Journal of the American Medical Informatics Association, 2024, **31(9)**, 1833–1843
https://doi.org/10.1093/jamia/ocae045
Advance access publication 13 April 2024
Research and Applications

Research and Applications

PMC-LLaMA: toward building open-source language models for medicine

Chaoyi Wu, BEng[1,2], Weixiong Lin, BEng[1,2], Xiaoman Zhang, BEng[1,2], Ya Zhang, PhD[1,2], Weidi Xie, PhD[1,2], Yanfeng Wang, PhD[1,2,]*

[1]Cooperative Medianet Innovation Center (CMIC), Shanghai Jiao Tong University, Shanghai, 200240, China, [2]Shanghai AI Laboratory, Shanghai, 200232, China

*Corresponding author: Yanfeng Wang, PhD, Cooperative Medianet Innovation Cente (CMIC), Shanghai Jiao Tong University, Shanghai, 200240, China (wangyanfeng622@sjtu.edu.cn)

Chaoyi Wu and Weixiong Lin contributed equally to this work.

Abstract

Objective: Recently, large language models (LLMs) have showcased remarkable capabilities in natural language understanding. While demonstrating proficiency in everyday conversations and question-answering (QA) situations, these models frequently struggle in domains that require precision, such as medical applications, due to their lack of domain-specific knowledge. In this article, we describe the procedure for building a powerful, open-source language model specifically designed for medicine applications, termed as PMC-LLaMA.

Materials and methods: We adapt a general-purpose LLM toward the medical domain, involving data-centric knowledge injection through the integration of 4.8M biomedical academic papers and 30K medical textbooks, as well as comprehensive domain-specific instruction fine-tuning, encompassing medical QA, rationale for reasoning, and conversational dialogues with 202M tokens.

Results: While evaluating various public medical QA benchmarks and manual rating, our lightweight PMC-LLaMA, which consists of only 13B parameters, exhibits superior performance, even surpassing ChatGPT. All models, codes, and datasets for instruction tuning will be released to the research community.

Discussion: Our contributions are 3-fold: (1) we build up an open-source LLM toward the medical domain. We believe the proposed PMC-LLaMA model can promote further development of foundation models in medicine, serving as a medical trainable basic generative language backbone; (2) we conduct thorough ablation studies to demonstrate the effectiveness of each proposed component, demonstrating how different training data and model scales affect medical LLMs; (3) we contribute a large-scale, comprehensive dataset for instruction tuning.

Conclusion: In this article, we systematically investigate the process of building up an open-source medical-specific LLM, PMC-LLaMA.

Key words: large language models; biomedical NLP; generative language models; ChatGPT.

Objective

The rapid advancement of large language models (LLMs), for example, OpenAI's GPT-3.5[1] and GPT-4[2] has truly revolutionized the natural language processing research,[3,4] sparking AI applications for numerous daily scenarios. Unfortunately, the training details and model architectures for the GPT-series remain unclear. The open-source LLMs, for example, LLaMA-series,[5,6] also show comparable performance with GPT-3.5 in the general domain. However, though the LLMs demonstrate proficiency in everyday conversations, in the medical domain which requires high precision, they often produce seemingly accurate output but lead to incorrect conclusions, which could be highly fatal. We conjecture this is due to their lack of comprehensive medical knowledge.

Existing works have also explored several ways for adapting general-purpose LLMs toward the medicine domain, like Med-Alpaca,[7] Chat-Doctor,[8] and Med-PaLM 2.[9] Among these, Med-PaLM 2 is the only work successfully outperforming GPT-3.5 while their training details, for example, training data, model architecture, remain unclear. Thus, systematic investigation of the medical domain adaptation for LLMs still needs to be discussed further, especially in the open-source community.

Our goal is to systematically adapt an open-source general LLM, that is, LLaMA, toward the medicine domain from the following aspects. First, we adopt data-centric medical-specific knowledge injection for the language model with a large-scale free text medical corpora. We claim that language models can accumulate enough medical knowledge in this step and build up a better embedding space for domain-specific complex terminologies. Second, augmenting the reasoning capabilities of the proposed model. This empowers the model to link its medical knowledge with provided case information and provide well-justified recommendations. Lastly, enhancing the alignment ability of LLMs. Robust alignment with various instructions facilitates effective instruction prompting adaptation to a diverse spectrum of tasks.

In conclusion, in this article, we systematically build up an LLM for medicine through data-centric knowledge injection and medical-specific instruction tuning and release an open-source lightweight medical-specific language model, PMC-LLaMA. Specifically, we first collect a large medical-specific corpus, named MedC-K, consisting of **4.8M** biomedical academic papers and **30K** textbooks for knowledge injection. We then adopt medical-specific instruction tuning on a new

Received: December 15, 2023; **Revised:** February 6, 2024; **Editorial Decision:** February 16, 2024; **Accepted:** March 5, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of the American Medical Informatics Association.
All rights reserved. For permissions, please email: journals.permissions@oup.com

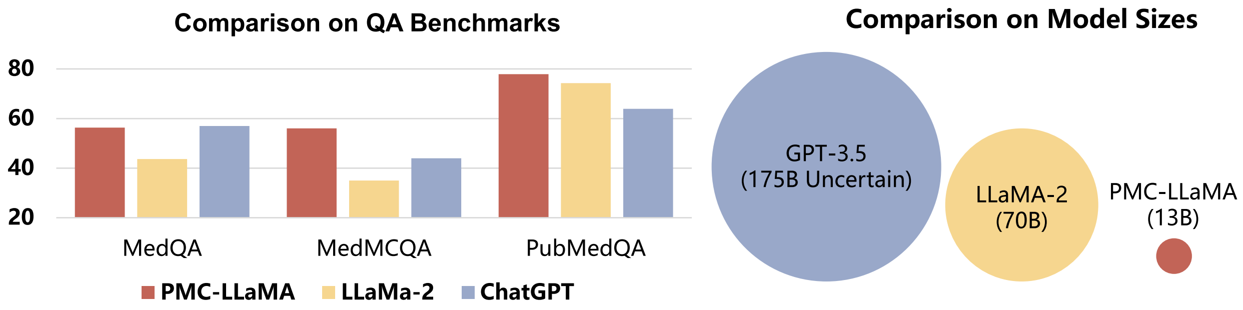


Figure 1. In the left, we show the general comparison between our PMC-LLaMA with LLaMA-2 and GPT-3.5. On the right, we visually show the advantages of our model in model sizes. PMC-LLaMA is much smaller than the others.

medical knowledge-aware instruction dataset, termed MedC-I, consisting of medical QA, rationale, and conversation with **202M** tokens in total. We evaluate PMC-LLaMA on various medical QA benchmarks, surpassing GPT-3.5 and LLaMA-2 as shown in Figure 1.

Background and significance

Large language model

Recently, the great success of LLMs,[1,2,9,10] has garnered significant attention within the field of natural language processing. For example, OpenAI's strides with GPT-3.5 and GPT-4 have showcased remarkable capabilities in various tasks, including text generation, language translation, question answering (QA), and more. However, intricate details concerning their training methodologies and weight parameters remain undisclosed. LLaMA[5] serves as an open-source alternative for the foundational language model, ranging from 7 billion to 65 billion parameters. In light of these advancements, there has been a surge of interest in tailoring language models for specific biomedical domains. Most of these models are prompt-tuned using LLaMA on a small medical corpus, resulting in a deficiency of comprehensive medical knowledge integration.

Instruction tuning

For LLMs to follow natural language instructions and complete real-world tasks, instruction tuning has been widely used for alignment.[11,12] This involves fine-tuning the model on a collection of tasks described via instructions, to effectively improve the zero-shot and few-shot generalization abilities of LLMs.[13,14] Building on the publicly accessible language models, Alpace[15] and Vicuna[16] are proposed by finetuning on the machine-generated instruction-following samples, showing promising performance. In the medical domain, Chat-Doctor[8] and Med-Alpaca[7] are instruction-tuned for medical QA and dialogue applications. Notably, Med-PaLM[4] represents the pinnacle of LLMs in the medical field, trained with intensive instruction tuning on the strong PaLM model (with 540 billion parameters). However, its code and data remain inaccessible to the public.

Medical foundational language model

In addition to instruction tuning, there have been extensive efforts in training foundation model for medicine, for example, BioBERT, BioMedGPT, etc.[17–19] However, these models exhibit certain limitations, first, most domain-specific models have been exclusively trained on medical corpora. The lack of exposure to diverse knowledge domains beyond medicine can impede the model's capability to perform reasoning or context understanding; second, these models are limited in model scale and are predominantly designed to be based on BERT, thus imposing restrictions on their utility for a wide array of downstream tasks under instruction prompting. In this work, we aim to resolve these 2 limitations by adapting a general LLM toward medicine with knowledge injection, followed by medical-specific instruction tuning.

Methods

In this section, we will describe our method in detail. Generally speaking, our method is data-driven, focusing on the procedure for building models by acquiring appropriate data. Next, we describe the problem formulation for generative language models and then introduce our dataset construction procedure. Finally, we will introduce the training implementation details.

Problem formulation

In this article, our goal is to systematically investigate the procedure for steering a pretrained foundational language model to the knowledge-intense domain, that is, medicine. The training process can be divided into 2 stages: first, a data-centric knowledge injection stage, that aims to enrich the language model with fundamental medical knowledge; second, a medical-specific instruction tuning stage, that tailors the model to align with clinical use cases.

At the training stage, assuming the text input as a sequence of tokens, for example, $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$, where each u_i is a text token and N is the total sequence length, the training objective is to minimize auto-regressive loss, with the major difference on whether to compute loss on the entire sequence or only subsequence, as detailed in the following.

Data-Centric Knowledge Injection. For the knowledge injection step, we simply minimize the default auto-regressive loss, all free-form texts on medical knowledge can be used, for the model to accumulate sufficient medical-specific knowledge contexts, formulated as:

$$L(\Phi) = -\sum \log \Phi(u_i|u_{<i}). \tag{1}$$

where $u_{<i}$ indicates the tokens appear before index i and Φ denotes our model.

Medical-Specific Instruction Tuning. At this stage, the token sequence is further split into instruction $\mathcal{I}$, and response $\mathcal{R}$, the former is to mimic the user's query, thus the loss is ignored at training time, denoted as:

$$L(\Phi) = -\sum_{u_i \in \mathcal{R}} \log \Phi(u_i | u_{<i}, \mathcal{I}). \quad (2)$$

At inference time, the common use case is a conversation, where the user normally provides the question as instruction $\mathcal{I}$, and the output of the model serves as the answer.

To support our 2-stage training, namely data-centric knowledge injection, and medical-specific instruction tuning for alignment, we herein detail the procedure for constructing the high-quality language datasets.

Dataset-I: fundamental medical knowledge

To steer a general-purpose foundational language model for medical scenario, we propose to first conduct data-centric knowledge injection, that aims to expose the model with medical-related terminologies and definitions. We primarily focus on 2 key data sources, namely, biomedical papers and textbooks.

Papers. As a valuable knowledge resource, academic papers naturally contain high-quality, cutting-edge medical knowledge. We start with the S2ORC[20] Datasets with 81.1M English-language academic papers, and pick out those biomedical-related papers depending on whether having corresponding PubMed Central (PMC) IDs. As a result, there are around 4.8M biomedical papers left, totaling over 75B tokens.

Books. We collect 30K textbooks sourced from various outlets, for example, open libraries, university library, and reputable publishers, covering a wide range of medical specialties as shown in Figure 2. For preprocessing, we first extract the text content from the book PDF, then carry out data cleaning via deduplication and content filtering. Specifically, we eliminate extraneous elements such as URLs, author lists, superfluous information, document contents, references, and citations. Additionally, we have also removed any references to images and tables within the paragraphs, for example, "Figure 1," using *clean-text* (https://pypi.org/project/clean-text/) package. After this thorough cleaning process, there are approximately 4B tokens left. **Note that,** In books or papers, the patient privacy has already been well-resolved by original authors.

Combination. The 2 corpora encompass distinct types of medical knowledge, while papers predominantly capture cutting-edge insights, books capture more fundamental medical knowledge, which is more crucial for pretrained general-purpose language models. Hence, when blending these 2 datasets for knowledge injection training, we use a ratio of 15:4:1 at each training batch, by that we mean to emphasize "book" tokens more. Specifically, we sample more tokens from books, ensuring they occupy 15 parts per batch and sample tokens from "papers" less so that they occupy 4 parts per batch. For the remaining 1 occupation, we sample from a general language corpus, RedPajama-Data[21] to form a complete batch. This mainly aims to avert catastrophic forgetting of previously acquired general text knowledge after extensive knowledge injection on large-scale medical-specific data.

Knowledge Injection Training. Till here, we have constructed a large-scale language dataset of fundamental medical knowledge, termed as **MedC-K**. With such corpus, we conduct data-centric knowledge injection with auto-regressive training, resulting in a language model for medicine, named **PMC-LLaMA$_\mathrm{K}$**, as the largest number of tokens are from PubMed Central academic papers.

Dataset-II: medical instructions

Here, we proceed to carry out instruction tuning with the goal of guiding the model to respond to various instructions, by exploiting the medical knowledge embedded in PMC-LLaMA$_\mathrm{K}$ model. Generally speaking, our instruction tuning datasets are composed of 3 main parts, namely, medical consulting conversation, medical rationale QA, and medical knowledge graph prompting.

Medical Conversation. Considering there exists diverse doctor-patient dialogues in daily life, the questions raised by patients are naturally suitable as instructions and doctor responses as ground truth. We start with the data collected by Med-Alpaca[7] and Chat-Doctor,[8] and further expand the provided instructions into various synonymous sentences to improve model's robustness to diverse instructions. Specifically, we use the GPT-4 with the following query prompt:

> "Rewrite 10 sentences that convey similar meanings to what I've stated: {instruction seeds}.",

where {instruction seeds} denotes the provided instruction from ChatDoctor or MedAlpaca, and the query can be repeated until the desired prompt number. At training time, we randomly select one instruction from the instruction base to simulate the inputs from real users and avoid over-fitting on specific instruction templates.

Medical Rationale QA. Beyond daily conversations, we also consider equipping our model with reasoning ability with professional medical knowledge. We start with the training sets of the open-source medical multi-choice QA datasets, such as USMLE,[22] PubMedQA,[23] and MedMCQA.[24] Despite the questions in them naturally demanding medical-specific knowledge, most of these datasets only include plain choices, lacking detailed reasoning guidance. To complement such information, we prompt

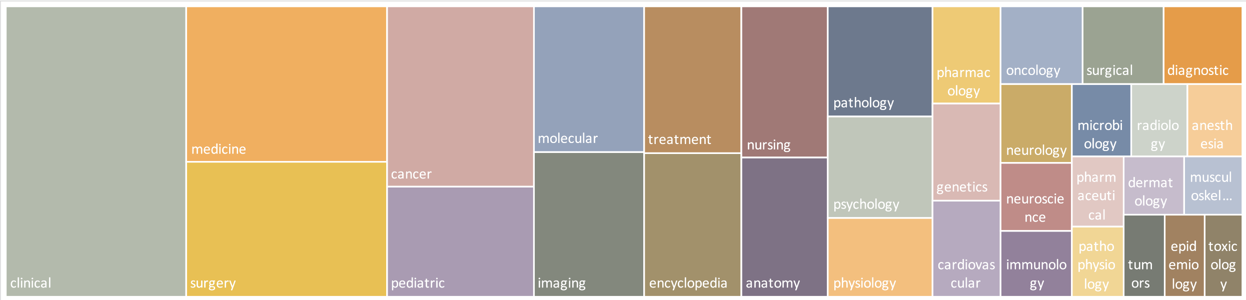
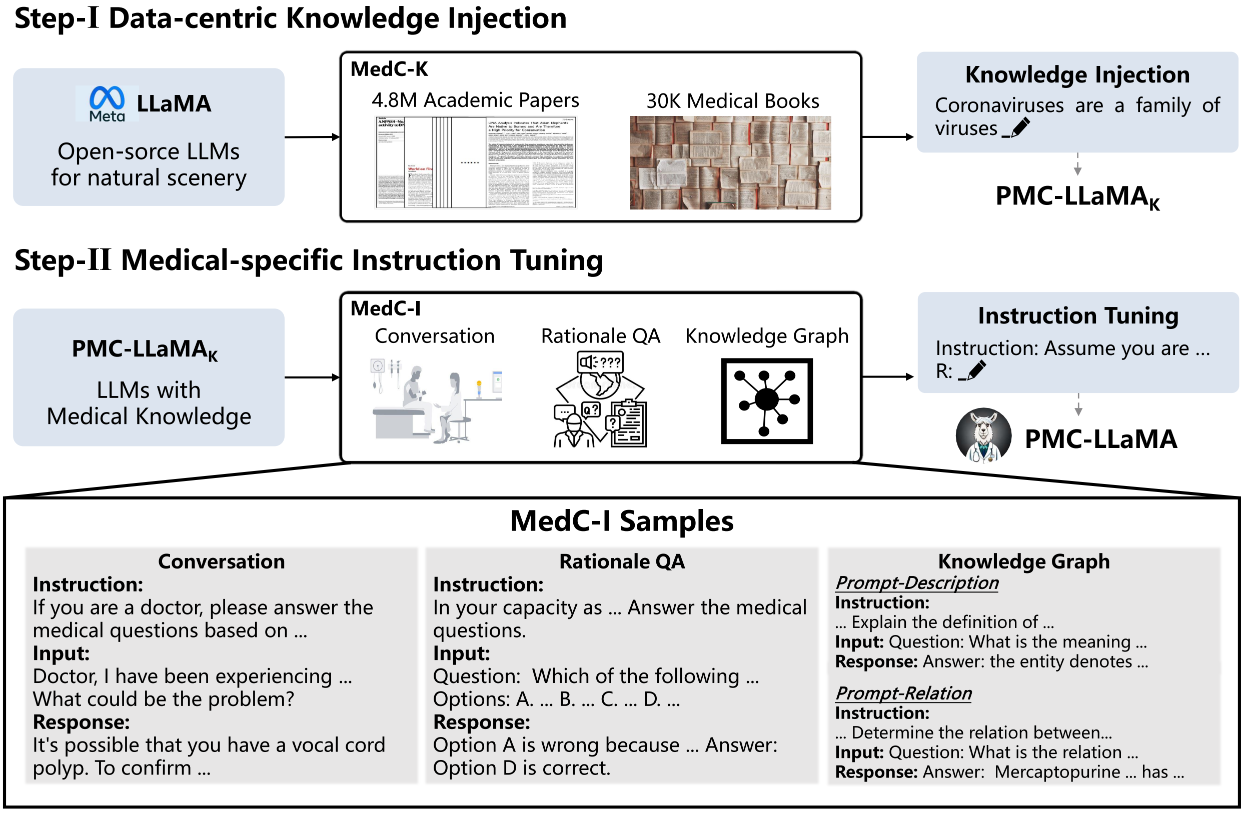


Figure 2. Distribution of medical textbooks categories. The box sizes denote the book numbers for different categories.


Step-I Data-centric Knowledge Injection
LLaMA
Meta
Open-sorce LLMs for natural scenery
MedC-K
4.8M Academic Papers
30K Medical Books
Knowledge Injection
Coronaviruses are a family of viruses _
PMC-LLaMA$_K$
Step-II Medical-specific Instruction Tuning
PMC-LLaMA$_K$
LLMs with Medical Knowledge
MedC-I
Conversation
Rationale QA
Knowledge Graph
Instruction Tuning
Instruction: Assume you are ...
R: _
PMC-LLaMA
MedC-I Samples

Conversation
Instruction:
If you are a doctor, please answer the medical questions based on ...
Input:
Doctor, I have been experiencing ... What could be the problem?
Response:
It's possible that you have a vocal cord polyp. To confirm ...

Rationale QA
Instruction:
In your capacity as ... Answer the medical questions.
Input:
Question: Which of the following ...
Options: A. ... B. ... C. ... D. ...
Response:
Option A is wrong because ... Answer: Option D is correct.

Knowledge Graph
Prompt-Description
Instruction:
... Explain the definition of ...
Input: Question: What is the meaning ...
Response: Answer: the entity denotes ...

Prompt-Relation
Instruction:
... Determine the relation between...
Input: Question: What is the relation ...
Response: Answer: Mercaptopurine ... has ...

Figure 3. The training pipeline of PMC-LLaMA. Our training flow can be separated into 2 parts, that is, data-centric knowledge injection and medical-specific instruction tuning. In knowledge injection, we collect 4.8M biomedical academic papers and 30K medical books for further injecting knowledge into LLaMA. In the instruction tuning stage, we mainly consider 3 aspects, medical conversation, medical rationale question answering, and knowledge graph, containing 202M tokens in total.

GPT-3.5[1] for causality analysis. Specifically, given a QA pair, we query GPT-3.5 to get rationale output (check Supplementary Material for details), and treat the output as an explanation with structured format shown at the bottom of Figure 3.

Medical Knowledge Graph Prompting. In addition to the aforementioned data, we also consider exploiting medical knowledge graphs UMLS,[25] to align with clinicians' experience. Specifically, to link the medical terminologies with their respective knowledge description or corresponding relationships, we construct QA pairs to translate the common knowledge graph. There are 2 main types contained in medical knowledge graph, that is, entity descriptions and entity relationships. We add 2 different prompts for them as shown at the bottom of Figure 3, that demand the model to output descriptions for a certain entity or predict the relationship between 2 entities.

Medical-Specific Instruction Tuning. By combining the above 3 parts together, we form a large-scale, high-quality, medical-specific instruction tuning dataset, **MedC-I**, consisting **202M** tokens. We further tune PMC-LLaMA$_K$ on it, resulting in our final model—**PMC-LLaMA.**

Training details

We start by carrying out knowledge injection on open-source LLaMA model, optimizing an auto-regressive loss. Specifically, at training time, the max context length is set as 2048, with a batch size to be 3200, and the model is trained with AdamW optimizer[26] with a learning rate 2e-5. We adopt the Fully Sharded Data Parallel acceleration strategy, bf16 (Brain Floating Point) data format, and gradient checkpointing.[27] Since we sample more tokens from books in each batch, the model will finish seeing all book tokens earlier. Thus, we here define 1 epoch for seeing all book tokens instead of seeing all mixed tokens. The model is trained with knowledge injection for 5 epochs with 32 A100 GPUs. Then we carry out medical-specific instruction tuning on MedC-I, for 3 epochs with 256 batch size with 8 A100 GPUs. Note that, at instruction tuning stage, each epoch refers to looping through all sequences. The first knowledge injection stage may take around 2 weeks on the 13B model and 1 week on the 7B model with 32 A100 GPUs. The following instruction tuning step may take around 3 days on 13B models with 8 A100 GPUs, and around 1 day on 7B models under the same setting.

Results

In this section, we will discuss our final results. Specifically, we start by introducing our benchmarks, followed by various baseline models and evaluation settings. Then we present our quantitative and qualitative results in detail.

Benchmarks

QA Benchmarks. In the literature, the primary method for measuring the ability of medical language models is based on multiple-choice QA, which uses accuracy as the main metric. Following the convention, we adopt 3 prominent medical QA benchmarks for evaluation.

- PubMedQA[23] is a biomedical QA dataset collected from PubMed abstracts. The task of PubMedQA is to answer research questions with yes/no/maybe, which can be considered as the multiple-choice question. It is split into 3 subsets: 1k manually labeled pairs (PQA-L), 61.2k unlabeled pairs (PQA-U), and 211.3k artificially generated pairs (PQA-A). Following former works,[28] we view PQA-A as the train set, PQA-L as the test set, and discard the PQA-U parts.
- MedMCQA[24] is a dataset of multiple choice questions, that are sourced from mock exams and past exams of 2 Indian medical school entrance exams called AIIMS and NEET-PG.[24] The train split contains 182 822 questions, and the test split contains 4183 questions. Each question has 4 choices.
- USMLE[22] is a dataset of multiple choice questions (4 choices per question), based on the United States Medical License Exams. The dataset is collected from the professional medical board exams, covering 3 languages: English, simplified Chinese, and traditional Chinese, containing 12 724, 34 251, and 14 123 questions, respectively. Here, we use the English parts and split it into 10 178 questions for training, 1273 for validation, and 1273 for testing, following the official splits.

Manual Evaluation. We adopt manual rating to evaluate the quality of long sentence generations. Two types of contents are considered, that is, rationale and conversation. Rationale refers to the detailed reasoning contexts for QA questions, and conversation is the response to free-form dialogue in medical consulting. Specifically, we sample 100 cases from the medical QA test sets for rationale and 50 samples from iCliniq (an online medical consulting website) for conversation. We report results for 3 metrics, namely, CU (Contextual Understanding), KC (Knowledge Correlation), and LF (Language Fluency). CU measures whether the model can correctly extract and understand the key information from a given context, and KC measures whether the model's knowledge-wise output is related to the answer. The users are provided with the output of 6 evaluated models, that is, GPT-3.5, LLaMA-2-13B, LLaMA-2-70B, Med-Alpaca, Chat-Doctor, and ours, then asked to rate the generated sentences from different aspects, that is, CU KC and LF. The average ranking level on all test cases for these 3 aspects is viewed as 3 metrics. The rating procedure is double-checked by 2 clinical patient users and further checked by a 10-year experienced clinical physician for fact correctness.

Baseline models

LLaMA. LLaMA[5] is the most widely used open-source language model, it has been trained on a large text corpus with only auto-regressive learning, that is, no instruction tuning is involved.

LLaMA-2. LLaMA-2[6] is the improved version of LLaMA that has been further tuned with instructions. Its largest version (70B) is reported to be the best on natural scenery among the open-source LLMs.

GPT-3.5. GPT-3.5[1] is a commercial model released by OpenAI in November, 2022, that has shown remarkable performance on a wide range of NLP tasks in various domains, including medicine. Note that, since the exact details of GPT-3.5 are confidential, we follow the general presumption that GPT-3.5 is roughly the same as GPT-3 in model sizes (175B).[29] Note that, until now, no exact scale number is released by OpenAI.

Med-Alpaca. Med-Alpaca[7] is a model further fine-tuned on Alpaca[15] using medical instruction data. They focus on the task of assisting medical dialogues and QA.

Chat-Doctor. Chat-Doctor[8] is a language model aiming for health assistants, that is designed to provide users with medical information, advice, and guidance. For training, it has leveraged the dialogue-based instruction tuning data.

GatorTronGPT. GatorTronGPT[30] is a 20B GPT-like model trained with 277B words of mixed clinical and general English corpus. Its architecture is based on GPT-3.[31] GatorTronGPT targets on improving Biomedical NLP for medical research and shows superior performance on various benchmarks.

BioMedLM. BioMedLM[32] is a 2.7B BERT-like model trained with biomedical papers.

Med42. Med42[33] is an open-access clinical LLM developed by M42 with 70B parameters based on LLaMA-2 to expand access to medical knowledge.

MEDITRON. MEDITRON[32] is a 70B medical LLM built on LLaMA-2 by further training on papers and instruction tuning sets.

Evaluation settings

In this part, we describe the evaluation in detail to compare the above language models on the QA benchmarks. **Note that**, we do not claim the presented comparison to be completely fair, as a number of training details, for example, data, and architecture remain undisclosed for the commercial model. Therefore, we only treat these baseline models for reference, and are more focused on presenting our procedure for building on a powerful language model for medicine, demonstrating how different data sources and model scales affect the LLM performance in medicine.

Our evaluation settings can be divided into 2 types: task-specific fine-tuning evaluation and instruction prompting evaluation.

Task-Specific Fine-Tuning Evaluation. In this evaluation setting, we use the combination of 3 QA training sets to further fine-tune a language model and then evaluate it.

Instruction Prompting Evaluation. In this evaluation setting, we directly test the model by giving a medical QA instruction, for example, "Make a choice based on the question and options.", without doing any task-specific fine-tuning.

It is worth emphasizing that task-specific fine-tuning evaluation is considerably an easier task, as it simply requires the model to fit one specific dataset, while the latter requires the model to do instruction prompting transfer on different tasks based on various instructions.

In our evaluation, we test all models that are capable of following free text instructions in "Instruction Prompting Evaluation" setting, for example, LLaMA-2, GPT-3.5, GPT-4, Med-Alpaca, Med42, MEDITRON, and our final model PMC-LLaMA. For others that have not experienced instruction tuning, we adopt the "Task-specific Fine-tuning Evaluation" setting, as these models are unable to give answers without finetuning, for example, LLaMA, GatorTronGPT, BioMedLM, and our intermediate model PMC-LLaMA$_K$ without instruction tuning.

Table 1. The comparison of different PEFT methods for instruction tuning on PMC-LLaMA$_K$ at 13B scales.

Methods	Param (M)	RAM (GB)	Time (GPU·H)	MedQA	MedMCQA	PubMedQA
LoRA-16	6.7	74	82	45.64	46.66	77.1
LoRA-32	13.4	74	82	46.35	46.52	77.5
LoRA-64	26.8	75	83	45.87	47.85	77.0
P-Tuning-4	7.8	76	97	37.41	39.04	73.5
P-Tuning-8	7.8	76	97	39.26	41.45	72.4
P-Tuning-16	7.8	76	98	43.34	42.91	72.3
Full-Finetuning	13 312	224	384	56.36	56.04	77.90

The "Param" denotes the number of trainable parameters measured in million (M), GPU RAM denotes the training consumption with one batch size per measured by Gigabyte (GB). The training time denotes the total GPU hours measured with one GPU. We adopt 2 main PEFT methods for ablation, that is, LoRA and P-Tuning. For LoRA, we change the matrix rank (16, 32, 64) to control the total parameters, for P-Tuning, we change the trainable token number (4, 8, 16). The ACC scores for 3 medical QA benchmarks are reported.
Abbreviations: QA, question answering; PEFT, parameter-efficient tuning.

Table 2. Ablation study on QA benchmarks.

		Knowledge Injection		Instruction Tuning					
Method	Model Size	Papers	Textbooks	Rationale	Conversation	KG	MedQA	MedMCQA	PubMedQA
Baseline (LLaMA)	7B	✗	✗	✗	✗	✗	44.54	48.51	73.40
Baseline (LLaMA)	13B	✗	✗	✗	✗	✗	45.48	51.42	76.40
PMC-LLaMA$_K$	7B	✓	✗	✗	✗	✗	44.70	50.54	69.50
	7B	✗	✓	✗	✗	✗	45.03	51.83	73.80
	7B	✓	✓	✗	✗	✗	45.56	51.45	74.60
	13B	✓	✓	✗	✗	✗	48.15	54.15	77.10
PMC-LLaMA	13B	✓	✓	✓	✗	✗	49.32	54.56	77.20
	13B	✓	✓	✓	✓	✗	54.43	55.77	77.00
	13B	✓	✓	✓	✓	✓	**56.36**	**56.04**	**77.90**

ACC scores are reported in the table. Note that for the models without the ability to follow instructions, we task-specific fine-tune them on the combination of the 3 downstream training sets to get the number. The best results are bolded.
Abbreviations: KG, knowledge graph; QA, question answering.

Ablation study

As shown in Tables 1 and 2, we systematically study the different design choices on various medical QA benchmarks, for example, instruction tuning training strategies, the effect of the model scale, data-centric knowledge injection, and medical-specific instruction tuning.

PEFT or Full-Finetuning. We first investigate the instruction tuning training strategies, that is, parameter-efficient tuning (PEFT)[34] or full-finetuning. We adopt 2 classical PEFT methods, LoRA[35] and P-Tuning,[36] for comparison. For LoRA, we try different ranks as recommended by the original paper to control the training parameters, and for P-Tuning, we try different trainable tokens. The results are shown in Table 1. As indicated by the results, generally, LoRA performs better than P-Tuning. For LoRA, increasing training parameters does not have a significant effect on the results, while for P-Tuning, increasing trainable token numbers can further improve the results. However, despite PEFT can reduce the training cost, it still shows suboptimal results than full-finetuning. In the following section, we all adopt the full-finetuning for our experiments and comparison.

Model Scale. The scaling law[37] can also be observed in the medical corpus, for example, as shown in the table, when switching the model size from 7B to 13B, performance on all benchmarks has been improved. This phenomenon holds for both the baseline LLaMA model and PMC-LLaMA$_K$, which has further trained with fundamental medical knowledge.

Data-Centric Knowledge Injection. Compared with the 7B baseline LLaMA model, integrating biomedical papers brings a performance gain from 44.54% to 44.70% and 48.51% to 50.54% on MedQA and MedMCQA, respectively, knowledge injection with textbooks brings more performance boost, that is, obtaining 0.49%, 3.32%, and 0.4% on MedQA, MedMCQA, and PubMedQA, respectively, indicating that the knowledge from textbooks is more effective for enhancing general LLMs, thus inspiring us to weigh books more in mixing. By mixing papers with books together, the final performance will be further improved. Both observations have shown the importance of injecting fundamental medical knowledge.

Medical-Specific Instruction Tuning. We start instruction tuning with only rationale QA data. In this case, since only QA task is considered, the difference from task-specific fine-tuning only lies in whether to give a rationale sentence as supervision signals. We observe that simply incorporating rationale cases can lead to enhanced QA results compared to task-specific fine-tuning on plain choice data, showcasing an improvement of 1.17% on the MedQA dataset.

Furthermore, integrating conversations with rationale QA for instruction tuning can produce substantial enhancements, with performance boosts from 49.32% to 54.43% on MedQA. This demonstrates the pivotal role played by the diversity of question types during the instruction tuning stage, as all involved questions will be limited on medical choice tests without conversation. In addition, the incorporation of a knowledge graph introduces a further improvement of 1.93% on the MedQA dataset, demonstrating the importance of using explicit instructions to emphasize the key medical concepts.

Comparison with baselines

In Table 3, we conduct a comparative analysis of our model against SOTA baseline models on 3 QA benchmark datasets

Table 3. Evaluation on QA benchmarks.

Methods	Model Size	MedQA	MedMCQA	PubMedQA	Average
Human (pass)[a]	—	50.0	—	60.0	—
Human (expert)[a]	—	87.0	90.0	78.0	85.0
Close-source					
GPT-4[a]	—	81.38	72.36	74.4	76.04
Med-Palm 2[a]	540B	86.5	72.3	81.8	80.20
GPT-3.5[a]	175B (Uncertain)	57.0	44.0	63.9	54.97
Open-source					
MEDITRON[a]	70B	59.8	53.3	**79.8**	64.30
Med42[a]	70B	**63.9**	**59.2**	61.2	**64.43**
LLaMA-2	70B	43.68	35.02	74.3	51.00
LLaMA-2	13B	42.73	37.41	68.0	49.40
Med-Alpaca	13B	30.85	31.13	53.2	38.38
MEDITRON[a]	7B	37.4	36.3	69.3	47.67
Chat-Doctor	7B	33.93	31.10	54.3	39.78
GatorTronGPT[a,b]	20B	45.1	42.9	77.6	55.2
BioMedLM[a,b]	2.7B	50.3	N/A	74.4	N/A
PMC-LLaMA_{K}[b]	13B	48.15	54.15	77.1	59.80
PMC-LLaMA	13B	56.36	56.04	77.9	**64.43**

ACC scores are reported. Average refers to the average of the 3 datasets. The best one of the open-source model is bolded and the second one is underlined.
Abbreviation: QA, question answering.

[a] These related results are borrowed from other papers. The “Human” and GPT-3.5 scores are from LMFlow[28] and the other results are from their original paper, respectively.

[b] These models are specific to multi-choice QA tasks and could not respond correctly to other instructions.

for evaluation. We also show a qualitative case study to demonstrate the conversation and rationale ability.

Medical QA Ability. While comparing with *other LLMs* on medical QA benchmarks, PMC-LLaMA achieves superior results on most of them. Specifically, in the open-source LLMs, PMC-LLaMA achieves an average accuracy of 64.43%, which is the best among all the open-sourced LLMs, significantly surpassing those with similar parameter scale (<70B) and comparable to the models with much more parameters (70B), like Med42 and MEDITRON-70B, better than the close-source model GPT-3.5. It is worth emphasizing that, Med42 and MEDITRON become visible online after ours, and our corpus also plays a role in inspiring the development of MEDITRON.

For those *models specifically trained* on each QA dataset from corresponding papers, for example, GatorTron,[38] BioMedLM,[32] our model beats the benchmark-specific models with instruction prompting evaluation. In addition, the medical QA datasets were originally used as medical exams. “Human(pass)” denotes the *human passing scores*, “Human (expert)” denotes the average *expert scores*. The “Human” scores and GPT-3.5 instruction prompting scores are inherited from LMFlow.[28]

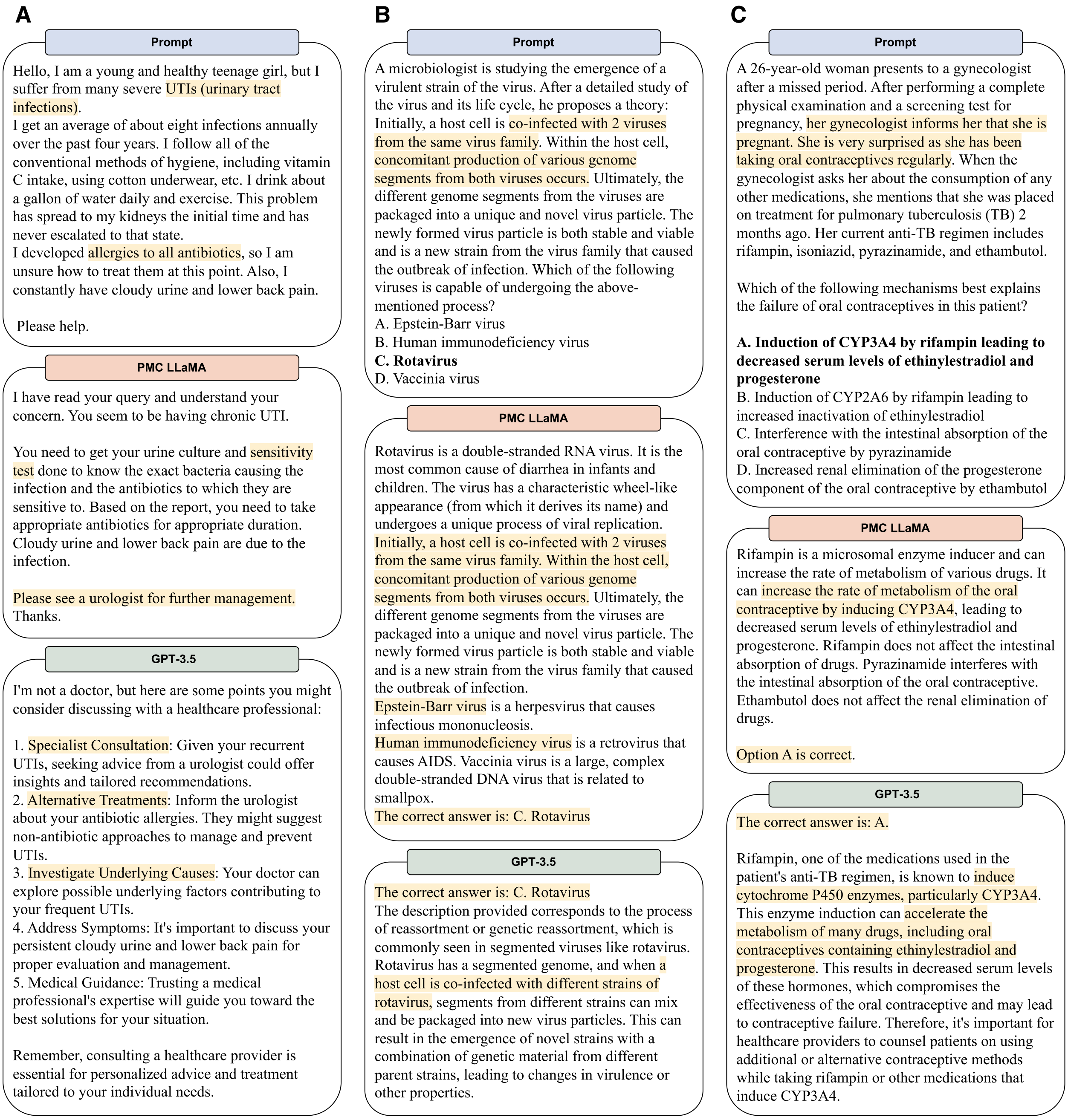
For qualitative results, we conduct the human evaluation of the model’s ability at rationale and conversation. In Table 4, we report results for 3 metrics, namely, CU (Contextual Understanding), KC (Knowledge Correlation), and LF (Language Fluency). CU measures whether the model can extract key information from a given context and understand it correctly, while KC measures whether the model knowledge-wise output is related to the answer. The final rating is double-checked by 2 clinical users. As shown in Table 4, comparing with open-source models, our model demonstrates superior results on CU and KC both for rationale and conversation, which is the best among the open-source LLMs and comparable to GPT-3.5. On language fluency (LF), our model still has room for improvement.

Table 4. Human evaluation.

	Rationale			Conversation		
Methods	CU ↓	KC ↓	LF ↓	CU ↓	KC ↓	LF ↓
GPT-3.5 (close source)	2.50	2.39	3.40	2.14	2.26	3.04
LLaMA-2-13B	3.61	3.72	3.99	3.70	3.96	4.36
LLaMA-2-70B	3.45	3.53	3.56	3.28	2.84	**3.00**
Med-Alpaca	5.15	4.68	3.53	4.80	4.98	3.50
Chat-Doctor	3.88	3.94	**2.93**	4.42	4.20	3.12
Ours	**2.41**	**2.74**	3.59	**2.66**	**2.76**	3.98

We manually rated the 6 models (the listed 5 together with GPT-3.5) on the rationale and conversation abilities. The average rating ranks are reported, lower is better. We sample 100 cases from the medical question-answering test sets for rationale and 50 samples from iCliniq (an online medical consulting website) for Conversation. The average ranking for the 5 open-sourced models and the close source GPT-3.5 are shown in the table, the best among the open-source models is bolded.
Abbreviations: CU, Contextual Understanding; KC, Knowledge Correlation; LF, Language Fluency.

Case Study for Instruction Prompting. In Figure 4, we show qualitative examples with the instruction prompting prediction from PMC-LLaMA and GPT-3.5 to verify the quality of prediction, covering patient-physician conversation and rationale QA. The query in Figure 4A is raised online after our data collection, thus, none of the models have seen this question at training time. Based on the patient’s description, both PMC-LLaMA and GPT-3.5 recognize the symptom of recurrent UTIs (urinary tract infections), while PMC-LLaMA proposes a sensitivity test as the specific advice, rather than the general suggestion (investigate the underlying causes) given by GPT-3.5. Figure 4B shows a QA case of microbiology. As can be seen, PMC-LLaMA not only produces an accurate answer but also briefly analyzes the wrong options, forming a more comprehensive rationale. Another case that focuses on pharmacology knowledge is illustrated in Figure 4C. Both PMC-LLaMA and GPT-3.5 have shown to



A

Prompt

Hello, I am a young and healthy teenage girl, but I suffer from many severe UTIs (urinary tract infections).
I get an average of about eight infections annually over the past four years. I follow all of the conventional methods of hygiene, including vitamin C intake, using cotton underwear, etc. I drink about a gallon of water daily and exercise. This problem has spread to my kidneys the initial time and has never escalated to that state.
I developed allergies to all antibiotics, so I am unsure how to treat them at this point. Also, I constantly have cloudy urine and lower back pain.

Please help.

PMC LLaMA

I have read your query and understand your concern. You seem to be having chronic UTI.

You need to get your urine culture and sensitivity test done to know the exact bacteria causing the infection and the antibiotics to which they are sensitive to. Based on the report, you need to take appropriate antibiotics for appropriate duration. Cloudy urine and lower back pain are due to the infection.

Please see a urologist for further management.
Thanks.

GPT-3.5

I'm not a doctor, but here are some points you might consider discussing with a healthcare professional:

1. Specialist Consultation: Given your recurrent UTIs, seeking advice from a urologist could offer insights and tailored recommendations.
2. Alternative Treatments: Inform the urologist about your antibiotic allergies. They might suggest non-antibiotic approaches to manage and prevent UTIs.
3. Investigate Underlying Causes: Your doctor can explore possible underlying factors contributing to your frequent UTIs.
4. Address Symptoms: It's important to discuss your persistent cloudy urine and lower back pain for proper evaluation and management.
5. Medical Guidance: Trusting a medical professional's expertise will guide you toward the best solutions for your situation.

Remember, consulting a healthcare provider is essential for personalized advice and treatment tailored to your individual needs.

B

Prompt

A microbiologist is studying the emergence of a virulent strain of the virus. After a detailed study of the virus and its life cycle, he proposes a theory: Initially, a host cell is co-infected with 2 viruses from the same virus family. Within the host cell, concomitant production of various genome segments from both viruses occurs. Ultimately, the different genome segments from the viruses are packaged into a unique and novel virus particle. The newly formed virus particle is both stable and viable and is a new strain from the virus family that caused the outbreak of infection. Which of the following viruses is capable of undergoing the above-mentioned process?
A. Epstein-Barr virus
B. Human immunodeficiency virus
C. Rotavirus
D. Vaccinia virus

PMC LLaMA

Rotavirus is a double-stranded RNA virus. It is the most common cause of diarrhea in infants and children. The virus has a characteristic wheel-like appearance (from which it derives its name) and undergoes a unique process of viral replication. Initially, a host cell is co-infected with 2 viruses from the same virus family. Within the host cell, concomitant production of various genome segments from both viruses occurs. Ultimately, the different genome segments from the viruses are packaged into a unique and novel virus particle. The newly formed virus particle is both stable and viable and is a new strain from the virus family that caused the outbreak of infection.
Epstein-Barr virus is a herpesvirus that causes infectious mononucleosis.
Human immunodeficiency virus is a retrovirus that causes AIDS. Vaccinia virus is a large, complex double-stranded DNA virus that is related to smallpox.
The correct answer is: C. Rotavirus

GPT-3.5

The correct answer is: C. Rotavirus
The description provided corresponds to the process of reassortment or genetic reassortment, which is commonly seen in segmented viruses like rotavirus. Rotavirus has a segmented genome, and when a host cell is co-infected with different strains of rotavirus, segments from different strains can mix and be packaged into new virus particles. This can result in the emergence of novel strains with a combination of genetic material from different parent strains, leading to changes in virulence or other properties.

C

Prompt

A 26-year-old woman presents to a gynecologist after a missed period. After performing a complete physical examination and a screening test for pregnancy, her gynecologist informs her that she is pregnant. She is very surprised as she has been taking oral contraceptives regularly. When the gynecologist asks her about the consumption of any other medications, she mentions that she was placed on treatment for pulmonary tuberculosis (TB) 2 months ago. Her current anti-TB regimen includes rifampin, isoniazid, pyrazinamide, and ethambutol.

Which of the following mechanisms best explains the failure of oral contraceptives in this patient?

A. Induction of CYP3A4 by rifampin leading to decreased serum levels of ethinylestradiol and progesterone
B. Induction of CYP2A6 by rifampin leading to increased inactivation of ethinylestradiol
C. Interference with the intestinal absorption of the oral contraceptive by pyrazinamide
D. Increased renal elimination of the progesterone component of the oral contraceptive by ethambutol

PMC LLaMA

Rifampin is a microsomal enzyme inducer and can increase the rate of metabolism of various drugs. It can increase the rate of metabolism of the oral contraceptive by inducing CYP3A4, leading to decreased serum levels of ethinylestradiol and progesterone. Rifampin does not affect the intestinal absorption of drugs. Pyrazinamide interferes with the intestinal absorption of the oral contraceptive. Ethambutol does not affect the renal elimination of drugs.

Option A is correct.

GPT-3.5

The correct answer is: A.

Rifampin, one of the medications used in the patient's anti-TB regimen, is known to induce cytochrome P450 enzymes, particularly CYP3A4. This enzyme induction can accelerate the metabolism of many drugs, including oral contraceptives containing ethinylestradiol and progesterone. This results in decreased serum levels of these hormones, which compromises the effectiveness of the oral contraceptive and may lead to contraceptive failure. Therefore, it's important for healthcare providers to counsel patients on using additional or alternative contraceptive methods while taking rifampin or other medications that induce CYP3A4.

Patient-Physician Conversation Microbiology Question Pharmacology Question

Figure 4. Examples of 3 instruction prompting cases from PMC-LLaMA and GPT-3.5. (A) Compares their responses to the patient's query, where PMC-LLaMA proposes more concrete suggestions. (B) Shows the probing of microbiology knowledge. PMC-LLaMA analyzes both correct and incorrect options, enhancing the comprehensiveness of the analysis. Example (C) examines the models' grasp of Pharmacology, and they respond with roughly equivalent answers. The correct options are marked **bold**.

properly understand Rifampin's efficacy and mechanism of side effects.

Discussion

An open-source light-weight language model for medicine

The motivation for this study is to build an open-source LLM for medicine. Although many open-source LLMs have been proposed in the research community, none of them is comparable to GPT-3.5 on medical NLP benchmarks. To bridge the gap, we construct PMC-LLaMA-13B and conduct an investigation of the dataset construction strategies, for example, the contribution of textbooks, conversation dialogue, rationales, etc.

As shown in Table 2, we eventually add training data to show the performance gain. Two critical observations can be observed in the table: *first,* textbooks help general LLMs adapt to the medical domain more than academic papers; *second,* combining different tasks together is beneficial, *third*, the well-known scaling law is also observed in the medical domain, that is, models with more trainable parameters demonstrate superior performance. In comparison to existing open-source models[7,8] tuned with instructions only, we show evidence that further pretraining in the medical domain is critical for the performance of LLMs in downstream tasks and our analysis on training data also accelerates the development progress of a larger scale open-source medical LLM, like MEDITRON-70B.[39]

Open-source versus close-source

Despite the strong ability of close-source LLMs, for example, GPT-4 and Med-PaLM 2, we believe that open-source medical LLMs remain critical for the following reasons:

- **Safety.** In contrast to close-source LLMs, which require users to upload private information online, open-source LLMs can be executed offline without information leaking. This is critical in medical scenarios, for better-protecting patients' privacy. In addition, open-source models potentially enable researchers or clinicians to better understand the fail mode, ensuring that they are put into appropriate clinical scenarios, while for the close-source LLMs, their boundaries of the capabilities can only be partially understood through limited interactions, which may lead to severe over-trust of LLMs in clinical practice.
- **Customization.** Open-source LLMs provide the flexibility to be customized based on specific healthcare needs or regional medical practices. Developers and researchers can modify these models to fit particular languages, medical terminologies, or unique healthcare challenges for different communities, which is often not possible with close-source models.
- **Accessibility.** Being open-source, these models are more accessible to a diverse range of users, including researchers, clinicians, and institutions in low-resource settings. This democratizes access to advanced AI technology, allowing for wider adoption and benefiting various healthcare systems, preventing monopolization of high-quality healthcare resources.

Future directions

As future work, we will continually improve or exploit PMC-LLaMA toward following directions:

- **Multi-Modal Foundation Models.** PMC-LLaMA, as a powerful generative language model, can be extended to generate free-form text response based on the two multi-modal inputs, as demonstrated in many following works.[40–43]
- **Retrieval-Augmented Methods.** Hallucination is a fatal weakness for LLMs, which is even more catastrophic for medical applications. PMC-LLaMA also suffers from such issue. A promising direction is to explore retrieval-augmentation methods.[44–46] Generally speaking, by retrieving 100% correct, and useful facts from an external knowledge memory basis, the LLMs may have the potential to avoid fact hallucination and can edit their knowledge base faster.
- **Tool-Augmentation System (Language Agents).** PMC-LLaMA also has the potential to be customized into a medical agent controller, to support interaction with other accurate medical tools.[47–49] The key insight is that LLMs can decompress a complex problem into a series of subtasks that can be solved by more accurate specific agents. This is also crucial for healthcare, as the specific agents can often be more precise.

Limitations

There are also some limitations in this work. *First*, our experiments are mostly carried on 7B and 13B models, we hope to try out on even larger models when granted with more computation resources. *Second*, despite being more diverse than all existing work, the medical tasks involved in instruction tuning are still limited. *Third*, we will improve PMC-LLaMA to pick up in-context learning abilities, and improve performance on more medical tasks.

Conclusion

In this article, we propose to systematically investigate the procedure for building up a medical-specific LLM based on an open-source LLM, including data-centric knowledge injection and medical-specific instruction tuning. As a result, our proposed PMC-LLaMA is the first, open-source medical-specific language model, that demonstrates superior performance on various medical benchmarks, surpassing GPT-3.5 and LLaMA-2 with much fewer parameters.

Author contribution

All listed authors clearly meet the ICMJE 4 criteria. Specifically, C.W., W.L., X.Z., Y.Z., Y.W., and W.X. all make contributions to the conception or design of the work, and C.W., W.L. further perform acquisition, analysis, or interpretation of data for the work. In writing, C.W., W.L. draft the work and X.Z., Y.Z., Y.W., and W.X. review it critically for important intellectual content. All authors approve of the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. C.W. and W.L. contribute equally to the work and Y.W. is the corresponding author.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported by the National Key R&D Program of China, grant number 2022ZD0160702, Science and Technology Commission of Shanghai Municipality (STCSM) grant number 22511106101, 18DZ2270700, 21DZ1100100, Higher Education Discipline Innovation Project 111 plan grant number BP0719010, and State Key Laboratory of UHD Video and Audio Production and Presentation.

Conflicts of interest

None declared.

Data availability

All our data, models, and codes are open-source. Specifically, the academic papers used in our knowledge injection step can be found at https://allenai.org/data/s2orc. The textbooks are listed at https://github.com/chaoyi-wu/PMC-LLaMA/blob/main/MedicalBook.xlsx. Considering the copyrights, we cannot release the raw book contents, and our followers may buy and collect the books themselves to reproduce our experiments. The instruction data are directly released at https://huggingface.co/datasets/axiong/pmc_llama_instructions. All source code can be found at https://github.com/chaoyi-wu/PMC-LLaMA. Model checkpoints can also be found following the GitHub "ReadMe" guidance.

References

1. OpenAI. OpenAI. Introducing ChatGPT. OpenAI; 2023. Accessed April 2, 2024. https://openai.com/blog/chatgpt/
2. OpenAI. GPT-4 technical report. arXiv:2303.08774. 2023, preprint: not peer reviewed.
3. Nori H, King N, et al. Capabilities of GPT-4 on medical challenge problems. arXiv 230313375. 2023, preprint: not peer reviewed.
4. Singhal K, Azizi S, Tu TAO, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180.
5. Touvron H, Lavril T, Izacard G, Martinet , et al. LLaMA: open and efficient foundation language models. arXiv 230213971. 2023, preprint: not peer reviewed.
6. Touvron H, Martin L, Stone KR, et al. LLaMA 2: open foundation and fine-tuned chat models. arXiv abs/2307.09288. 2023, preprint: not peer reviewed. https://api.semanticscholar.org/CorpusID:259950998
7. Han T, Adams LC, et al. MedAlpaca - an open-source collection of medical conversational AI models and training data. arXiv 230408247. 2023, preprint: not peer reviewed.
8. Yunxiang L, et al. Chatdoctor: a medical chat model fine-tuned on LLaMA model using medical domain knowledge. arXiv 230314070. 2023, preprint: not peer reviewed.
9. Anil R, Dai AM, et al. PaLM 2 technical report. arXiv abs/2305.10403. 2023, preprint: not peer reviewed. https://api.semanticscholar.org/CorpusID:258740735
10. Du Z, Qian Y, Liu X, et al. *GLM: General Language Model Pretraining with Autoregressive Blank Infilling*. ACL; 2022. Accessed April 2, 2024. https://aclanthology.org/2022.acl-long.26
11. Ouyang L, et al. Training language models to follow instructions with human feedback. *NIPS*. 2022;35:27730-27744.
12. Peng B, et al. Instruction tuning with GPT-4. arXiv 230403277. 2023, preprint: not peer reviewed.
13. Chung HW, et al. Scaling instruction-finetuned language models. arXiv 221011416. 2022, preprint: not peer reviewed.
14. Iyer S, et al. OPT-IML: scaling language model instruction meta learning through the lens of generalization. arXiv 221212017. 2022, preprint: not peer reviewed.
15. Taori R, et al. *Stanford Alpaca: An Instruction-Following LLaMA Model*. GitHub; 2023. Accessed April 2, 2024. https://github.com/tatsu-lab/stanford_alpaca
16. Chiang WL, Li Z, et al. *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality*. LMSYS Org; 2023. Accessed April 2, 2024. https://vicuna.lmsys.org
17. 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.
18. Zhang K, et al. BiomedGPT: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. arXiv 230517100. 2023, preprint: not peer reviewed.
19. 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.
20. Lo K, Wang LL, Neumann M, Kinney R, Weld D. *S2ORC: The Semantic Scholar Open Research Corpus*. ACL; 2020:4969-4983. Accessed April 2, 2024. https://aclanthology.org/2020.acl-main.447. doi: 10.18653/v1/2020.acl-main.447
21. Together Computer. *RedPajama: An Open Dataset for Training Large Language Models*. Together AI; 2023. Accessed April 2, 2024. https://github.com/togethercomputer/RedPajama-Data
22. Jin D, Pan E, Oufattole N, et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl Sci*. 2021;11(14):6421.
23. Jin Q, et al. PubMedQA: a dataset for biomedical research question answering. *ACL*. 2019:2567-2577.
24. Pal A, Umapathi LK, et al. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. In: *Conference on Health, Inference, and Learning*. PMLR; 2022:248-260.
25. Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. *Yearb Med Inform*. 1993;2(1):41-51.
26. Loshchilov I, Hutter F. *Decoupled Weight Decay Regularization*. ICLR; 2018.
27. Chen T, et al. Training deep nets with sublinear memory cost. arXiv 160406174. 2016, preprint: not peer reviewed.
28. Diao S, et al. *LMFlow: An Extensible Toolkit for Finetuning and Inference of Large Foundation Models*. GitHub; 2023. Accessed April 2, 2024. https://optimalscale.github.io/LMFlow/
29. 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*. 2022;2(2):e0000198. https://api.semanticscholar.org/CorpusID:254876189
30. Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. *npj Digit. Med*. 2023;6:210. https://doi.org/10.1038/s41746-023-00958-w
31. Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. *Minds Mach*. 2020;30(4):681-694.
32. Venigalla A, Frankle J, Carbin M. *BioMedLM: A Domain-Specific Large Language Model for Biomedical*. Stanford University; 2022. Accessed April 2, 2024. https://www.mosaicml.com/blog/introducing-pubmed-gpt
33. Christophe C, Gupta A, Hayat N, et al. *Med42 - A Clinical Large Language Model*. M42; 2023.
34. Mangrulkar S, Gugger S, Debut L, Belkada Y, Paul S. *PEFT: State-of-the-Art Parameter-Efficient Fine-Tuning Methods*. Hugging Face; 2022. Accessed April 2, 2024. https://github.com/huggingface/peft
35. Hu EJ, Shen Y, et al. *LoRA: Low-Rank Adaptation of Large Language Models*. ICLR; 2021.
36. Liu X, Ji K, Fu Y, et al. *P-Tuning: Prompt Tuning Can Be Comparable to Fine-Tuning Across Scales and Tasks*. ACL; 2022:61-68. Accessed April 2, 2024. https://aclanthology.org/2022.acl-short.8
37. Kaplan J, McCandlish S, Henighan TJ, et al. Scaling laws for neural language models. arXiv abs/2001.08361. 2020, preprint: not peer reviewed. https://api.semanticscholar.org/CorpusID:210861095
38. Yang X, et al. GatorTron: a large clinical language model to unlock patient information from unstructured electronic health records. arXiv abs/2203.03540. 2022, preprint: not peer reviewed. https://api.semanticscholar.org/CorpusID:247157824
39. Chen Z, Cano AH, Romanou A, et al. MEDITRON-70B: scaling medical pretraining for large language models. arXiv abs/2311.16079. 2023, preprint: not peer reviewed. https://api.semanticscholar.org/CorpusID:265456229
40. Moor M, Huang Q, Wu S, et al. Med-Flamingo: a multimodal medical few-shot learner. In: *Proceedings of the 3rd Machine Learning for Health Symposium*. PMLR; 2023:225:353-367. https://proceedings.mlr.press/v225/moor23a.html
41. Li C, Wong C, Zhang S, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. *NIPS*. 2023;36;28541-28564.
42. Wu C, Zhang X, Zhang Y, Wang Y, Xie W. Towards generalist foundation model for radiology. arXiv abs/2308.02463. 2023,

preprint: not peer reviewed. https://api.semanticscholar.org/CorpusID:260611504

43. Tu T, Azizi S, Driess D, et al. Towards generalist biomedical AI. *NEJM AI*. 2024;1(3):AIoa2300138.
44. Jin Q, Yang Y, Chen Q, Lu Z. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*. 2024;40(2):btae075.
45. Zakka C, Shad R, Chaurasia A, et al. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*. 2024;1(2):AIoa2300068. https://ai.nejm.org/doi/abs/10.1056/AIoa2300068
46. Liu J, Jin J, Wang Z, Cheng J, Dou Z, Rong Wen J. RETA-LLM: a retrieval-augmented large language model toolkit. arXiv abs/2306.05212. 2023, preprint: not peer reviewed. https://api.semanticscholar.org/CorpusID:259108339
47. Schick T, Dwivedi-Yu J, Dessì R, et al. Toolformer: language models can teach themselves to use tools. *NIPS*. 2024;36:68539-68551.
48. Yang R, Song L, Li Y, et al. GPT4Tools: teaching large language model to use tools via self-instruction. *NIPS*. 2023;36:71995-72007.
49. Sumers TR, Yao S, Narasimhan K, Griffiths TL. Cognitive architectures for language agents. arXiv abs/2309.02427. 2023, preprint: not peer reviewed. https://api.semanticscholar.org/CorpusID:261556862

© The Author(s) 2024. Published by Oxford University Press on behalf of the American Medical Informatics Association.
All rights reserved. For permissions, please email: journals.permissions@oup.com
Journal of the American Medical Informatics Association, 2024, 31, 1833–1843
https://doi.org/10.1093/jamia/ocae045
Research and Applications