LlamaCare: an Instruction Fine-Tuned Large Language Model for Clinical NLP

Rumeng Li[1,2], **Xun Wang**[3], **Hong Yu**[1,2,4]
[1]Umass Amherst, [2]VA Bedford Healthcare System,[3]Microsoft, [4]Umass Lowell
[1]Amherst, MA, USA; [2]Bedford, MA, USA; [3]Redmond, WA, USA;[4]Lowell, MA, USA
rli@umass.edu, wangxun.pku@gmail.com, hong_yu@uml.edu

Abstract

Large language models (LLMs) have shown remarkable abilities in generating natural texts for various tasks across different domains. However, applying LLMs to clinical settings still poses significant challenges, as it requires specialized knowledge, vocabulary, as well as reliability. In this work, we propose a novel method of instruction fine-tuning for adapting LLMs to the clinical domain, which leverages the instruction-following capabilities of LLMs and the availability of diverse real-world data sources. We generate instructions, inputs, and outputs covering a wide spectrum of clinical services, from primary cares to nursing, radiology, physician, and social work, and use them to fine-tune LLMs. We evaluated the fine-tuned LLM, LlamaCare, on various clinical tasks, such as generating discharge summaries, predicting mortality and length of stay, and more. Using both automatic and human metrics, we demonstrated that LlamaCare surpasses other LLM baselines in predicting clinical outcomes and producing more accurate and coherent clinical texts. We also discuss the challenges and limitations of LLMs that need to be addressed before they can be widely adopted in clinical settings.

Keywords: LLM, clinical NLP, Llama 2, instruction tuning

1. Introduction

Natural language understanding and generation are essential for many applications in the clinical domain, such as summarizing patient records, answering queries, or predicting patient outcomes. However, these applications require processing and producing medical texts that are rich in domain-specific knowledge, terminology, and logic. Large language models (LLMs), as "foundation models" (Bommasani et al., 2021) that can learn general representations from large-scale corpora and adapt to various domains and tasks, have shown great potential in such natural language tasks. They have demonstrated the potential to transform modern medicine by offering new tools and insights for healthcare, such as Chat-GPT and GPT-4, which promise to revolutionize clinical decision support, clinical trial recruitment, clinical data management, research support, patient education, etc. (Xue et al., 2023).

However, applying LLMs to the clinical domain faces several challenges, such as data scarcity, ethical issues, and lacking of domain-specific knowledge. Several LLMs have been compared on clinical language understanding tasks and it is showen that LLMs need domain adaptation, task-specific learning, and clinical pretraining to handle the complexity and incompleteness of medical texts. Moreover, LLMs need to be carefully evaluated and monitored to ensure their reliability and safety in healthcare settings (Wang et al., 2023b; Lehman et al., 2023; Nori et al., 2023).

Recent work has explored parameter-efficient and domain-aware methods for adapting LLMs to specific domains. For example, a recent work has proposed a two-step parameter-efficient fine-tuning (PEFT) (Hu et al., 2021) framework that uses adapter layers to fine-tune LLMs on both clinical data and downstream tasks (Gema et al., 2023). ClinicalGPT (Wang et al., 2023a), a LLM that is explicitly designed and optimized for clinical scenarios by incorporating diverse real-world data, such as medical records, domain-specific knowledge, and multi-round dialogue consultations in the pretraining process, has also been developed. These works demonstrate the effectiveness and potential of domain adaptation methods for LLMs in the clinical domain.

Building on the insights from the referenced literature, a primary challenge in tailoring large language models (LLMs) to clinical applications lies in steering these models to grasp the intricacies of clinical tasks and enhance their memory and reasoning with domain-specific knowledge. In this work, we propose a novel and efficient method of instruction fine-tuning for this purpose. The proposed method leverages the LLMs' ability of generating diverse instructions and the availability of diverse real-world data sources to mimic the complexity and diversity of clinical tasks. We extract inputs, and outputs for various clinical services, from primary cares to nursing, radiology, physician, and social work from real-world clinical data, and use LLMs to generate instructions of different styles and lengths to vary the inputs, which increases the diversity of the fine tuning dataset. The created dataset thus can be more representative and inclusive.

Our method is inspired by the recent work on instruction-tuning and self-instruction. Instruction-tuning is a method to improve the instruction-following capabilities of pre-trained language models by using generated instructions, input, and output samples. It has advantages over parameter tuning of the entire model or parameter efficient tuning, such as better generalization, higher efficiency and scalability, and more natural and intuitive interaction (Wang et al., 2022b). Self-Instruct is a nearly annotation-free method that bootstraps off the generations of a language model and use them to fine-tune the original model. Self-Instruct outperforms existing public instruction datasets by a large margin on novel tasks (Wang et al., 2022a). We employ these methods to further enhance data diversity and apply them to the clinical domain. Specifically, we use the GPT-4 (OpenAI, 2023) to generate instructions and extract inputs and outputs from the MIMIC-III dataset (Johnson et al., 2016), a large-scale collection of de-identified clinical notes. We fine-tune Llama 2 (Touvron et al., 2023) on the instruction fine-tuned dataset we created and evaluate its performance on two types of tasks: text generation and text classification. By comparing the fine-tuned model with the original Llama 2 and other baselines, we demonstrate the effectiveness of the self-instruction fine tuning pipelines for enhancing the model quality in the clinical domain.

The main topics covered in this paper that may be of interests to the Clinical NLP research community are as follows:

- We propose and implement a method for instruction fine tuning of Llama 2 for the clinical domain, using the MIMIC-III dataset as a source of training data and GPT-4 for diverse instruction generation.
- We show that the fine-tuned model surpasses the original model and other baselines on both text generation tasks and text classification tasks.
- We analyze the impact of instruction fine tuning on the model output quality and diversity, and provide insights into the strengths and weaknesses of the method.

2. Background and Related Work

Large language models (LLMs) have emerged as a powerful paradigm for natural language processing and understanding, achieving remarkable results on a variety of tasks and domains. However, applying LLMs to the biomedical and clinical domains poses several challenges, such as the need for domain-specific knowledge, the scarcity of annotated data, and the ethical and legal implications of handling sensitive health information.

One of the main approaches for adapting LLMs to BioNLP tasks is to pretrain them on large collections of biomedical and clinical texts, such as scientific articles, electronic health records (EHRs). Some popular models include BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), ClinicalBERT (Alsentzer et al., 2019), and BioMedRoBERTa (Gururangan et al., 2020).

Another approach for leveraging LLMs for BioNLP tasks is to fine-tune them with natural language instructions that describe the desired behavior or output for a given task. This way, the LLMs can generalize to new or unseen tasks without requiring task-specific architectures or labeled data. Several instruction-tuned models have been proposed for BioNLP (Kamble and AlShikh; Karn et al., 2023).

The self-instruct method, in particular, provides an effort-free way for generating instructions (Wang et al., 2022a). However, the LLM generated instructions may be homogeneous and cannot reflect the complexity and diversity of real-world scenarios.

3. Methodology

In this section, we describe the instruction fine-tuning method in details, including 1) the generation of instruction fine-tuning dataset, i.e., how to generate diverse instructions using GPT-4, and how to extract and format the inputs and outputs from the MIMIC-III dataset; 2) how to fine-tune the Llama 2 model. We also explain the design choices and the hyperparameters of the method, and justify them with empirical or theoretical evidence.

3.1. Automatic Instruction Data Generation

We created two types of instruction dataset aiming to guide the model perform well in different clinical tasks including both text classification tasks and text generation task. Each example in the instruction data contains three fields, namely, the instruction, the input and the output.

3.1.1. Instruction Generation

We firstly generated a set of instructions for each type of service that appears in the MIMIC-III notes. A service type is a category of clinical activity, such as Radiology, Respiratory, Rehab, etc. We provided a seed instruction for each type of service, which is a short sentence that describes the main goal or action of the service type, such as "Write a note summarizing the patient's status after the [service name]". We then used GPT-4 to rephrase the seed instruction in different ways, varying the wording, style, and level of detail. Up to 20 instructions for each service type are generated, then we manually checked and filtered out any instructions that are irrelevant, ambiguous, or contradictory.

Similarly, we manually created an instruction for generating ICD codes from discharge summaries: "Assign ICD codes based on the discharge summary", and asked GPT-4 to produce more variations of this instruction, and up to 20 instructions for the ICD coding task are collected.

3.1.2. Input/Output Extraction

We used MIMIC-III as the source of data for input and output extraction. For each patient, we extracted their demographic information, such as age and sex, from the structured data. We also extracted the chief complaint section (if available) and the preliminary diagnosis section from the notes and combined them as the input.

The output depends on the type of service and the task. For the note generation task, The output consists of the remaining parts of the EHR note that match the service type and the instruction, excluding the sections that were used as input. For ICD coding, the input is the discharge summary, which is a free-text document that summarizes the patient's hospital course and discharge plan. The output is a list of ICD codes that represent the diagnoses and procedures that occurred during the hospitalization.

The building of the instruction dataset requires minimal human annotation by leveraging the diversified and creative outputs of GPT-4 and the demographic and notes information extracted from MIMIC-III, except the seed instructions which need to be provided by experts. The generated dataset is designed to guide the model to perform two types of clinical tasks on various clinical scenarios, including primary care, ECG, consult, etc.

Figure 1 shows an illustrative example on how we generate the instruction data.

3.2. Llama 2 Instruction Fine-tuning for Clinical Domain

After creating the instruction data, we used it to fine-tune the original Llama 2 model. The Llama 2 include a series of models of different sizes and trained with/without chat data. We used the Llama 2 with 7 billion parameters chat version as the base model. We concatenated the instruction and instance input as a prompt and trained the model to generate the instance output in a standard supervised way.

To make the model robust to different formats, we used multiple templates to encode the instruction and instance input together. For example, the instruction can be prefixed with "Task:" or not, the input can be prefixed with "Input:" or not, "Output:" can be appended at the end of the prompt or not, and different numbers of break lines can be put in the middle, etc.

We used LoRA (Hu et al., 2021) for the model fine-tuning which greatly reduced the number of parameters that need to be tuned. LoRA is much more efficient and economical than full fine-tuning. During inference, LoRA weights are added to the base model to create the fine-tuned model. The details of LoRA fine tuning settings are shown in Table 2.

4. Experiments

4.1. Dataset

As mentioned above, we used the MIMIC-III as the source of the inputs and outputs for the fine-tuning and the evaluation of the Llama 2 model. Amongst the tables present in this database, we used the 'noteevents' table, which contains various notes for patients. There are 15 types of notes present, including Discharge summary, Echo, ECG, Nursing etc. For text generation fine-tuning, we used all types of the notes available. For ICD coding fine-tuning, we used the Discharge Summary notes only.

The statistics of our instruction dataset are presented in Table 1. We randomly split the dataset into three subsets: train, validation, and test, with the ratio of 90%, 5%, and 5%, respectively. We ensured that the subsets have no overlapping patients or notes, and that they have similar distributions of note types, note contents, and note metadata. We used the train subset to fine-tune the Llama 2 model, the validation subset to tune the hyperparameters and select the best model, and the test subset to evaluate the performance and compare the models.

4.2. Training Configurations

The Parameter-Efficient Fine-Tuning (PEFT) package is used for Llama 2 fine tuning [1]. We used 8x NVIDIA A100-80GB GPUs for model fine tuning. It takes 68 hrs to finish the training. We present the parameter settings in Table 2.

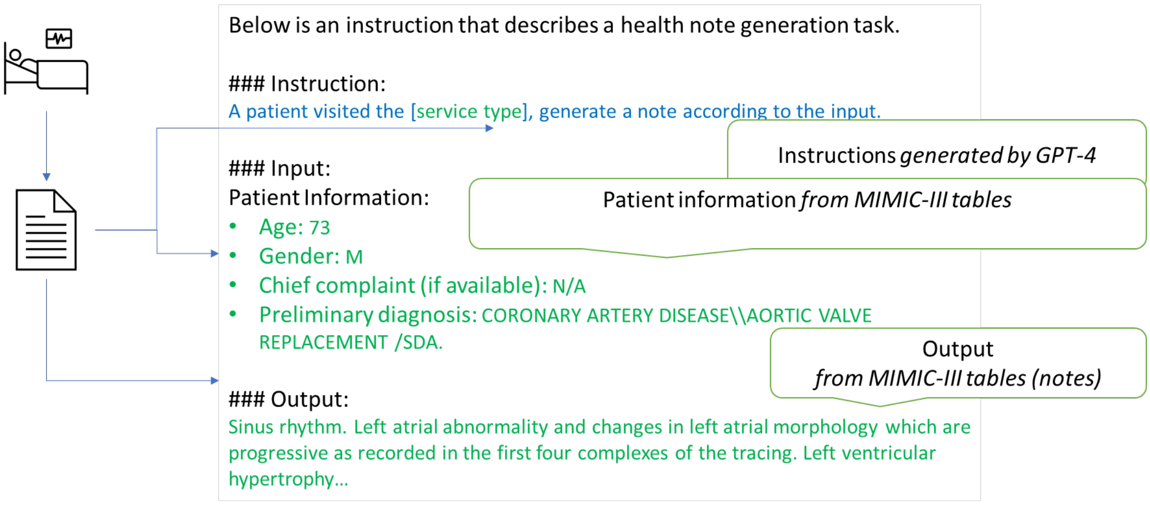
4.3. Tasks

We conducted two kinds of tasks including text generation and text classification to evaluate the performances of the fine tuned model.

4.3.1. Text Generation

- Discharge summary generation (DSG): Automatically generate discharge summaries from

[1] https://huggingface.co/docs/peft/index



Below is an instruction that describes a health note generation task.

Instruction:
A patient visited the [service type], generate a note according to the input.

Input:
Patient Information:
- Age: 73
- Gender: M
- Chief complaint (if available): N/A
- Preliminary diagnosis: CORONARY ARTERY DISEASE\\AORTIC VALVE REPLACEMENT /SDA.

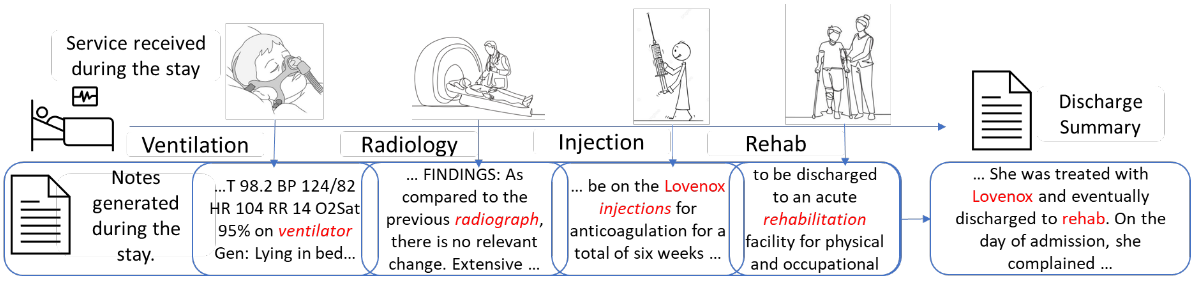
Output:
Sinus rhythm. Left atrial abnormality and changes in left atrial morphology which are progressive as recorded in the first four complexes of the tracing. Left ventricular hypertrophy...

Instructions *generated by GPT-4*
Patient information *from MIMIC-III tables*
Output
from MIMIC-III tables (notes)

Figure 1: Overview of the instruction data generation, using note generation as an example. For ICD coding, the inputs are discharge summaries and outputs are ICD-9 codes (diagnosis and procedure).

Service Type	# of Instances	Ave. Input Length	Ave. Output Length	Total # of Tokens
Discharge summary	59652	68.3	1804.92	1.11e+08
Echo	45794	41.98	311.78	1.56e+07
ECG	209051	24.24	24.13	7.54e+06
Nursing	223556	14.72	319.43	7.20e+07
Physician	141624	16.14	1117.02	1.59e+08
Rehab Services	5431	37.19	531.69	3.02e+06
Case Management	967	50.6	189.65	2.20e+05
Respiratory	31739	19.85	191.37	6.31e+06
Nutrition	9418	32.34	406.99	4.02e+06
General	8301	35.08	263.17	2.37e+06
Social Work	2670	43.13	365.81	1.06e+06
Pharmacy	103	52.24	412.28	4.66e+04
Consult	98	42.44	978.97	9.89e+04
Radiology	522279	17.34	172.17	9.26e+07
Nursing/other	822497	14.79	163.65	1.37e+08

Table 1: Statistics of the automatically generated instruction dataset. The dataset contains 15 types of notes with about 6e+08 tokens.



Service received during the stay
Ventilation
Radiology
Injection
Rehab
Discharge Summary
Notes generated during the stay.

...T 98.2 BP 124/82 HR 104 RR 14 O2Sat 95% on *ventilator* Gen: Lying in bed...

... FINDINGS: As compared to the previous *radiograph*, there is no relevant change. Extensive ...

... be on the Lovenox *injections* for anticoagulation for a total of six weeks ...

to be discharged to an acute *rehabilitation* facility for physical and occupational

... She was treated with Lovenox and eventually discharged to rehab. On the day of admission, she complained ...

Figure 2: Illustration of the discharge summary note generation task, which aims to encapsulate the key details of a patient's hospitalization. It synthesizes information from prior EHR notes and other sources, along with expert observations, assessments, and plans. In this automated task, the model processes all pertinent notes from a patient's admission to produce the discharge summary.

LoRA parameters	
LoRA_alpha	16
LoRA_dropout	0.1
LoRA_rank	64
bias	None
Other parameters	
num_train_epochs	1
per_device_train_batch_size	4
gradient_accumulation_steps	2
optimizer	paged_adamw
learning_rate	2e-4
tf32	True
max_grad_norm	0.3
warmup_ratio	0.01
max_length	4k
lr_scheduler_type	constant

Table 2: Parameters for Model Fine-tuning on 8x A100 80GB GPUs

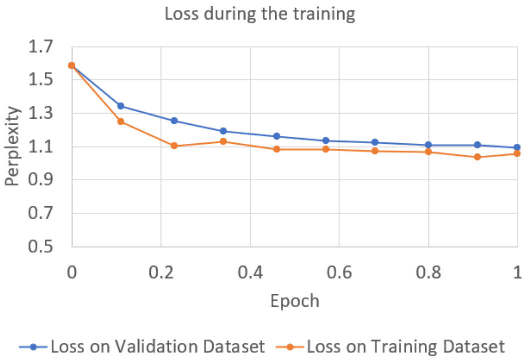


Figure 3: The perplexity curve reported on the training and validation dataset. The loss keeps decreasing on both the train and validation dataset during the training.

prior notes for a clinical encounter (Shing et al., 2021).

4.3.2. Text Classification

- In-hospital mortality (MOR): a binary classification task of predicting the survival outcome of a patient during their hospital stay (Van Aken et al., 2021).

- Length of stay (LOS): a task of assigning a time-bin label to the duration of a patient's hospital stay using multiclass classification. The time-bin labels for the LOS task are: less than three days, three to seven days, one to two weeks, and more than two weeks (Van Aken et al., 2021).

- Diagnoses (DIAG): a multilable classification task of predicting the possible diagnoses related to a patient's condition. The diagnoses for the DIAG task are represented by simplified ICD-9 codes (Van Aken et al., 2021).

- Procedures (PROC): a multilable classification task of predicting the possible diagnostics or treatments that a patient received. The procedures for the PROC task are represented by simplified ICD-9 codes (Van Aken et al., 2021).

4.4. Baselines

We compared our fine-tuned LlamaCare model with the original Llama 2 model (Touvron et al., 2023) and other baselines.

- LlamaCare: Our proposed clinical domain instruction fine-tuned model based on Llama 2-chat, the 7 billion parameters chat version.

- Llama 2-chat, the 7 billion parameters chat version, the fine-tuned model which leverages publicly available instruction datasets and over 1 million human annotations.

- Llama, the 7 billion parameters version.

- PMC-LLaMA (Wu et al., 2023), the 7 billion parameters version of a domain-adapted Llama model that was pretrained on 4.8 million biomedical academic papers from PubMed Central.

To ensure a fair comparison, all our baselines are Llama-based (version 1 or version 2) models with 7 billion parameters, and we do not compare them with larger LLMs such as the GPT family. Although some other BERT-based models have been fine-tuned for the clinical domain, they are not the main focus of our work (Alsentzer et al., 2019; Lee et al., 2020; Peng et al., 2020; Van Aken et al., 2021; Michalopoulos et al., 2020). Moreover, There are already existing studies that compare BERT-like models and Llama models (Gema et al., 2023) which show that fine tuned Llama outperforms fine tuned bert-like models.

The classification task-specific data was built following the description in (Van Aken et al., 2021). We fine tuned the models on the downstream task data using LoRA.

For the text generation task, we collected the related notes of an admission as inputs and used the models to generate the discharge summary. The discharge summary generation procedure is depicted in Figure 2. In this work, we only kept admissions whose total word counts of all related notes do not exceed the max length and fine tuned the model using LoRA. Note that tables like medications/lab tests/outputs/etc. are not used. After data cleaning and selection, we collected 500 for testing purpose and used 5000 for training. This setting is adopted for simplicity and poses impacts on the experiment results which will be explained in the Results and Analysis section.

4.5. Evaluation

The evaluation for text classification tasks are reported using the Area Under the Receiver Operating Characteristic Curve (AUROC) scores. Additionally, we report the macro-averaged AUROC score across all clinical tasks as commonly done in NLP benchmarking tasks (Peng et al., 2019; Gema et al., 2023)

To evaluate the quality of the generated text, we use two automatic metrics: BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). These metrics measure the similarity between the generated text and the reference text based on the n-gram overlap. We also show some examples of the generated notes and compare them with the baselines and the reference notes.

In addition, we performed human evaluation on 100 randomly selected discharge summary notes generated by our model. We asked two experts who have extensive experience in annotating clinical data to rate the usefulness of the generated text on a scale of 1 to 5. Usefulness is a criterion that measures how well the generated text can serve as a replacement for human written notes (Table 3). We reported the average ratings and the inter-annotator agreement. We also calculated the correlations between the human ratings and automatic metrics.

5. Results and Analysis

5.1. Training Perplexity

Table 4 provides comparisons of our proposed LlamaCare model and the baseline models in terms of perplexity score. Figure 3 shows the perplexity loss in train/validation dataset. The loss decreases as the model sees more data. LlamaCare, a clinical domain-specific model, outperforms the baselines, Llama 2 and PMC-LLaMA, in perplexity, demonstrating its effectiveness. Note that the perplexity score is a direct metric of model optimization and act as a proxy for model quality here. We will further evaluate the effectiveness of LlamaCare on downstream tasks.

5.2. Performance Comparison on Text Classification Tasks

Table 7 shows the results on 4 tasks and the macro average AUROC. The LlamaCare is able to beat all baselines, boosting the performances by 2 to 5 points on AUROC, depending on the task. This verifies the advantages of LlamaCare on these clinical tasks and is consistent with its low perplexity on the clinical data.

5.3. Performance Analysis on text Generation Tasks

For the text generation task, Table 5 shows that the LlamaCare model outperforms baseline models on all the metrics, followed by the fine-tuned PMC_LLaMA model and the original Llama 2 model. The original Llama model has the lowest scores on these automatic evaluation metrics. The human evaluation on a small dataset agrees with this ranking and shows strong correlations with ROUGE (Pearson correlation $r = 0.72$) and BLEU (Pearson correlation $r = 0.65$), validating the reliability of these automatic metrics. We also present some examples of the generated texts in Table 6. As shown in the table, the outputs of Llama and its variant PMC_LLaMA are repetitive and do not capture the essential service that the patient received. Llama 2 model, on the other hand, can generate fluent sentences and cover some aspects of the admission. The LlamaCare model, builds upon the foundation of Llama 2, significantly improves output quality by delivering text that is more detailed, consistent, and precise.

The human annotators also raised some questions that need to be addressed as follows:

- How to handle information that is not derived from the related notes, but from previous admissions or non-text documents. For example, the discharge summary may include the patient's disease history, medicine records, tables, forms, images, and so on, that are not present in the notes during the patient's hospital stay.
- How to handle information that is not factual, but tentative or provisional. For example, the discharge notes may include possible plans or diagnoses that depend on expert knowledge and personal insights. LLMs and humans may have different opinions or preferences for these kinds of information.

These challenges have been discussed by previous works (Ando et al., 2022) and further confirmed by the experiments in this study. These challenges raise questions about the accuracy, completeness, and consistency of LLM-generated texts in clinical settings. And they are essential for ensuring the quality and safety of patient care and communication. These challenges must be addressed before LLMs can be widely adopted in clinical settings.

The LlamaCare model shows better performance on both text generation and text classification tasks, which indicates that instruction fine-tuning improves

Rating	Criteria
5	The generated text can replace the human-written references without any changes.
4	The generated text can mostly replace the human-written references with some minor modifications.
3	The generated text is somewhat useful but requires significant modifications.
2	The generated text is mostly not useful except for a few sentences.
1	The generated text is not useful at all.

Table 3: Human Evaluation Criteria of Usefulness

Model	Train Perplx.	Test Perplx.
Llama-LoRA	1.858	2.244
PMC_LlaM-LoRA	1.938	2.404
LlamaCare	1.057	1.09

Table 4: Domain-adaptive Pretraining results of Llama and PMC-LLaMA on MIMIC-IV clinical notes (Gema et al., 2023), and LlamaCare trained on MIMIC-III clinical notes with a language modelling objective. Lower perplexity scores indicate better language modelling performance.

Model	Rouge-L	BLEU-4	Human Rate
Llama 2-chat	25.4	12.3	2.6
Llama	20.3	10.0	1.9
PMC_LLaMA	22.4	13.2	2.5
LlamaCare	27.2	18.8	3.2

Table 5: Text generation results on the extracted test subset of the MIMIC-III dataset. Note that the human rates are based on 100 test examples.

the quality of the Llama 2 model on the clinical domain. We hypothesize that instruction fine-tuning helps the Llama 2 model to learn the clinical vocabulary, syntax, style, and logic, and to generate texts that are more relevant and accurate for the clinical instructions. We also hypothesize that instruction fine-tuning boosts the learning ability of the Llama 2 model, and enables it to perform or adapt to different types of instructions or scenarios.

6. Limitations

Our work also has some limitations and lead to some open questions, including but not limited to:

- The effectiveness of instructions and outputs hinges on their quality and variety, which can be compromised by data biases, errors, or inconsistencies. Ensuring their integrity and managing issues like data noise, scarcity, or imbalance is crucial.

- In the note generation task, we excluded examples that exceed the maximum length limits of LLMs. This typically pertains to patients with extensive notes, indicative of complex health issues and numerous services received. Omitting these cases introduces a data selection bias, resulting in inflated performance metrics that warrant further investigation in subsequent research.

- It is subject to the ethical and the social issues, and may generate texts that are inappropriate, misleading, or harmful for the domain or the task, or may affect the decisions, the actions, or the outcomes of the users or the patients.

- Privacy remains a significant concern for customers of LLMs. To mitigate this, the utilization of GPT-* is confined solely to generating instructions, while the rest of the processes are managed by open-source models, reducing the likelihood of data breaches.

7. Conclusion

In this paper, we proposed and implemented a method for instruction fine tuning of Llama 2 for the clinical domain, using GPT-4 and the MIMIC-III dataset as a source of instructions, inputs and outputs. We evaluated the fine-tuned model on both text generation and text classification tasks, and showed that it outperformed the original model and other baselines on both tasks. Human annotators evaluated the model output quality for discharge summary generation task, and provided insights into the strengths and weaknesses of the method. We demonstrated the potential of instruction fine tuning for enhancing language models for domain-specific and task-oriented text generation.

Our work has several implications and applications for the clinical domain, such as:

- It can potentially facilitate the development and the deployment of natural language systems for the clinical domain, such as summarizing patient records, answering queries, or generating diagnoses, by reducing the need for domain adaptation, data annotation, or model fine-tuning.

- It enhances clinician-patient communication and collaboration by producing texts that are more pertinent, precise, and comprehensible within the clinical context, and by allowing the

Model	Output
Llama 2, chat	Service: Medicine Allergies: HPI: 35 yo female with h/o HTN, hyperlipidemia, and hypothyroidism who presented to the ED with chest pain. She had been at work and was sitting at her desk when she felt a sudden onset of chest pain. She had no associated symptoms such as SOB, palpitations, nausea, or diaphoresis...
Llama	Service: Medicine Allergies: She has had intermittent chest pain for the past three weeks. She also has bilateral shoulder pain for two days. Her blood pressure is 140/80. # ALT: Blood pressure is 140/80. 140/80 is a normal blood pressure...
PMC_LLaMA	Chief Complaint: chest pain due to stress. HPI: She has been having chest pain sometimes for the last three weeks. She also has pain in both shoulders for two days. Her blood pressure is normal at 140/80...
LlamaCare	Service: MEDICINE Allergies: No Known Allergies to Drugs Chief Complaint: Chest pain Major Surgical or Invasive Procedure: None History of Present Illness: 35 yo female with h/o anxiety and depression who presents with chest pain.She was at work today and felt a sharp chest pain. in the ED, VS were 100.7, 88, 140/80, 97%. Normal EKG and chest auscultation.
Reference	Service: Medicine Allergies: HPI: By report, she was at work and sitting at her desk when she felt a sudden onset of chest pain. She had no other symptoms such as shortness of breath, palpitations, nausea, or sweating. Pt. reports to have a similar episode of chest pain 2 years ago, which was relieved with nitroglycerin. In the ED, initial VS were: 100.7, 88, 140/80, 97% RA General: Obese young women in no acute distress, alert, oriented...

Table 6: Examples generated by different models. This discharge summary is generated for a patient who had a sudden pain in chest. The outputs have been processed and formatted for readability.

Model	MOR	LOS	DIAG	PROC	Macro Avg.
Llama	73.67	65.40	75.83	86.22	75.28
PMC_LLaMA	74.39	66.67	76.01	86.97	76.01
Llama 2, chat	76.03	66.70	78.47	88.31	77.38
LlamaCare	77.62	68.76	79.16	90.76	79.08

Table 7: Text classification results (AUROC) on the test subset of the MIMIC-III dataset. For fair comparison, all models are fine tuned on the downstream task using LoRA.

users to specify or modify the goal of the text generation task.

- It can bolster the quality, dependability, and accountability of natural language systems within the clinical field. This is achieved by producing text that is consistent, coherent, and meaningful, while also complying with the clinical domain's standards and ethical principles.

For future work we will continue exploring from the following directions:

- Extending the instruction fine tuning method to other task like text summarization, question answering, table to text generation, investigating how the method varies or aligns across different domains or tasks.
- Exploring the interpretability and the robustness of the LlamaCare model, and understanding how the model responds to the variations of the instructions and the outputs.
- Integrating the LlamaCare model with other clinical systems or platforms, such as EHRs, clinical decision support systems, or telemedicine systems, and evaluating the utility, or the impact of the fine-tuned Llama 2 model on the clinical workflows, the clinical outcomes, or the patient satisfaction.

8. Acknowledgement

This study was partially supported by the National Institute on Aging of the National Institutes of Health (NIH) under award number 5R01AG080670. The authors are solely responsible for the content and do not represent the official views of the NIH. We thank the reviewers for their insightful comments and constructive suggestions. Their guidance was instrumental in refining this paper to its present form.

9. References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Kenichiro Ando, Takashi Okumura, Mamoru Komachi, Hiromasa Horiguchi, and Yuji Matsumoto. 2022. Is artificial intelligence capable of generating hospital discharge summaries from inpatient records? *PLOS Digital Health*, 1(12):e0000158.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Aryo Gema, Luke Daines, Pasquale Minervini, and Beatrice Alex. 2023. Parameter-efficient fine-tuning of llama for the clinical domain. *arXiv preprint arXiv:2307.03042*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Kiran Kamble and Waseem AlShikh. Palmyra-med: Instruction-based fine-tuning of llms enhancing medical domain performance.

Sanjeev Kumar Karn, Rikhiya Ghosh, Oladimeji Farri, et al. 2023. shs-nlp at radsum23: Domain-adaptive pre-training of instruction-tuned llms for radiology report impression generation. *arXiv preprint arXiv:2306.03264*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J. Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models?

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2020. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. *arXiv preprint arXiv:2010.10391*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020. An empirical study of multi-task learning on bert for biomedical text mining. *arXiv preprint arXiv:2005.02799*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Han-Chin Shing, Chaitanya Shivade, Nima Pourdamghani, Feng Nan, Philip Resnik, Douglas Oard, and Parminder Bhatia. 2021. Towards clinical encounter summarization: Learning to compose discharge summaries from prior notes. *arXiv preprint arXiv:2104.13498*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Betty Van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix A Gers, and Alexander Loeser. 2021. Clinical outcome prediction from admission notes using self-supervised knowledge integration. *arXiv preprint arXiv:2102.04110*.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023a. Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Yuqing Wang, Yun Zhao, and Linda Petzold. 2023b. Are large language models ready for healthcare? a comparative study on clinical language understanding. *arXiv preprint arXiv:2304.05368*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Vivian Weiwen Xue, Pinggui Lei, and William C Cho. 2023. The potential impact of chatgpt in clinical and translational medicine. *Clinical and Translational Medicine*, 13(3).