Artificial Intelligence In Medicine 158 (2024) 103007

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Exploring the effectiveness of instruction tuning in biomedical language processing

Omid Rohanian [a,b,*,1], Mohammadmahdi Nouriborji [b,c,1], Samaneh Kouchaki [d], Farhad Nooralahzadeh [e], Lei Clifton [f], David A. Clifton [a,g]

[a] Department of Engineering Science, University of Oxford, Oxford, UK
[b] NLPie Research, Oxford, UK
[c] Sharif University of Technology, Tehran, Iran
[d] Dept. Electrical and Electronic Engineering, University of Surrey, Guildford, UK
[e] University of Zürich and University Hospital of Zürich, Zürich, Switzerland
[f] Nuffield Department of Population Health, University of Oxford, Oxford, UK
[g] Oxford-Suzhou Centre for Advanced Research, Suzhou, China

ARTICLE INFO

Keywords:
Instruction tuning
Biomedical NLP
Named entity recognition
Relation extraction
Medical NLI
Llama2-MedTuned

ABSTRACT

Large Language Models (LLMs), particularly those similar to ChatGPT, have significantly influenced the field of Natural Language Processing (NLP). While these models excel in general language tasks, their performance in domain-specific downstream tasks such as biomedical and clinical Named Entity Recognition (NER), Relation Extraction (RE), and Medical Natural Language Inference (NLI) is still evolving. In this context, our study investigates the potential of instruction tuning for biomedical language processing, applying this technique to two general LLMs of substantial scale. We present a comprehensive, instruction-based model trained on a dataset that consists of approximately 200,000 instruction-focused samples. This dataset represents a carefully curated compilation of existing data, meticulously adapted and reformatted to align with the specific requirements of our instruction-based tasks. This initiative represents an important step in utilising such models to achieve results on par with specialised encoder-only models like BioBERT and BioClinicalBERT for various classical biomedical NLP tasks. Our work includes an analysis of the dataset's composition and its impact on model performance, providing insights into the intricacies of instruction tuning. By sharing our codes, models, and the distinctively assembled instruction-based dataset, we seek to encourage ongoing research and development in this area.[2]

1. Introduction

Transformers have become the cornerstone of modern NLP, providing the backbone for a wide array of applications including machine translation, question-answering, and text summarisation [1]. Their self-attention mechanisms and parallelised architecture have proven to be highly effective in capturing the nuances of human language [2].

Autoregressive language models, exemplified by the Generative Pre-trained Transformer series like GPT [3] and GPT-3 [4], have revolutionised the way NLP is approached. These models, operating as decoder-only transformers, excel at generating text in a sequential, token-by-token manner, leveraging their attention mechanisms to focus on relevant segments of input text. Models based on this architecture, such as GPT-4 have demonstrated a remarkable ability to perform a variety of language tasks without the need for task-specific fine-tuning, showcasing strong zero-shot and few-shot learning capabilities. This feature allows these models to effectively respond to text-based prompts, including those with a limited number of examples or instructions, thereby enabling a more interactive and dynamic text generation process.

Medical language models, particularly encoder-only models like BioBERT and ClinicalBERT, have been instrumental in advancing tasks such as medical diagnosis, biomedical literature mining, and clinical information extraction [5,6]. Excelling in areas like classification and Named Entity Recognition (NER), these models have significantly contributed to biomedical NLP. However, they often lack inherent

* Corresponding author.
E-mail address: omid.rohanian@eng.ox.ac.uk (O. Rohanian).
[1] These authors contributed equally to this work.
[2] Our code repository is available at https://github.com/nlpie-research/BioInstTune-LLM

https://doi.org/10.1016/j.artmed.2024.103007
Received 2 April 2024; Received in revised form 4 October 2024; Accepted 23 October 2024
Available online 7 November 2024

capabilities in interpreting and executing natural language instructions or generating reports from medical Electronic Health Records (EHRs). This limitation has spurred research into developing generative Large Language Models (LLMs) capable of handling more dynamic tasks, aiming to parallel the performance of specialised encoder-only models in the biomedical domain. Yet, as indicated by studies such as Lehman et al. [7], encoder-only models continue to lead in clinical NLP, underscoring the challenges in tailoring general-domain LLMs for specialised medical applications.[3] Our research aims to contribute to this area by introducing a dataset that integrates various clinical and biomedical datasets. Utilising this resource, we apply instruction tuning to two publicly available general LLMs, with the objective of exploring its potential in enhancing the performance of these LLMs for downstream medical tasks. This approach represents an initial step towards understanding the effectiveness of instruction tuning in this domain, with the dataset serving as an additional tool to facilitate this exploration for future work.

The primary contributions of our work are as follows. First, we introduce Llama2-MedTuned, developed in two variants: one fine-tuned on the Llama2 7B model[4] and the other on the Llama2 13B model.[5] These are specialised models designed explicitly for instruction-based tasks in the medical domains. Second, we present a dataset that amalgamates various publicly available datasets into a novel configuration, creating a rich and diverse training environment specifically compiled for the Llama2-MedTuned models. Our comparative experimental results highlight the effectiveness of our approach in comparison to current state-of-the-art models in a number of classical tasks in biomedical and clinical NLP (see Fig. 1).

2. Related works

2.1. Autoregressive language models

Autoregressive Language Models (ALMs), exemplified by GPT and its different variants, constitute a class of transformers pre-trained on a language modelling objective, namely, predicting the subsequent token given a particular context [3,8]. Noteworthy instances of ALMs include GPT-3.5 and GPT-4 by OpenAI, trained on extensive datasets harvested from the web for the language modelling objective [4]. Google's Bard/Gemini and Anthropic's Claude are also notable contributions to this field, demonstrating the growing exploration and advancements of autoregressive language models for diverse applications.

2.2. Instruction-based language models

Instruction-based language models, a novel category within autoregressive models, have been shown to improve significantly when fine-tuned with instructions. Traditional autoregressive models, while adept at sequential text generation, often struggle with comprehending and executing complex instructions. Fine-tuning such models on natural language instructions and human-generated responses can markedly enhance their ability to follow instructions accurately [9]. This advancement is exemplified in models like Instruct-GPT [10], Falcon [11], and Llama [12], which are fine-tuned to respond more effectively to instruction-based prompts, thus enabling more dynamic and interactive text generation capabilities.

2.3. Clinical LLMs

With the advent of instruction-based LLMs, their adaptation to the clinical domain has been explored, using instruction-based datasets specific to this area. ChatDoctor [13], a fine-tuned clinical chatbot, has been trained on real conversations between doctors and patients, showcasing its efficacy in clinical settings. Similarly, Med-Alpaca [14] and Clinical Camel [15] follow this trend by adapting open LLMs to the clinical domain. PMC-Llama [16] is another significant model, initially pre-trained on a biomedical/clinical corpus, and subsequently trained on an instruction dataset primarily containing medical question answering and reasoning tasks.

3. Method

In this work, we train an instruction-based language model for the medical domain which is able to target tasks such as Named Entity Recognition, Relation Extraction, Document Classification, Question Answering, and Natural Language Inference (see Fig. 2). In order to train this model, we compiled a new medical instruction-based dataset called **Llama2-MedTuned-Instructions**.[6]

3.1. Prompting template

To transform the original datasets into instruction-based formats, we adopted the prompting strategy used in the Alpaca dataset. Our prompts are composed of three parts: Instruction, Input, and Output. In the Instruction section, we developed 5 to 10 different instructions for each dataset, detailing the target tasks and the labelling scheme for the model. One instruction is randomly chosen for each sample during the conversion to the instruction-based dataset. The Input is the dataset's original input, while the Output is the expected output that the model should predict, consistent with the format described in the instructions. Fig. 3 presents some samples from our instruction dataset.

3.2. Tasks and datasets

As mentioned earlier, various tasks are used in this work to diversify the training corpus used for training our language model. Training subsets from several well-known datasets were selected for each task to assemble the dataset employed in our study.

3.2.1. Named entity recognition

For the task of Named Entity Recognition, we used the NCBI-disease, BC5CDR-disease [17], BC5CDR-chem [18], BC2GM [19], JNLPBA [19], and i2b2-2012 dataset [20]. For the first five datasets, we use the BIO labelling scheme with no additional label names. However, for the i2b2-2012 dataset, 6 different categories are used along with BIO labelling.

3.2.2. Relation extraction

We used the i2b2-2010 [21] and GAD [22] datasets for relation extraction. For both datasets we follow the same pre-processing method used in [23] and [24], which uses specific tags (e.g. test$, problem$, etc.) for tagging medical concepts in the text, in order to frame the relation extraction as a sentence classification task.

3.2.3. Natural language inference

For Natural Language Inference, we used the MedNLI dataset [25], which is composed of pairs of medical sentences labelled with Entailment, Contradiction, or Neutral to indicate the type of relationship between them.

[3] This could be attributed to the general nature of LLMs, which often lack the specialised biomedical domain knowledge encoded in models like BioBERT. Additionally, while LLMs benefit from instruction-tuning, they may struggle with structured tasks requiring precise token-level predictions, where smaller, task-specific models tend to excel.

[4] Llama2-MedTuned-7b: https://huggingface.co/nlpie/Llama2-MedTuned-7b.

[5] Llama2-MedTuned-13b:https://huggingface.co/nlpie/Llama2-MedTuned-13b.

[6] https://huggingface.co/datasets/nlpie/Llama2-MedTuned-Instructions.

Named Entity Recognition

Instruction:
Your task is to identify clinical Named Entities within the text and apply the BIO labeling scheme. Use the following labels to categorize each entity:

OCCURRENCE: If the entity represents a clinical incident or event.
PROBLEM: If the entity indicates a medical problem.
TEST: If the entity pertains to a medical test.
TREATMENT: If the entity refers to a medical treatment.
EVIDENTIAL: If the entity provides evidence.
CLINICAL_DEPT: If the entity relates to a clinical department.
O: If the entity doesn't fit into any of the above categories.

Input:
The urine contained yeast and she was treated with amphotericin .

Output:
The : B-TEST
urine : I-TEST
contained : O
yeast : B-PROBLEM
and : O
she : O
was : O
treated : O
with : O
amphotericin : B-TREATMENT
. : O

General Medical Instructions

Instruction:
In your capacity as a medical assistant, help the user's with their medical inquiries, and follow their instructions.

Input:
One of my friends has recently have a heart-storke, Give me a list of foods that might be dangerous for him, in an ordered list.

Output:

Following is a list of foods that should be avoided or taken in moderation after a heart attack:

1. Red meat and high fat milk products: These should be taken in moderation.
2. Fried foods: Deep-fried foods and foods cooked in a lot of oil should be taken in moderation.
3. High fat foods: Foods that are high in saturated or trans-fats should be taken in moderation.
4. Processed meats: Bacon, sausage, bologna, salami, etc. should be taken in moderation.
5. Canned foods: Canned foods that are high in sodium should be taken in moderation.
6. Sweets: Cakes, cookies, pies, and other sweets should be taken in moderation.
7. Salt: High amounts of salt should be avoided.
8. Alcohol: Alcohol should be avoided or taken in moderation.
9. Caffeine: Caffeine should be taken in moderation.
10. Cholesterol: Foods that are high in cholesterol should be taken in moderation.

Fig. 1. Example outputs from Llama2-MedTuned-7B for biomedical tasks (left) and general medical instructions (right). The model demonstrates the application of instruction-based learning in NER by correctly labelling biomedical entities (left) and providing a relevant list in response to a medical inquiry (right).

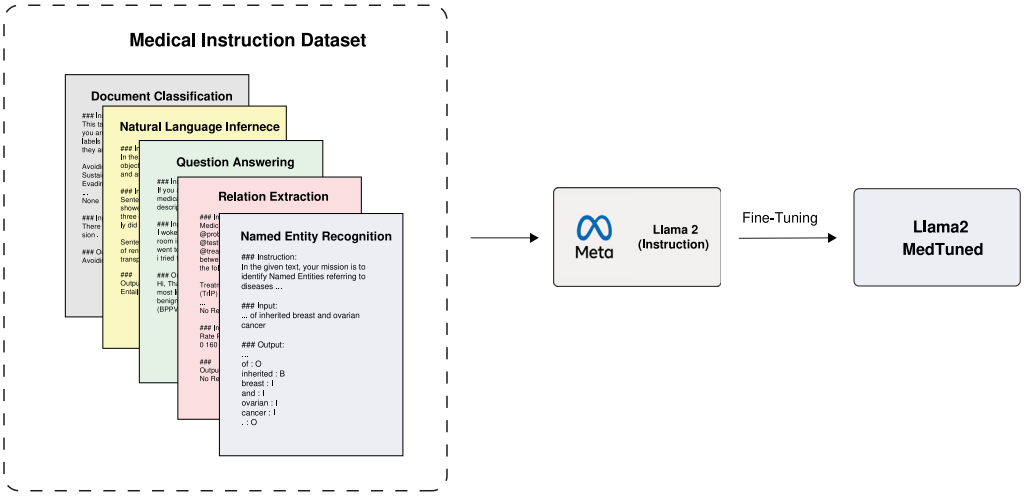

Fig. 2. Schematic representation of the process for fine-tuning Llama2 models with the proposed medical instruction dataset.

3.2.4. Document classification

We used the hallmarks of cancer (HoC) dataset [26] for the task of Document Classification which is a well-known multi-class classification dataset in the medical domain.

3.2.5. Question answering

For question answering, we used two prominent datasets, ChatDoctor [13], and Pmc-Llama-Instructions [16]. ChatDoctor consists of 100k samples taken from the ChatDoctor website that are real conversations between patients and doctors, In our dataset we randomly sampled 50K samples from this dataset. PMC-Llama-Instructions is a large dataset consisting of multiple QA datasets such as MedQA [27], PubMedQA [28], etc. For our work, we randomly sampled 50K samples from this dataset.

3.2.6. Llama2-MedTuned instructions

Finally, we concatenate all of the datasets mentioned earlier in this section and shuffle them to obtain our final medical instruction-based dataset which consists of approximately 200K samples.

The final portions of the tasks within the fine-tuning dataset are summarised in Table 1:

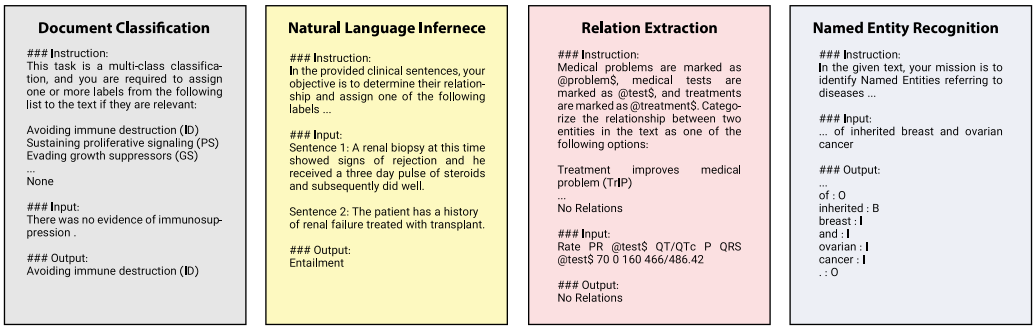

Fig. 3. Overview of some of the prompt templates used in our instruction dataset.

Table 1
Portions of the tasks within the fine-tuning dataset.

Task source	Percentage
Chatdoctor	28.0%
PMC-Llama	25.0%
i2b2–2010	11%
JNLPBA	7.3%
BC2GM	6.3%
HoC	6.1%
MedNLI	5.6%
i2b2–2012	3.4%
NCBI-Disease	2.7%
BC5CDR-Disease	2.3%
BC5CDR-Chem	2.3%

3.3. Training configuration

In order to train our models, we used 10 V100 GPUs with a batch size of 4 per GPU. We used the deepspeed zero 3 config without CPU offloading, with a learning rate of $1e-5$ and 500 warmup steps along with a linear learning rate scheduler. The models were trained for three epochs.

4. Results

Assessing the instruct-tuned models, Llama2-MedTuned 7B and 13B, against their foundational counterparts, Llama2 7B and 13B, presents several challenges. As shown in Figs. A.4 and A.5 in the Appendix, the outputs from the base Llama2 models for NER are often inconsistent and difficult to evaluate. The exception to this pattern is the MedNLI task, where Llama2 produced more consistent and stable outputs. For the remaining tasks, we compare the performance of our instruction-tuned models with conventional baselines such as DistilBERT and BioBERT across NER, RE, and NLI tasks.

Our study focuses on zero-shot learning scenarios, as we found that adding a few examples in few-shot learning or rewording prompts to encourage chain-of-thought reasoning did not significantly alter the results. Therefore, we prioritised the zero-shot template to maintain simplicity and consistency across model comparisons.

Thanks to instruction-tuning, we were able to systematically interpret our models' outputs into a structured format, suitable for evaluation using conventional metrics like F1 or Accuracy. The results for the biomedical NER are available in Table 2, Where the 13B model is generally better than our 7B model. Additionally, the results of the clinical tasks are available in Table 3.

Generally, interpreting the outputs of Llama2 on most structured tasks proved to be challenging as the outputs tended to deviate from the expected format. We have provided examples of output generations from both our model and Llama2 in Figs. A.4 and A.5. [29] reports results for the NER datasets on a number of closed and open LLMs including LLama2. Please refer to Table 5 for a baseline reference to the reported results on the NER tasks in the literature. Llama2, on the other hand, did yield consistent outputs on the MedNLI task. Upon evaluation, the Llama2 model scored an accuracy of 37.20 on the MedNLI evaluation subset, significantly lower than the 89.46 score achieved by Llama2-MedTuned-13b.

5. Ablation studies

To maintain the general capabilities of our model on tasks such as Question Answering and general instructions we use additional instruction-based data along with our NER, RE, and CLS instructions. We tested two strategies to create our final dataset. First, we randomly sampled 50K samples from the PMC-Llama instructions, and 50K from the ChatDoctor. For the second approach, we employed a more balanced sampling by taking 50K samples from PubMedQA, 50K from MedQA, 100% of UMLS relations, and UMLS, which resulted in 200K samples from the PMC-Llama instructions, along with 50K samples from ChatDoctor. The ablation study results, presented in Table 4, reveal that the model trained on the larger PMC-Llama dataset exhibited inferior performance in biomedical downstream tasks compared to the model trained on the smaller dataset.

6. Conclusions & future works

In our study, we focused on instruction tuning of the Llama 2 model using a bespoke biomedical dataset, specifically curated for specialised biomedical NLP tasks like Named Entity Recognition (NER), Relation Extraction (RE), and medical Natural Language Inference (NLI). This process led to the creation of Llama2-MedTuned-7B and Llama2-MedTuned-13B, which represent adaptations of the original Llama 2 models. These tuned versions showed significant improvements in handling the complexities of medical NER, RE, and NLI, indicating the efficacy of instruction tuning in aligning general-purpose language models with specialised task requirements.

While the most substantial performance gains were observed in tasks such as MedNLI, where Llama2-MedTuned demonstrated a significant margin over baseline models, we acknowledge that the improvements were more modest in other tasks like NER and RE. However, the aim of this work was not solely to surpass smaller models across all tasks, but rather to explore the broader applicability of instruction-tuned LLMs in the biomedical domain. Llama2-MedTuned showed competitive results in some tasks, such as JNLPBA, while offering unique advantages such as scalability, flexibility, and adaptability to new tasks, which smaller, encoder-only models may lack. For instance, LLMs can accommodate new NER tags or entities unseen during training, making them well-suited for data-poor scenarios where traditional models may struggle.

We do not advocate for instruction-based tuning in all scenarios but highlight its potential to bridge the gap between large-scale, versatile

[7] The results are taken from [29].

Table 2
Test results on the biomedical downstream tasks.

Type	Task	DistilBERT	BioBERT-v1.1	Llama2-MedTuned-7b	Llama2-MedTuned-13b
NER	NCBI-Disease	86.38	88.62	87.18	85.69
NER	BC5CDR-Disease	82.01	86.67	83.92	85.46
NER	BC5CDR-Chem	92.50	94.73	93.88	94.51
NER	BC2GM	84.61	87.62	76.46	79.12
NER	JNLPBA	79.14	80.33	82.30	81.31

Table 3
Test results on the clinical downstream tasks.

Type	Task	DistilBERT	BioClinicalBERT	Llama2-MedTuned-7b	Llama2-MedTuned-13b
NER	i2b2–2012	79.15	82.98	80.67	80.64
RE	i2b2–2010	92.75	93.58	89.35	93.14
NLI	MedNLI	73.41	82.41	79.21	89.46

Table 4
Ablation study results using the instruction-based dataset.

Type	Task	Llama2-MedTuned	Llama2-MedTuned[a]
NER	NCBI-Disease	85.69	83.59
NER	BC5CDR-Disease	85.46	84.30
NER	BC5CDR-Chem	94.51	93.77
NER	BC2GM	79.12	78.51
NER	JNLPBA	81.31	78.91

[a] Denotes the model trained with the expanded instruction dataset.

Table 5
Baseline results of different language models on the biomedical NER tasks.[7]

Dataset	GPT-3.5	Llama-2	Claude-2
NCBI-disease	33.39	4.58	45.75
BC2GM	31.99	5.95	40.45
BC5CDR-chem	41.25	12.21	58.05
BC5CDR-disease	32.26	5.68	50.13
JNLPBA	31.89	4.30	34.62

models and specialised biomedical tasks that require structured output. We believe that as LLM fine-tuning for domain-specific applications evolves, further improvements in instruction-tuning techniques and task-specific datasets will enhance the performance of these models beyond what was achieved in this initial study.

Future work will focus on expanding our dataset to include a wider variety of biomedical and clinical tasks. Additionally, we plan to explore integrating more recent advancements in language models to continually refine our approach and better address the evolving challenges of biomedical NLP applications.

CRediT authorship contribution statement

Omid Rohanian: Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Mohammadmahdi Nouriborji:** Software, Resources, Methodology, Data curation, Conceptualization. **Samaneh Kouchaki:** Writing – review & editing, Supervision, Resources. **Farhad Nooralahzadeh:** Writing – review & editing, Formal analysis, Conceptualization. **Lei Clifton:** Supervision. **David A. Clifton:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

Funding

This work was supported in part by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC), and in part by an InnoHK Project at the Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE). OR acknowledges the support of the Medical Research Council (grant number MR/W01761X/). DAC was supported by an NIHR Research Professorship, an RAEng Research Chair, COCHE, the UKRI, and the Pandemic Sciences Institute at the University of Oxford. The views expressed are those of the authors and not necessarily those of the NIHR, MRC, COCHE, UKRI, or the University of Oxford.

Limitations

Our exploration into the application of large autoregressive language models like Llama2 for NLP tasks such as NER and RE unveiled significant challenges. The base Llama2 models, without fine-tuning, struggled to generate coherent and appropriately formatted outputs for these tasks. This underscores the difficulty in applying general-purpose LLMs to domain-specific tasks that demand organised responses. However, our instruct-tuned models, Llama2-MedTuned 7B and 13B, showed improved performance, successfully generating outputs in the necessary structured format. Despite this advancement, they did not outperform specialised models like BioBERT, highlighting a need for further development to meet the precision requirements of specific biomedical NLP tasks.

Declaration of competing interest

The authors declare the following financial relationships: This research was partially funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and by an InnoHK Project at the Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE). Omid Rohanian has received grant support from the Medical Research Council (grant number MR/W01761X/1). David A. Clifton is supported by an NIHR Research Professorship, an RAEng Research Chair, the UKRI, COCHE, and the Pandemic Sciences Institute at the University of Oxford.

None of the authors have any employment, consultancies, stock ownership, honoraria, paid expert testimony, or patent applications/registrations that could be considered as potential conflicts of interest affecting this work. The funding sources had no involvement in the study design, in the collection, analysis, and interpretation of data, in the writing of the report, or in the decision to submit the article for publication.

None beyond the funding sources mentioned.

Appendix

See Figs. A.4–A.7.

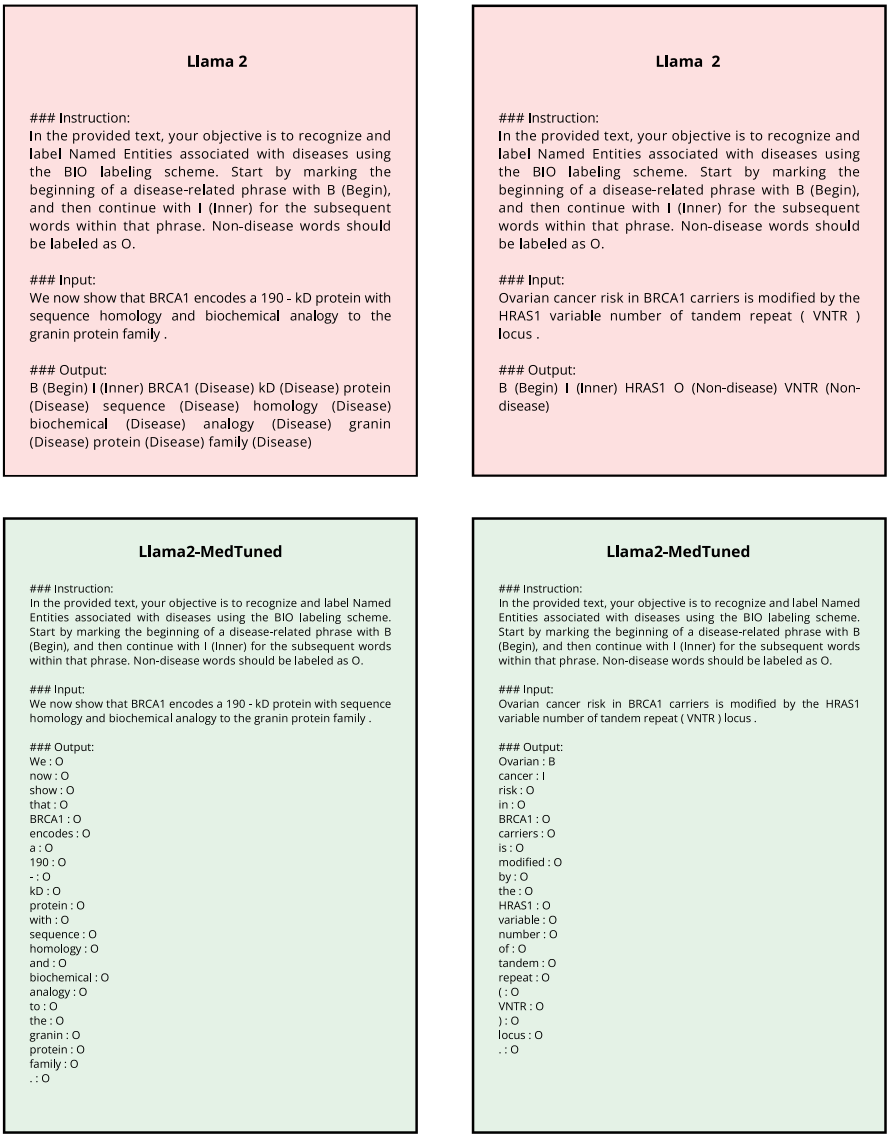

Fig. A.4. Sample outputs of the Llama2 model and Llama2-MedTuned on Named Entity Recognition.

Llama 2

Instruction:
Your task is to determine the relationships between medical problems, treatments, and tests within the clinical text. Medical problems are marked as @problem$, medical tests are marked as @test$, and treatments are marked as @treatment$. Categorize the relationship between two entities in the text as one of the following options:
Treatment improves medical problem (TrIP)
Treatment worsens medical problem (TrWP)
Treatment causes medical problem (TrCP)
Treatment is administered for medical problem (TrAP)
Treatment is not administered because of medical problem (TrNAP)
Test reveals medical problem (TeRP)
Test conducted to investigate medical problem (TeCP)
Medical problem indicates medical problem (PIP)
No Relations

Input:
She has an @problem$ controlled with @treatment$

Output:
Treatment improves medical problem (TrIP)
Treatment worsens medical problem (TrWP)
Treatment causes medical problem (TrCP)
Treatment is administered for medical problem (TrAP)
...

Llama 2

Instruction:
Your task is to determine the relationships between medical problems, treatments, and tests within the clinical text. Medical problems are marked as @problem$, medical tests are marked as @test$, and treatments are marked as @treatment$. Categorize the relationship between two entities in the text as one of the following options:
Treatment improves medical problem (TrIP)
Treatment worsens medical problem (TrWP)
Treatment causes medical problem (TrCP)
Treatment is administered for medical problem (TrAP)
Treatment is not administered because of medical problem (TrNAP)
Test reveals medical problem (TeRP)
Test conducted to investigate medical problem (TeCP)
Medical problem indicates medical problem (PIP)
No Relations

Input:
This is an 83 y/o female with moderate AS , s/p MVR 27 years ago , s/p PPM , @problem$, CHF , p/w dsypnea and @problem$ x 4-5 days

Output:
Treatment improves medical problem (TrIP)
Treatment worsens medical problem (TrWP)
Treatment causes medical problem (TrCP)
...

Llama2-MedTuned

Instruction:
Your task is to determine the relationships between medical problems, treatments, and tests within the clinical text. Medical problems are marked as @problem$, medical tests are marked as @test$, and treatments are marked as @treatment$. Categorize the relationship between two entities in the text as one of the following options:

Treatment improves medical problem (TrIP)
Treatment worsens medical problem (TrWP)
Treatment causes medical problem (TrCP)
Treatment is administered for medical problem (TrAP)
Treatment is not administered because of medical problem (TrNAP)
Test reveals medical problem (TeRP)
Test conducted to investigate medical problem (TeCP)
Medical problem indicates medical problem (PIP)
No Relations

Input:
She has an @problem$ controlled with @treatment$

Output:
TrIP

Llama2-MedTuned

Instruction:
Your task is to determine the relationships between medical problems, treatments, and tests within the clinical text. Medical problems are marked as @problem$, medical tests are marked as @test$, and treatments are marked as @treatment$. Categorize the relationship between two entities in the text as one of the following options:

Treatment improves medical problem (TrIP)
Treatment worsens medical problem (TrWP)
Treatment causes medical problem (TrCP)
Treatment is administered for medical problem (TrAP)
Treatment is not administered because of medical problem (TrNAP)
Test reveals medical problem (TeRP)
Test conducted to investigate medical problem (TeCP)
Medical problem indicates medical problem (PIP)
No Relations

Input:
This is an 83 y/o female with moderate AS , s/p MVR 27 years ago , s/p PPM , @problem$, CHF , p/w dsypnea and @problem$ x 4-5 days

Output:
No Relations

Fig. A.5. Sample outputs of the Llama2 model and Llama2-MedTuned on Relation Extraction.

General Medical Instructions

Instruction:
In your capacity as a medical expert, follow the user's instructions and help them with their medical inquiries.

Input:
Describe the clinical features, diagnostic criteria, and management strategies for gastroesophageal reflux disease (GERD). Emphasize lifestyle modifications and acid-suppressing medications, listing them in numbered order.

Output:

GERD is a chronic digestive disease that occurs when stomach acid or, occasionally, stomach content, flows back into a persons' food pipe (esophagus). GERD may cause symptoms such as heartburn, chest pain, chest tightness, chest pressure, bloating, burping, and sour water feeling in the back of the throat. GERD can lead to complications such as esophageal inflammation (esophagitis), esophageal ulcers, and Barrett's esophagus, which can increase the risk of adenocarcinoma of the esophagus.

Diagnostic criteria for GERD include:

1. Symptoms of heartburn, chest pain, chest tightness, chest pressure, bloating, burping, or sour water feeling in the back of the throat at least two times a week for the past six weeks.
2. Duration of symptoms of two or more times a week for at least six weeks.
3. Relief of heartburn, chest pain, chest tightness, chest pressure, bloating, burping, or sour water feeling after taking an antacid or acid-suppressing medication.
4. Presence of acid in the esophagus, as demonstrated by esophageal pH testing.
5. Presence of esophageal inflammation, as demonstrated by esophageal biopsy.

Treatment options for GERD include:

1. Lifestyle modifications, such as:

a. Eating smaller, more-frequent meals.
b. Avoiding spicy foods, caffeine, nicotine, and alcohol.
c. Waiting three to four hours after a meal before lying down.
d. Raising the head of the bed about eight inches with blocks.
e. Avoiding sleeping or lying down within two hours of a meal.
f. Reducing weight if overweight.
g. Quitting smoking and avoiding exposure to secondhand smoke.
h. Avoiding stress.

2. Acid-suppressing medications, such as:

a. Antacids, such as omeprazole (Prilosec), lansoprazole (Prevail), pantoprazole (Protonix), and rabeperazole (Risperidone).
b. H2-receptor antagonists, such as cimetidine (Tagamet), ranitidine (Ranitidine), famotidine (Pepcid), and nizatidine (Axid).
c. Proton pump inhibitors (PPIs), such as omeprazole (Prilosec), lansoprazole (Prevail), pantoprazole (Protonix), and rabeperazole (Risperidone).

Fig. A.6. Llama2-MedTuned output sample on general medical instructions.

Chat Doctor

Instruction:
Your identity is a doctor, kindly provide answers to the medical questions with consideration of the patient's description. Determine the relation between the two medical terms.

Input:
Hi doctor, My last USG report showed intrauterine pregnancy with 8 mm gestation sac, no fetal and yolk sac. Last week I had brown discharge with mild lower abdominal pain and was prescribed Susten 200 mg (oral). Now, I am feeling the pain frequently in the right lower abdomen with mild bleeding and pain. It is less when I lie on bed. What to do?

Output:
Hello dear, I understand your concern. In my opinion the bleeding along with abdominal pain might be due to miscarriage or abnormal pregnancy. I suggest you to get a repeat ultrasound done to know the pregnancy status. If the gestational sac is present without fetal and yolk sac it suggests miscarriage or blighted ovum. If the sac is empty it also suggests miscarriage. So I suggest you to get a repeat ultrasound done to know the pregnancy status. Avoid physical stress and intercourse. Take rest, Don't worry.

Best regards...

UMLS

Instruction:
In your capacity as a doctor, it is expected that you answer the medical questions relying on the patient's description. Explain the definition of the medical term.

Input:
What is the meaning of milligram/day?

Output:
A unit of mass commonly used to express the dose of a substance to be administered over a period of time. One milligram/day is equal to one thousandth of a gram per day.

UMLS Relation

Instruction:
Your identity is a doctor, kindly provide answers to the medical questions with consideration of the patient's description. Determine the relation between the two medical terms.

Input:

###Question:
What is the relation between L-DOPS and Product containing precisely droxidopa 100 milligram/1 each conventional release oral capsule (clinical drug) ?

Output:

###Answer: Product containing precisely droxidopa 100 milligram/1 each conventional release oral capsule (clinical drug) has active ingredient: L-DOPS

Fig. A.7. Llama2-MedTuned outputs on a few medical generation tasks.

References

[1] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst 2017;30.

[2] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019, p. 4171–86. http://dx.doi.org/10.18653/v1/N19-1423, URL https://aclanthology.org/N19-1423.

[3] Radford A, Narasimhan K, Salimans T, Sutskever I, et al. Improving language understanding by generative pre-training. 2018.

[4] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877–901.

[5] Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J-N, Laleh NG, Löffler CML, Schwarzkopf S-C, Unger M, Veldhuizen GP, et al. The future landscape of large language models in medicine. Commun Med 2023;3(1):141.

[6] Kormilitzin A, Vaci N, Liu Q, Nevado-Holgado A. Med7: A transferable clinical natural language processing model for electronic health records. Artif Intell Med 2021;118:102086.

[7] Lehman E, Hernandez E, Mahajan D, Wulff J, Smith MJ, Ziegler Z, Nadler D, Szolovits P, Johnson A, Alsentzer E. Do we still need clinical language models? In: Mortazavi BJ, Sarker T, Beam A, Ho JC, editors. Proceedings of the conference on health, inference, and learning. Proceedings of machine learning research, vol. 209, PMLR; 2023, p. 578–97, URL https://proceedings.mlr.press/v209/eric23a.html.

[8] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI Blog 2019;1(8):9.

[9] Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, Du N, Dai AM, Le QV. Finetuned language models are zero-shot learners. 2021, arXiv preprint arXiv:2109.01652.

[10] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst 2022;35:27730–44.

[11] Penedo G, Malartic Q, Hesslow D, Cojocaru R, Cappelli A, Alobeidli H, Pannier B, Almazrouei E, Launay J. The RefinedWeb dataset for falcon LLM: outperforming curated corpora with web data, and web data only. 2023, arXiv preprint arXiv:2306.01116. URL https://arxiv.org/abs/2306.01116.

[12] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, et al. Llama: Open and efficient foundation language models. 2023, arXiv preprint arXiv:2302.13971.

[13] 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).

[14] Han T, Adams LC, Papaioannou J-M, Grundmann P, Oberhauser T, Löser A, Truhn D, Bressem KK. MedAlpaca–an open-source collection of medical conversational AI models and training data. 2023, arXiv preprint arXiv:2304.08247.

[15] Toma A, Lawler PR, Ba J, Krishnan RG, Rubin BB, Wang B. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. 2023, arXiv preprint arXiv:2305.12031.

[16] Wu C, Lin W, Zhang X, Zhang Y, Wang Y, Xie W. PMC-llama: Towards building open-source language models for medicine. 2023, arXiv:2304.14454.

[17] Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inform 2014;47:1–10.

[18] Li J, Sun Y, Johnson RJ, Sciaky D, Wei C-H, Leaman R, Davis AP, Mattingly CJ, Wiegers TC, Lu Z. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database 2016;2016.

[19] Smith L, Tanabe LK, Kuo C-J, Chung I, Hsu C-N, Lin Y-S, Klinger R, Friedrich CM, Ganchev K, Torii M, et al. Overview of BioCreative II gene mention recognition. Genome Biol 2008;9(2):1–19.

[20] Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. J Am Med Inform Assoc 2013;20(5):806–13.

[21] Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552–6.

[22] Bravo À, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinform 2015;16:1–17.

[23] Rohanian O, Nouriborji M, Kouchaki S, Clifton DA. On the effectiveness of compact biomedical transformers. Bioinformatics 2023;39(3):btad103. http://dx.doi.org/10.1093/bioinformatics/btad103, arXiv:https://academic.oup.com/bioinformatics/article-pdf/39/3/btad103/49571998/btad103.pdf.

[24] Rohanian O, Nouriborji M, Jauncey H, Kouchaki S, Group ICC, Clifton L, Merson L, Clifton DA. Lightweight transformers for clinical natural language processing. 2023, arXiv:2302.04725.

[25] Romanov A, Shivade C. Lessons from natural language inference in the clinical domain. 2018, arXiv:1808.06752 [cs]. URL http://arxiv.org/abs/1808.06752.

[26] Baker S, Silins I, Guo Y, Ali I, Högberg J, Stenius U, Korhonen A. Automatic semantic classification of scientific literature according to the hallmarks of cancer. Bioinformatics 2015;32(3):432–40.

[27] Jin D, Pan E, Oufattole N, Weng W-H, Fang H, Szolovits P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Appl Sci 2021;11(14):6421.

[28] Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. PubMedQA: A dataset for biomedical research question answering. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). 2019, p. 2567–77.

[29] Jahan I, Laskar MTR, Peng C, Huang J. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. 2023, arXiv preprint arXiv:2310.04270.