiple outputs rather than individual outputs.
4. Preference feedback: This technique involves the model generating several outputs and human experts selecting among them, leading the model to modify its behavior accordingly. This method is useful when assigning a numeric value (reward) to an output is difficult. It is an effective method of fine-tuning the model in practical applications.

FINE-TUNING PIPELINE

To carry out a fine-tuning process for a specialized use case, there are several generic steps that include the following[50]:

1. Data preparation: Data set preparation for LLM model fine-tuning is entirely task-specific. In general, the model must be presented with some blocks of text. Many data sets are available to fine-tune an LLM.[51-53] One must follow the instructions given for each data set to prepare the data for fine-tuning. For custom data sets, depending on the task, the data set preparation may include data cleaning, normalizing for missing values, and formatting the text to align with the model's input requirements.
2. Selecting the appropriate pretrained model: There are several LLMs to date (BERT, Cohere, GPT-4, LLaMA, Mistral, etc; access to nonopen source models will require working with the owners of the models), and choosing the appropriate one that complies with the demands of the target task is essential. For fine-tuning an LLM model for a specific data set or task, one must have a good grasp of the model architecture and the input and output requirements. Depending on the available resources, the model weights and number of parameters should be considered when selecting a model. Finally, the performance of the model on the relevant target task should be considered during model selection.
3. Fine-tuning the model: LLM fine-tuning also includes basic hyperparameter tuning, adjustments of the learning rate, batch size, regularization, optimizer, number of epochs, and so on. As LLMs are trained on vast amounts of data, overfitting a small data set for a specific task could be a likely event. Careful tuning of the hyperparameters guarantees that the model learns efficiently and does not overfit when applied to new data.
4. Validation: Validation of an LLM is complex. In predictive AI, a specific input and output are expected. For example, a neural network might assess a medical image such as an electrocardiogram, or a medical video such as an echocardiogram, and be tasked with determining the ejection fraction (heart pump strength). The ejection fraction can be manually measured by a human to assess the performance of the AI tool. In contrast, generative AI, such as LLMs, generate new text or images, the performance of which may be harder to grade. If asked to create a poem, how does one assess the quality of it?

In general practice, there are 2 types of validation to perform: (1) internal validation, which is used to select the best model, monitor the model's learning process, and callback and stop model training with certain criteria; and (2) validation on a holdout test set to evaluate model performance for real-world applications. Although, in general, some commonly used metrics for model evaluation include accuracy, area under the curve, precision, recall, and so on, LLM model evaluation may require some careful task-specific metric selection.[54] Some key performance metrics used in LLM evaluation are as follows:

- Accuracy: measures the model's ability to produce correct responses to prompts.
- Perplexity: measures uncertainty in predicting the next token.
- ROUGE scores: compares an LLM's output with a set of reference summaries.
- Diversity: evaluates the variety of responses generated.
- Disparity analysis: identifies and mitigates biases within model responses.
- Coh-Metrix: analyzes logical consistency and clarity over longer stretches of text.
- Human evaluation: subjective assessment by human judges.

In medical LLM development, in cases for which models are fine-tuned for clinical prediction tasks in which the ground truth labels are well defined (eg, predicting discharge events), evaluation typically involves statistical performance metrics like accuracy, precision, and so on. For more generative tasks in which the ground truth labels are not well defined (eg, medical report summarization), human or domain expert evaluation is crucial to ensure that model outputs are clinically accurate and safe for real-world applications. For example, Singhal et al[36] developed Med-PaLM for answering medical questions and had clinicians review outputs to ensure that the responses were medically sound and factually accurate. Similarly, Serapio et al[55] fine-tuned LLMs for generating radiological impressions from chest computed tomography scans and had model outputs assessed by board-certified radiologists.

Beyond these general approaches, a number of specific techniques can be applied to fine-tune LLMs for specialized use cases. Some example techniques include the following:

1. In-context learning: In this approach, a pretrained LLM is induced to perform a task using prompted examples. An example of this is few-shot learning, which involves giving the model a few shots or instances to learn a new task during inference. Few-shot learning aims to direct the model's predictions by providing examples and context specifically in the prompt but importantly does not involve gradient-based training.[56]
2. Hyperparameter tuning: This is a straightforward method that consists of manually modifying basic hyperparameters (ie, learning rate, batch size, optimizer, and number of epochs) of the model until the desired performance is obtained. This changes how the model learns; that is, how fast it learns, how to decide when training is completed, and so on.
3. Parameter-efficient fine-tuning: Parameter-efficient fine-tuning (PEFT) is an efficient technique in which only a small portion of the parameters of an LLM are selectively modified during fine-tuning, typically by adding new layers or modifying existing ones in a task-specific manner. This method drastically lowers computational and storage needs although keeping performance comparable with complete fine-tuning. Some PEFT techniques are low-rank adaptation,[57] quantized low-rank adaptation (QLoRA),[58] Prefix tuning, and Prompt tuning.

Quantized low-rank adaptation is a very popular technique used for LLM fine-tuning owing to its power of using much smaller amounts of memory than a full fine-tuning approach with the price of sacrificing some performance. For example, a full fine-tuning of the LLaMA 65B parameter model requires more than 780 GB of GPU memory, whereas using the QLoRA technique requires only 48GB of GPU memory.[58] This powerful technique is on the basis of these following highly technical ingredients:

i. 4-bit NormalFloat representation of model parameters, whereas typically, parameters of trained models are stored in a 32-bit format. This technique divides model parameters into equally-sized buckets instead of equally-spaced buckets.
ii. Double quantization is a method that quantizes the quantization constants. In general, quantization converts datatypes with a larger number of bits to fewer bits (eg, FP32 to 8-bit Integers). Quantized low-rank adaptation uses the blockwise quantization technique that requires more memory than standard quantization but reduces bias significantly, thus retaining good performance.
iii. Low-rank adaptation,[57] which freezes the pretrained model weights and injects trainable rank decomposition matrices into each layer of the transformer architecture, greatly reduces the number of trainable parameters for downstream tasks.

In simpler terms, low-rank adaptation finds a more compressed version of the LLM weights and updates those weights. Although the compression may lose some data, under the assumption a lot of the model weights are redundant, leading only to a small decrease in performance relative to savings in memory and required compute power.

4. Retrieval augmented generation: Retrieval augmented generation (RAG) is a technique that combines the capabilities of neural language models with information retrieval systems to enhance the generation of contextually rich and accurate responses. In RAG, when a query is received, the model first uses a retrieval system to fetch relevant documents or snippets from a large corpus, such as a database of scientific literature. These retrieved texts are then fed into a generative model, typically a transformer-based neural network, which integrates the retrieved information with its pretrained knowledge to produce a coherent and informed response. This approach is particularly useful in domains

where accuracy and specificity are critical, such as scientific research or technical support, because it allows the model to base its answers on up-to-date and source-specific data, providing citations and grounding its responses in existing literature. A comparison of the outputs generated by presenting the same medical question to a search engine, LLM, and RAG-based system is shown in Figure 2.

SPECIALIZED USE CASES IN MEDICINE

Across many of the major subspecialties of medicine, LLMs are being fine-tuned to address specific issues, and practitioners are posing questions about how fine-tuned LLMs could revolutionize their fields. These subspecialties include but are not limited to cardiology,[59-61] dermatology,[62] digital pathology,[63] gastroenterology and hepatology,[64-66] hematology,[67] neurology,[68-70] obstetrics and gynecology,[71,72] oncology,[73-75] ophthalmology,[76] orthopedics,[77] pediatrics,[78,79] psychiatry,[80-82] radiology,[83,84] operation,[85,86] and urology.[87,88] Although there are nuances according to specific subspecialties, many practitioners highlight the potential of fine-tuned LLMs to aid clinicians in areas such as clinical decision support, treatment planning, and patient consultation, as well as alleviate administrative burden associated with tasks such as generating clinical notes, discharge reports, and medical billing. At the same time, many are concerned about the ethical, legal, and social implications of using such models.

In the subsequent section, we review some of these concerns. Before that, below we highlight a few specific methodologies and use cases that illustrate the general framework outlined in the previous sections.

Using RLHF to Fine-Tune LLMs in Medicine

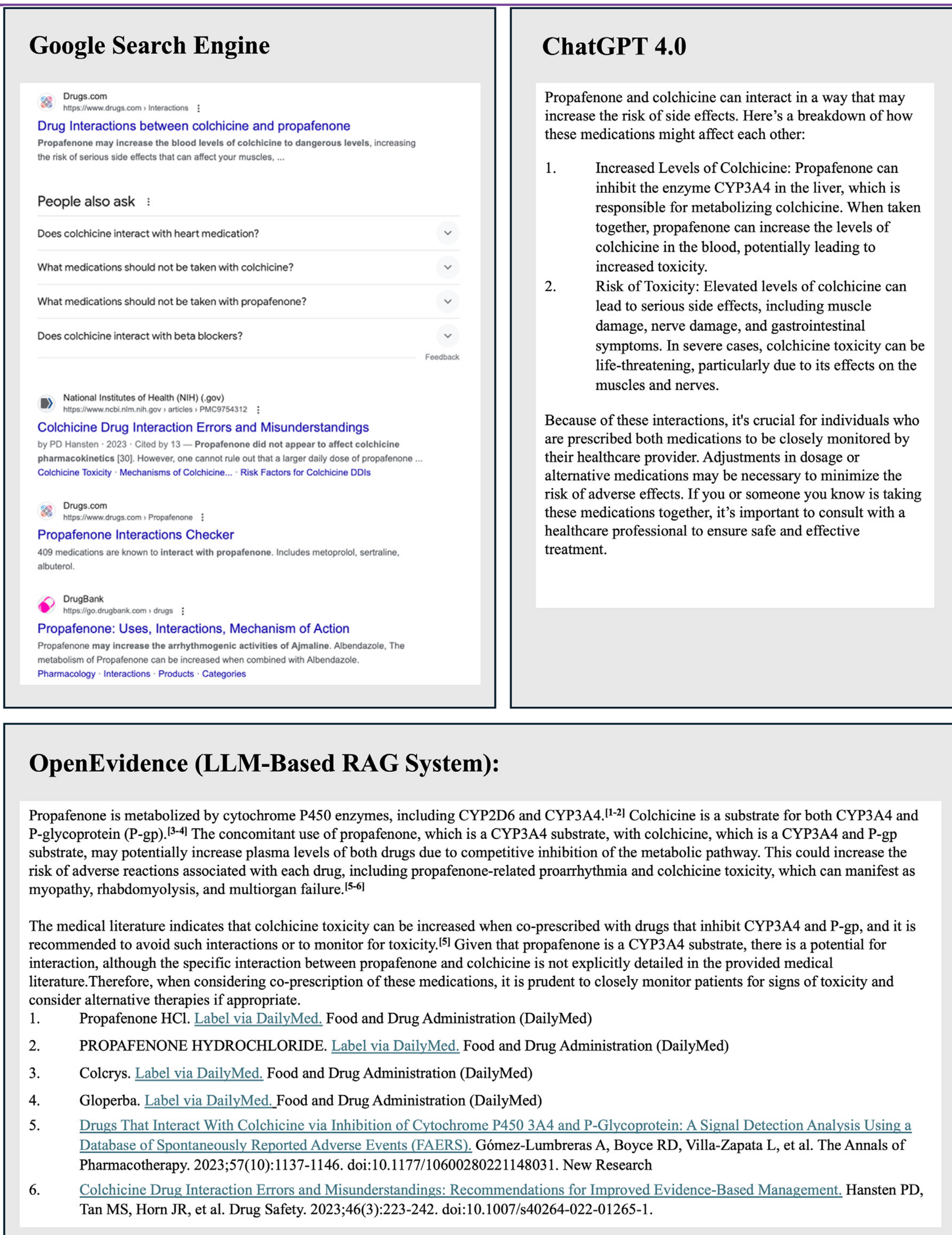
Mukherjee et al[89] developed a constellation system called Polaris, which was composed of several agents. Their primary agent (focused on patient-friendly conversation) was developed in 3 stages: general instruction-tuning, conversation and agent tuning, and RLHF. The RLHF step was performed by registered nurses, who gave preference feedback on multiple responses. Zhao et al[90] developed Aqulia-Med, a bilingual medical LLM, using supervised fine-tuning and RLHF to tackle medical challenges. It was trained on large-scale Chinese and English medical data sets, with RLHF further aligning the model to improve performance in medical dialogs and multiple-choice questions.

Integration of LLMs into Electronic Health Record Systems

Several LLMs have been applied to electronic health record (EHR) systems, providing benefits such as generating patient summaries from EHRs, assisting health care providers with more efficient decision making, named entity recognition, medical note summarization, and predictive diagnosis.[91] Zhang et al[92] investigated the application of LLM fine-tuning to EHR audit log data for clinical prediction tasks, with a focus on discharge predictions. Cui et al[93] evaluated the zero-shot and few-shot performance of LLMs on EHR-based disease prediction tasks and proposed a novel approach that leverages collaborative LLM agents to enhance predictive performance. Li et al[94] fine-tuned an LLM named LlamaCare and evaluated it on various clinical tasks, such as generating discharge summaries, predicting mortality and length of stay, and more.

Generation of Echocardiography Reports to Streamline Workflows

Echocardiography is one of the most widely used imaging techniques for gaining insights into the structure and function of the heart. A typical echocardiography report includes numerous measurements as well as text-based statements or findings. These findings are summarized by a clinician to give an overall set of final impressions for the study. This is a time-consuming and error-prone process. To address this issue, Chao et al[95] leveraged several open-source LLMs to generate echocardiography reports using either zero-shot learning (for Flan-T5, Med-Alpaca, Llama-2, and Zephyr) or QLoRA fine-tuning (Llama-2 and Zephyr). Using a training data set of 95,506 echocardiography reports, the authors observed that EchoGPT, which is a Llama-2 model trained using instruction fine-tuning with QLoRA, outperformed other LLMs on critical performance metrics. In addition, when 4 echocardiography board-certified cardiologists were asked to rate reports generated by EchoGPT for 30 randomly selected cases,



Google Search Engine

Drugs.com
https://www.drugs.com › Interactions

Drug Interactions between colchicine and propafenone

Propafenone may increase the blood levels of colchicine to dangerous levels, increasing the risk of serious side effects that can affect your muscles, ...

People also ask

Does colchicine interact with heart medication?

What medications should not be taken with colchicine?

What medications should not be taken with propafenone?

Does colchicine interact with beta blockers?

Feedback

National Institutes of Health (NIH) (.gov)
https://www.ncbi.nlm.nih.gov › articles › PMC9754312

Colchicine Drug Interaction Errors and Misunderstandings

by PD Hansten · 2023 · Cited by 13 — **Propafenone did not appear to affect colchicine pharmacokinetics** [30]. However, one cannot rule out that a larger daily dose of propafenone ...
Colchicine Toxicity · Mechanisms of Colchicine... · Risk Factors for Colchicine DDIs

Drugs.com
https://www.drugs.com › Propafenone

Propafenone Interactions Checker

409 medications are known to **interact with propafenone**. Includes metoprolol, sertraline, albuterol.

DrugBank
https://go.drugbank.com › drugs

Propafenone: Uses, Interactions, Mechanism of Action

Propafenone **may increase the arrhythmogenic activities of Ajmaline**. Albendazole, The metabolism of Propafenone can be increased when combined with Albendazole.
Pharmacology · Interactions · Products · Categories

ChatGPT 4.0

Propafenone and colchicine can interact in a way that may increase the risk of side effects. Here's a breakdown of how these medications might affect each other:

1. Increased Levels of Colchicine: Propafenone can inhibit the enzyme CYP3A4 in the liver, which is responsible for metabolizing colchicine. When taken together, propafenone can increase the levels of colchicine in the blood, potentially leading to increased toxicity.
2. Risk of Toxicity: Elevated levels of colchicine can lead to serious side effects, including muscle damage, nerve damage, and gastrointestinal symptoms. In severe cases, colchicine toxicity can be life-threatening, particularly due to its effects on the muscles and nerves.

Because of these interactions, it's crucial for individuals who are prescribed both medications to be closely monitored by their healthcare provider. Adjustments in dosage or alternative medications may be necessary to minimize the risk of adverse effects. If you or someone you know is taking these medications together, it's important to consult with a healthcare professional to ensure safe and effective treatment.

OpenEvidence (LLM-Based RAG System):

Propafenone is metabolized by cytochrome P450 enzymes, including CYP2D6 and CYP3A4.[1-2] Colchicine is a substrate for both CYP3A4 and P-glycoprotein (P-gp).[3-4] The concomitant use of propafenone, which is a CYP3A4 substrate, with colchicine, which is a CYP3A4 and P-gp substrate, may potentially increase plasma levels of both drugs due to competitive inhibition of the metabolic pathway. This could increase the risk of adverse reactions associated with each drug, including propafenone-related proarrhythmia and colchicine toxicity, which can manifest as myopathy, rhabdomyolysis, and multiorgan failure.[5-6]

The medical literature indicates that colchicine toxicity can be increased when co-prescribed with drugs that inhibit CYP3A4 and P-gp, and it is recommended to avoid such interactions or to monitor for toxicity.[5] Given that propafenone is a CYP3A4 substrate, there is a potential for interaction, although the specific interaction between propafenone and colchicine is not explicitly detailed in the provided medical literature.Therefore, when considering co-prescription of these medications, it is prudent to closely monitor patients for signs of toxicity and consider alternative therapies if appropriate.

1. Propafenone HCl. Label via DailyMed. Food and Drug Administration (DailyMed)
2. PROPAFENONE HYDROCHLORIDE. Label via DailyMed. Food and Drug Administration (DailyMed)
3. Colcrys. Label via DailyMed. Food and Drug Administration (DailyMed)
4. Gloperba. Label via DailyMed. Food and Drug Administration (DailyMed)
5. Drugs That Interact With Colchicine via Inhibition of Cytochrome P450 3A4 and P-Glycoprotein: A Signal Detection Analysis Using a Database of Spontaneously Reported Adverse Events (FAERS). Gómez-Lumbreras A, Boyce RD, Villa-Zapata L, et al. The Annals of Pharmacotherapy. 2023;57(10):1137-1146. doi:10.1177/10600280221148031. New Research
6. Colchicine Drug Interaction Errors and Misunderstandings: Recommendations for Improved Evidence-Based Management. Hansten PD, Tan MS, Horn JR, et al. Drug Safety. 2023;46(3):223-242. doi:10.1007/s40264-022-01265-1.

FIGURE 2. Comparison of responses using different systems to the same question: "What are the interactions between propafenone and colchicine?"

the generated reports were rated similarly to reports generated by cardiologists for these same cases (in completeness, conciseness, correctness, and clinical utility). On the basis of these results, the authors argue that EchoGPT could be used as a copilot for report generation, which would allow for considerable streamlining of the echocardiography report workflow. With that said, the authors stress that draft reports generated by EchoGPT should still be reviewed and approved by clinicians, noting that some hallucinations were observed in reports generated by EchoGPT (albeit not as many as were observed for zero-shot learning).

Identifying Eligible Patients for Clinical Trials

Randomized clinical trials are a cornerstone of medical research, yet it can be a challenge to identify patients who meet all inclusion and exclusion criteria for a clinical trial. To leverage the power of LLMs to assist with participant recruitment for clinical trials, Guan et al[96] developed CohortGPT, built on ChatGPT and GPT-4. CohortGPT can take input text from unstructured or semistructured data, such as clinical notes and radiology reports, to designate disease labels associated with the input text. To develop this model, the authors made use of a technique called chain-of-thought (CoT) prompting, a type of in-context learning that guides LLMs to learn task-specific logical chains, which detail how correct answers are deduced from given information. Using the CoT technique in conjunction with reinforcement learning, Guan et al[96] trained a policy model to dynamically select CoT samples. They then presented these CoT samples to a prompt model alongside knowledge graphs, which can be thought of as rules detailing the relationships between different concepts, such as that cardiomegaly is a type of heart disease or that scoliosis is a type of spine disease. Using thousands of publicly available radiology reports in the Indiana chest X-ray collection[97] and MIMIC-CXR[98] data sets, Guan et al[96] found that CohortGPT can reliably classify report text as being associated with specific disease labels. On the basis of these results, the authors argue that CohortGPT can be useful not only for patient recruitment for clinical trials but also for other medical applications such as diagnosis and treatment optimization. Furthermore, although CohortGPT was built on ChatGPT and GPT-4, the model can be implemented in any open-source LLM.

BENEFITS AND LIMITATIONS OF FINE-TUNING LLMS FOR SPECIALIZED USE CASES

To build a reliable real-world LLM-based application, fine-tuning is a necessary and crucial step because it fills in the gap between general knowledge and domain-specific expertise for that application. Some benefits of LLM fine-tuning are the following:

1. Domain-specific knowledge: general purpose LLMs may not have enough domain-specific knowledge.
2. Specific task optimization: general purpose LLMs can be optimized for specific tasks (eg, health report summarization and disease detection from a report).
3. Data efficiency: fine-tuning works well with smaller quantities of labeled data because it involves using pretrained LLM(s) trained on huge data sets.
4. Better performance: fine-tuning often leads to improved performance because the model learns domain-specific knowledge to perform relevant tasks although preserving out-of-domain knowledge.
5. Resource efficiency: fine-tuning requires less resources in terms of time and memory than training a general purpose LLM from scratch.

With that said, there are several critical limitations to LLMs to consider when fine-tuning for specialized tasks.[99,100] A few of these are as follows:

1. Hallucinations: these refer to situations in which model output contains inaccurate or nonfactual information.[42,99] In the medical domain for example, these could consist of findings that are not actually present in a study report. Addressing hallucinations could involve processes such as inducing a model to provide a reasoning process or confidence score associated with model output. For example, the Medical Domain Hallucination Test (Med-HALT) has been designed to evaluate and reduce hallucinations in the medical

domain and includes metrics for hallucinations associated with reasoning and memory.[101]

2. Legal and safety concerns: for example, in the medical domain, the data to be used for fine-tuning may contain sensitive patient information that needs to be safeguarded. In addition, if model output is used to guide treatment decisions for patients, incorrect output (such as hallucinations) could be harmful. This is why authors such as Chao et al[95] emphasize the critical need for human review of model output. In addition, cybersecurity measures such as the use of pseudonyms can enhance the privacy and security of patient data.[102]
3. Biases in training data sets: fine-tuned LLMs can inherit biases from the pretrained models on which they are built, and there is a critical need to use techniques that mitigate this bias.[103,104] In medicine, this bias has the potential to exacerbate health inequities if not addressed.[105] Some techniques for mitigating bias include prompt engineering, debiasing algorithms, and continuous monitoring of model performance.[106]
4. Lack of domain-specific data: depending on the extent to which a specific use case is specialized, there may not be sufficient quantities of domain-specific data to fine-tune an LLM using certain approaches. Here, techniques such as in-context learning or PEFT may be more appropriate than full fine-tuning.
5. Data leakage: many of the pretrained LLMs do not report which data were used for training, so if open data sets are used for fine-tuning, these data may have already been used for training the base model. This can lead to data leakage from the validation set to the model, resulting in overly optimistic performance. Addressing this concern will involve greater transparency on the part of developers when describing training data sets and careful selection of pretrained LLMs that provide information about the source, quality, and quantity of training data.[102]

CONCLUSION

Large language models are poised to transform medicine. In written form or verbally, they can summarize vast amounts of information, may prevent important pieces of information from being missed, and can meaningfully tap into vast stores of literature to inform clinicians at the point of care when meeting with a patient. However, much remains unproven including how to ensure the information is reliable, privacy is preserved, and answers are tuned to usefully guide medical professionals.

POTENTIAL COMPETING INTERESTS

Drs Anisuzzaman, Malins, Friedman, and Attia have invented algorithms licensed to Ultra-Sight and may benefit from algorithm commercialization via Mayo Clinic. None of these relations with industry are related in any way to the content of the current submission. Given their role as Editorial Board Members, Drs. Attia and Friedman had no involvement in the peer-review of this article and have no access to information regarding its peer-review. Drs Friedman and Attia report multiple patents owned by Mayo for AI ECG and stock or stock options in Anumana and XAI Health.

SUPPLEMENTAL ONLINE MATERIAL

Supplemental material can be found online at https://www.mcpdigitalhealth.org/. Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data.

Abbreviations and Acronyms: **AI,** artificial intelligence; **CoT,** chain-of-thought; **GPU,** graphical processing unit; **LLM,** large language model; **PEFT,** parameter-efficient fine-tuning; **PPO,** proximal policy optimization; **QLoRA,** quantized low-rank adaptation; **RAG,** retrieval augmented generation; **RLHF,** reinforcement learning from human feedback

Publication dates: Received for publication August 2, 2024; revisions received November 6, 2024; accepted for publication November 18, 2024.

Correspondence: Address to Zachi I. Attia, PhD, Department of Cardiovascular Medicine, Mayo Medical School, Artificial Intelligence in Cardiology, Mayo Clinic, 200 1st Street SW, Rochester, MN 55905 (attia.itzhak@mayo.edu).

ORCID

D.M. Anisuzzaman: https://orcid.org/0000-0001-8068-2571; Zachi I. Attia: https://orcid.org/0000-0002-9706-7900

REFERENCES

1. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models Preprint. Posted online August 16, 2021. arXiv 210807258. https://doi.org/10.48550/arXiv.2108.07258.
2. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inform Process Syst*. 2017;30.
3. Taori R, Gulrajani I, Zhang T, et al. *Stanford Alpaca: An Instruction-Following Llama Model*. Stanford University; 2023.
4. Devlin J, Chang M-W, Lee K, Toutanova K. *Bert: pre-training of deep bidirectional transformers for language understanding* Preprint. Posted online October 11, 2018. arXiv 181004805. https://doi.org/10.48550/arXiv.1810.04805.
5. Le Scao T, Fan A, Akiki C, et al. *Bloom: A 176b-parameter open-access multilingual language model* Preprint. Posted online November 9, 2022. arXiv 2211.05100. https://doi.org/10.48550/arXiv.2211.05100.
6. Anthropic. Introducing Claude. https://www.anthropic.com/news/introducing-claude. Accessed April 24, 2024.
7. Cohere. Cohere: the leading enterprise AI platform. https://cohere.com/. Accessed April 24, 2024.
8. BaiduResearch. ERNIE Bot: Baidu's knowledge-enhanced large language model built on full AI stack technology. http://research.baidu.com/Blog/index-view?id=183. Accessed April 25, 2024.
9. ZXhang YX, Haxo YM, Mat YX. Falcon LLM: a new frontier in natural language processing. *AC Investment Res J*. 2023; 220(44).
10. GoogleResearch. Introducing FLAN: more generalizable language models with instruction fine-tuning. https://research.google/blog/introducing-flan-more-generalizable-language-models-with-instruction-fine-tuning/. Accessed April 25, 2024.
11. Gemini Team Google, Anil R, Borgeaud S, Alayrac J-B, et al. *Gemini: a family of highly capable multimodal models* Preprint. Posted online December 19, 2023. arXiv 231211805. https://doi.org/10.48550/arXiv.2312.11805.
12. Gemma Team, Mesnard T, Hardin C, Dadashi R, et al. *Gemma: Open models based on gemini research and technology* Preprint. Posted online March 13, 2021. arXiv 240308295. https://doi.org/10.48550/arXiv.2403.08295.
13. OpenAI. Models—GPT 3.5 Turbo. https://platform.openai.com/docs/models/gpt-3-5-turbo. Accessed April 25, 2024.
14. OpenAI Achiam J, Adler S, Agarwal S, et al. *GPT-4 technical report* Preprint. Posted online March 15, 2023. arXiv 230308774. https://doi.org/10.48550/arXiv.2303.08774t.
15. Thoppilan R, De Freitas D, Hall J, et al. *LaMDA: language models for dialog applications* Preprint. Posted online January 20, 2022. arXiv 220108239. https://doi.org/10.48550/arXiv.2201.08239.
16. Touvron H, Lavril T, Izacard G, et al. *Llama: Open and efficient foundation language models* Preprint. Posted online February 27, 2023. arXiv 230213971. https://doi.org/10.48550/arXiv.2302.13971.
17. Jiang AQ, Sablayrolles A, Mensch A, et al. *Mistral 7B* Preprint. Posted online October 10, 2023. arXiv 231006825. https://doi.org/10.48550/arXiv.2310.06825.
18. HuggingFace. MPT. https://huggingface.co/docs/transformers/main/model_doc/mpt. Accessed April 25, 2024.
19. KDnuggets. Orca LLM: simulating the reasoning processes of ChatGPT. https://www.kdnuggets.com/2023/06/orca-llm-reasoning-processes-chatgpt.html. Accessed April 25, 2024.
20. Anil R, Dai AM, Firat O, et al. *Palm 2 technical report* Preprint. Posted online May 17, 2023. arXiv 230510403. https://doi.org/10.48550/arXiv.2305.10403.
21. Gunasekar S, Zhang Y, Aneja J, et al. *Textbooks are all you need* Preprint. Posted online June 20, 2023. arXiv 230611644. https://doi.org/10.48550/arXiv.2306.11644.
22. Bellagente M, Tow J, Mahan D, et al. *Stable LM 2 1.6 B technical report* Preprint. Posted online February 27, 2024. arXiv 240217834. https://doi.org/10.48550/arXiv.2402.17834.
23. Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res*. 2020;21(140):1-67.
24. Chiang W-L, Li Z, Lin Z, et al. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. https://vicuna.lmsys.org. Accessed April 14, 2023.
25. Tunstall L, Beeching E, Lambert N, et al. Zephyr: direct distillation of lm alignment Preprint. Posted online October 25, 2023. arXiv 231016944. https://doi.org/10.48550/arXiv.2310.16944.
26. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. *Health Care Sci*. 2023;2(4):255-263. https://doi.org/10.1002/hcs2.61.
27. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. https://doi.org/10.1038/s41591-023-02448-8.
28. Kraljevic Z, Bean D, Shek A, et al. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *Lancet Digit Health*. 2024;6(4):e281-e290. https://doi.org/10.1016/S2589-7500(24)00025-6.
29. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240. https://doi.org/10.1093/bioinformatics/btz682.
30. Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform*. 2022;23(6):bbac409. https://doi.org/10.1093/bib/bbac40.
31. Labrak Y, Bazoge A, Morin E, Gourraud P-A, Rouvier M, Dufour R. BioMistral: a collection of open-source pretrained large language models for medical domains Preprint. Posted online February 15, 2024. arXiv 240210373. https://doi.org/10.18653/v1/2024.findings-acl.348.
32. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. *Cureus*. 2023;15(6):e40985. https://doi.org/10.7759/cureus.40895.
33. Toma A, Lawler PR, Ba J, Krishnan RG, Rubin BR, Wang B. Clinical Camel: an open expert-level medical language model with dialogue-based knowledge encoding Preprint. Posted online May 19, 2023. arXiv 230512031. https://doi.org/10.48550/arXiv.2305.12031.
34. Xiong H, Wang S, Zhu Y, et al. *Doctorglm: fine-tuning your chinese doctor is not a herculean task* Preprint. Posted online April 3, 2023. arXiv 230401097. https://doi.org/10.48550/arXiv.2304.01097.
35. Han T, Adams LC, Papaioannou J-M, et al. *MedAlpaca—an open-source collection of medical conversational AI models and training data* Preprint. Posted online October 4, 2023. arXiv 230408247. https://doi.org/10.48550/arXiv.2304.08247.
36. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. https://doi.org/10.1038/s41586-023-06291-2.
37. Singhal K, Tu T, Gottweis J, et al. *Towards expert-level medical question answering with large language models* Preprint. Posted online May 16, 2023. arXiv 230509617. https://doi.org/10.48550/arXiv.2305.09617.
38. Christophe C, Kanithi PK, Raha T, Khan S, Pimentel MAF. *Med42-v2: a suite of clinical LLMs* Preprint. Posted online August 12, 2024. arXiv 240806142. https://doi.org/10.48550/arXiv.2408.06142.
39. Chen Z, Cano AH, Romanou A, et al. *MEDITRON-70b: scaling medical pretraining for large language models* Preprint. Posted online November 27, 2023. arXiv 231116079. https://doi.org/10.48550/arXiv.2311.16079.

40. Pal A, Sankarasubbu M. OpenBioLLMs: advancing open-source large language models for healthcare and life sciences. Hugging Face. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B. Accessed September 30, 2024.
41. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: toward building open-source language models for medicine. *J Am Med Inform Assoc*. 2024;31(9):1833-1843. https://doi.org/10.1093/jamia/ocae045.
42. Siontis KC, Attia ZI, Asirvatham SJ, Friedman PA. ChatGPT hallucinating: can it get any more humanlike? *Eur Heart J*. 2024;45(5): 321-323 https://doi.org/10.1093/eurheartj/ehad766.
43. Markey N, El-Mansouri I, Rensonnet G, van Langen C, Meier C. From RAGs to riches: using large language models to write documents for clinical trials Preprint. Posted online February 26, 2024. arXiv 240216406. https://doi.org/10.48550/arXiv.2402.16406.
44. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*. 2024;30(9):2613-2622. https://doi.org/10.1038/s41591-024-03097-1.
45. Ramjee P, Sachdeva B, Golechha S, et al. CataractBot: An LLM-Powered Expert-in-the-Loop Chatbot for Cataract Patients Preprint. Posted online February 7, 2024. arXiv 240204620. https://doi.org/10.48550/arXiv.2402.04620.
46. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: applications and implications. *JMIR Med Educ*. 2023;9:e50945. https://doi.org/10.2196/50945.
47. Wang Z, Liu L, Wang L, Zhou L. R2gengpt: Radiology report generation with frozen LLMs. *Meta-Radiology*. 2023;1(3): 100033. https://doi.org/10.1016/j.metrad.2023.100033.
48. Griewing S, Knitza J, Boekhoff J, et al. Evolution of publicly available large language models for complex decision-making in breast cancer care. *Arch Gynecol Obstet*. 2024;310(1):537-550. https://doi.org/10.1007/s00404-024-07565-4.
49. Gangavarapu A. Introducing L2M3, a multilingual medical large language model to advance health equity in low-resource regions Preprint. Posted online April 11, 2024. arXiv 240408705. https://doi.org/10.48550/arXiv.2404.08705.
50. Turing. Fine-tuning LLMS: overview, methods, and best practices. https://www.turing.com/resources/finetuning-large-language-models. Accessed April 26, 2024.
51. Zhao J. LLMDataHub: awesome datasets for LLM training. https://github.com/Zjh-819/LLMDataHub. Accessed April 26, 2024.
52. HuggingFace. Datasets (filter Other by name "llm"). https://huggingface.co/datasets?other=llm. Accessed April 26, 2024.
53. Liu Y, Cao J, Liu C, Ding K, Jin L. *Datasets for large language models: a comprehensive survey* Preprint. Posted online February 28, 2024. arXiv 240218041. https://doi.org/10.48550/arXiv.2402.18041.
54. Aisera. LLM evaluation metrics: performance benchmark. https://aisera.com/blog/llm-evaluation/. Accessed April 26, 2024.
55. Serapio A, Chaudhari G, Savage C, et al. An open-source fine-tuned large language model for radiological impression generation: a multi-reader performance study. *BMC Med Imaging*. 2024;24(1):254. https://doi.org/10.1186/s12880-024-01435-w.
56. Liu H, Tam D, Muqeeth M, et al. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Adv Neural Inform Process Syst*. 2022;35:1950-1965.
57. Hu EJ, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models Preprint. Posted online June 17, 2021. arXiv 210609685. https://doi.org/10.48550/arXiv.2106.09685.
58. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: efficient finetuning of quantized LLMs. *Adv Neural Inform Process Syst*. 2024;36:10088-10115.
59. Gendler M, Nadkarni G, Sudri K, et al. Large language models in cardiology: a systematic review Preprint. Posted online September 1, 2024. medRxiv 24312887. https://doi.org/10.1101/2024.09.01.24312887.
60. Novak A, Rode F, Lisii. The pulse of artificial intelligence in cardiology: a comprehensive evaluation of state-of-the-art large language models for potential use in clinical cardiology Preprint. Posted online January 30, 2024. medRxiv 23293689. https://doi.org/10.1101/2023.08.08.23293689.
61. Boonstra MJ, Weissenbacher D, Moore JH, Gonzalez-Hernandez G, Asselbergs FW. Artificial intelligence: revolutionizing cardiology with large language models. *Eur Heart J*. 2024;45(5):332-345. https://doi.org/10.1093/eurheartj/ehad838.
62. Gui H, Omiye JA, Chang CT, Daneshjou R. The promises and perils of foundation models in dermatology. *J Invest Dermatol*. 2024; 144(7):1440-1448. https://doi.org/10.1016/j.jid.2023.12.019.
63. Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagn Pathol*. 2024;19(1):43. https://doi.org/10.1186/s13000-024-01464-7.
64. Shahab O, El Kurdi B, Shaukat A, Nadkarni G, Soroush A. Large language models: a primer and gastroenterology applications. *Ther Adv Gastroenterol*. 2024;17:17562848241227031. https://doi.org/10.1177/17562848241227031.
65. Omar Sr M, Sharif Sr K, Glicksberg Sr BS, Nadkarni G, Klang E Sr. Emerging applications of NLP and large language models in gastroenterology and hepatology: a systematic review Preprint. Posted online June 27, 2021. medRxiv 24309567. https://doi.org/10.1101/2024.06.26.24309567.
66. Giuffre M, Kresevic S, Pugliese N, You K, Shung DL. Optimizing large language models in digestive disease: strategies and challenges to improve clinical outcomes. *Liver Int*. 2024; 44(9):2114-2124. https://doi.org/10.1111/liv.15974.
67. Mudrik A, Nadkarni GN, Efros O, Glicksberg BS, Klang E, Soffer S. Exploring the role of large language models (LLMs) in hematology: a systematic review of applications, benefits, and limitations. *Br J Haematol*. 2024;205(5):1685-1698. https://doi.org/10.1111/bjh.19738.
68. Barrit S, El Hadwe SE, Carron R, Madsen JR. Rise of large language models in neurosurgery. *J Neurosurg*. 2024;141(3):878-880. https://doi.org/10.3171/2024.3.JNS24610.
69. Chiang C-C, Fries JA. Exploring the potential of large language models in neurology, using neurologic localization as an example. *Neurol Clin Pract*. 2024;14(3):e200311. https://doi.org/10.1212/CPJ.0000000000200311.
70. Romano MF, Shih LC, Paschalidis IC, Au R, Kolachalama VB. Large language models in neurology research and future practice. *Neurology*. 2023;101(23):1058-1067. https://doi.org/10.1212/WNL.0000000000207967.
71. Bachmann M, Duta I, Mazey E, Cooke W, Vatish M, Jones GD. Exploring the capabilities of ChatGPT in women's health: obstetrics and gynaecology. *NPJ Womens Health*. 2024;2(1):26. https://doi.org/10.1038/s44294-024-00028-w.
72. Mudrik A, Tsur A, Nadkarni G, et al. Leveraging large language models in gynecologic oncology: a systematic review of current applications and challenges Preprint. Posted online August 9, 2024. medRxiv 24311699. https://doi.org/10.1101/2024.08.08.24311699.
73. Rydzewski NR, Dinakaran D, Zhao SG, et al. Comparative evaluation of LLMs in clinical oncology. *NEJM AI*. 2024;1(5). 10.1056/aioa2300151. https://doi.org/10.1056/aioa2300151.
74. Lawson McLean A, Wu Y, Lawson McLean AC, Hristidis V. Large language models as decision aids in neuro-oncology: a review of shared decision-making applications. *J Cancer Res Clin Oncol*. 2024;150(3):139. https://doi.org/10.1007/s00432-024-05673-x.
75. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. *JAMA Netw Open*. 2023;6(11):e2343689. https://doi.org/10.1001/jamanetworkopen.2023.43689.

76. Luo M-J, Pang J, Bi S, et al. Development and evaluation of a retrieval-augmented large language model framework for ophthalmology. *JAMA Ophthalmol*. 2024;142(9):798-805. https://doi.org/10.1001/jamaophthalmol.2024.2513.

77. Chatterjee S, Bhattacharya M, Pal S, Lee S-S, Chakraborty C. ChatGPT and large language models in orthopedics: from education and surgery to research. *J Exp Orthop*. 2023;10(1):128. https://doi.org/10.1186/s40634-023-00700-1.

78. Sisk BA, Antes AL, DuBois JM. An overarching framework for the ethics of artificial intelligence in pediatrics. *JAMA Pediatr*. 2024;178(3):213-214. https://doi.org/10.1001/jamapediatrics.2023.5761.

79. Wyatt KD, Alexander N, Hills GD, et al. Making sense of artificial intelligence and large language models—including ChatGPT—in pediatric hematology/oncology. *Pediatr Blood Cancer*. 2024;71(9):e31143. https://doi.org/10.1002/pbc.31143.

80. Obradovich N, Khalsa SS, Khan WU, et al. Opportunities and risks of large language models in psychiatry. *NPP Digit Psychiatry Neurosci*. 2024;2(1):8. https://doi.org/10.1038/s44277-024-00010-z.

81. Volkmer S, Meyer-Lindenberg A, Schwarz E. Large language models in psychiatry: opportunities and challenges. *Psychiatry Res*. 2024;339: 116026. https://doi.org/10.1016/j.psychres.2024.116026.

82. Omar M, Soffer S, Charney AW, Landi I, Nadkarni GN, Klang E. Applications of large language models in psychiatry: a systematic review. *Front Psychiatry*. 2024;15:1422807. https://doi.org/10.3389/fpsyt.2024.1422807.

83. Liu Z, Zhong A, Li Y, et al. Tailoring large language models to radiology: a preliminary approach to llm adaptation for a highly specialized domain. In: Cao X, Xu X, Rekik I, Cui Z, Ouyang X, eds. *Machine Learning in Medical Imaging. MLMI 2023*. Lecture Notes in Computer Science, vol 14348. Springer. doi:10.1007/978-3-031-45673-2_46.

84. D'Antonoli TA, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol*. 2024;30(2):80-90. https://doi.org/10.4274/dir.2023.232417.

85. Lee J, Sharma I, Arcaro N, et al. Automating surgical procedure extraction for society of surgeons adult cardiac surgery registry using pretrained language models. *JAMIA Open*. 2024;7(3): ooae054. https://doi.org/10.1093/jamiaopen/ooae054.

86. Oh N, Choi G-S, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Ann Surg Treat Res*. 2023;104(5):269-273. https://doi.org/10.4174/astr.2023.104.5.269.

87. Adhikari K, Naik N, Hameed BZ, Raghunath SK, Somani BK. Exploring the ethical, legal, and social implications of ChatGPT in urology. *Curr Urol Rep*. 2024;25(1):1-8. https://doi.org/10.1007/s11934-023-01185-2.

88. Gupta R, Pedraza AM, Gorin MA, Tewari AK. Defining the role of large language models in urologic care and research. *Eur Urol Oncol*. 2024;7(1):1-13. https://doi.org/10.1016/j.euo.2023.07.017.

89. Mukherjee S, Gamble P, Ausin MS, et al. Polaris: a safety-focused LLM constellation architecture for healthcare Preprint. Posted online March 20, 2024. arXiv 240313313. https://doi.org/10.48550/arXiv.2403.13313.

90. Zhao L, Zeng W, Shi X, Zhou H, Hao D, Lin Y. Aqulia-Med LLM: pioneering full-process open-source medical language models Preprint. Posted online June 18, 2024. arXiv 240612182. https://doi.org/10.48550/arXiv.2406.12182.

91. Li L, Zhou J, Gao Z, et al. A scoping review of using large language models (LLMs) to investigate electronic health records (EHRs) Preprint. Posted online May 5, 2024. arXiv 240503066. https://doi.org/10.48550/arXiv.2405.03066.

92. Zhang X, Yan C, Yang Y, et al. Optimizing large language models for discharge prediction: best practices in leveraging electronic health record audit logs Preprint. Posted online September 13, 2024. medRxiv 24313594. https://doi.org/10.1101/2024.09.12.24313594.

93. Cui H, Shen Z, Zhang J, et al. LLMs-based few-shot disease predictions using EHR: a novel approach combining predictive agent reasoning and critical agent instruction Preprint. Posted online March 19, 2024. arXiv 240315464. https://doi.org/10.48550/arXiv.2403.15464.

94. Li R, Wang X, Yu H. LlamaCare: an instruction fine-tuned large language model for clinical NLP. In: Calzolari N, Kan M-Y, Hoste V, Lenci A, Sakti S, Xue N, eds. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA/ICCL; 2024:10632-10641.

95. Chao C-J, Banerjee I, Arsanjani R, et al. EchoGPT: a large language model for echocardiography report summarization Preprint. Posted online January 20, 2024. medRxiv 24301503. https://doi.org/10.1101/2024.01.18.24301503.

96. Guan Z, Wu Z, Liu Z, et al. CohortGPT: an enhanced GPT for participant recruitment in clinical study Preprint. Posted online July 21, 2023. arXiv 230711346. https://doi.org/10.48550/arXiv.2307.11346.

97. Demner-Fushman D, Kohli MD, Rosenman MB, et al. Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inform Assoc*. 2016;23(2):304-310. https://doi.org/10.1093/jamia/ocv080.

98. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci data*. 2019;6(1):317. https://doi.org/10.1038/s41597-019-0322-0.

99. Zhou H, Gu B, Zou X, et al. A survey of large language models in medicine: progress, application, and challenge Preprint. Posted online November 9, 2023. arXiv 231105112. https://doi.org/10.48550/arXiv.2311.05112.

100. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). *NPJ Digit Med*. 2024;7(1):183. https://doi.org/10.1038/s41746-024-01157-x.

101. Pal A, Umapathi LK, Sankarasubbu M. Med-HALT: medical domain hallucination test for large language models Preprint. Posted online July 28, 2023. arXiv 230715343. https://doi.org/10.18653/v1/2023.conll-1.21.

102. Ong JCL, Chang SY-H, William W, et al. Ethical and regulatory challenges of large language models in medicine. *Lancet Digit Health*. 2024;6(6):e428-e432. https://doi.org/10.1016/S2589-7500(24)00061-X.

103. Goh E, Bunning B, Khoong E, et al. ChatGPT influence on medical decision-making, bias, and equity: a randomized study of clinicians evaluating clinical vignettes Preprint. Posted online November 27, 2023. medRxiv 23298844. https://doi.org/10.1101/2023.11.24.23298844.

104. Schmidgall S, Harris C, Essien I, et al. Addressing cognitive bias in medical language models Preprint. Posted online February 12, 2024. arXiv 240208113. https://doi.org/10.48550/arXiv.2402.08113.

105. Perez-Downes JC, Tseng AS, McConn KA, et al. Mitigating bias in clinical machine learning models. *Curr Treat Options Cardiovasc Med*. 2024;26(3):29-45. https://doi.org/10.1007/s11936-023-01032-0.

106. Omar Sr M, Sorin Sr V, Apakama DU, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review Preprint. Posted online October 1, 2024. medRxiv 24313295. https://doi.org/10.1101/2024.09.09.24313295.