Unified Approach for More Generalizable Medical Language Understanding through Instruction Tuning

Anonymous ACL submission

Abstract

Large language models (LLMs) such as ChatGPT are fine-tuned on large and diverse instruction-following corpora, and can generalize to new tasks. However, those instruction-tuned LLMs often perform poorly in specialized medical natural language understanding (NLU) tasks that require domain knowledge, granular text comprehension, and structured data extraction. To bridge the gap, we: (1) propose a unified prompting format for 7 important NLU tasks (2) curate an instruction-tuning dataset, MNLU-Instruct, utilizing diverse existing open-source medical NLU corpora, and (3) develop BioMistral-NLU, a generalizable medical NLU model, through fine-tuning BioMistral on MNLU-Instruct. We evaluate BioMistral-NLU in a zero-shot setting, across 6 important NLU tasks, from two widely adopted medical NLU benchmarks: BLUE and BLURB. Our experiments show that our BioMistral-NLU outperforms the original BioMistral, as well as the proprietary LLMs - ChatGPT and GPT-4. Our dataset-agnostic prompting strategy and instruction tuning step over diverse NLU tasks enhance LLMs' generalizability across diverse medical NLU tasks. Our ablation experiments show that instruction-tuning on a wider variety of tasks, even when the total number of training instances remains constant, enhances downstream zero-shot generalization. [1]

1 Introduction

Fine-tuning large language models (LLMs) on a diverse collection of instruction-following datasets enables LLMs to generalize across a wide range of new tasks in a zero- or few-shot setting (Chung et al., 2022; Chowdhery et al., 2023; Touvron et al., 2023). Following this instruction fine-tuning phase, medical foundation LLMs (Zhang et al., 2024; Saab et al., 2024) have demonstrated great performance in various medical tasks, which require in-depth medical domain knowledge and logical reasoning ability (Nori et al., 2023), such as medical exams (Nori et al., 2023), common sense reasoning (Labrak et al., 2024; Han et al., 2023) and diagnostic reasoning (Saab et al., 2024). This generalizability is particularly crucial for tasks with limited annotated data, where fine-tuning is infeasible.

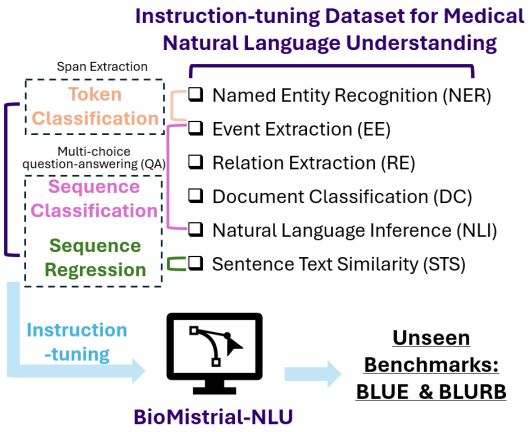

Figure 1: Instruction-tuning dataset (MNLU-Instruct), system development, and downstream evaluation for BioMistral-NLU.

Despite their superior generalizability in some areas, instruction-tuned LLMs can underperform smaller-scale, fine-tuned language models, in some specialized medical natural language understanding (NLU) tasks. These tasks require the model to understand, interpret, and respond to human language meaningfully (Wang et al., 2018). Examples of medical NLU tasks include information extraction (Xie et al., 2024; Hu et al., 2023) and sentence classification (Chen et al., 2024). The performance gap may be because the current foundation LLMs' instruction-tuning phase focuses primarily on natural language generation (NLG) tasks that allow for free-text, unconstrained outputs (Chung et al., 2022). Although many NLG tasks require complex logical reasoning, these skills do not directly translate to nuanced NLU tasks.

[1] We plan to release our code and the instruction-tuned system upon acceptance of this work.

To bridge this gap, we propose a unified prompting format for 7 important NLU tasks, employing span extraction and multi-choice question-answering (QA). Utilizing this unified format, we create an instruction-tuning dataset, MNLU-Instruct, from diverse existing open-source medical NLU corpora. We fine-tune a high-performing biomedical LLM, BioMistral (Labrak et al., 2024) on MNLU-Instruct, resulting in a new, generalizable medical NLU model we call BioMistral-NLU. We evaluate the generalizability of BioMistral-NLU, using zero-shot, dataset-agnostic prompts, on two widely adopted benchmark datasets: the Biomedical Language Understanding Evaluation (BLUE) (Peng et al., 2019) and the Biomedical Language Understanding and Reasoning Benchmark (BLURB) (Gu et al., 2021). Collectively, the benchmarks include 15 biomedical datasets with 6 important NLU task categories, across both clinical and biomedical domains.In our evaluation, BioMistral-NLU outperforms the original BioMistral, as well as ChatGPT, and GPT-4 on the macro average across all tasks. Our result demonstrated that instruction-tuning on diverse medical NLU datasets using our unified format is an effective approach to improving the generalizability on medical NLU.

2 Related work

2.1 Medical NLU

Within this broad category of medical NLU, there is extensive research on specific NLU tasks in clinical and biomedical domains, such as Information Extraction (IE) and Document Classification (DC) (Wu et al., 2020). To develop a comprehensive understanding of medical NLU, previous research curates two NLU benchmark datasets: the Biomedical Language Understanding Evaluation (BLUE) (Peng et al., 2019) and the Biomedical Language Understanding and Reasoning Benchmark (BLURB) (Gu et al., 2021). These two benchmarks encompass multiple important medical NLU tasks and are widely adopted to evaluate various LLMs for their medical NLU capabilities (Feng et al., 2024; Wang et al., 2023b; Chen et al., 2023).

Previous studies explore the ability of task-agnostic LLMs to perform medical NLU tasks. For example, Agrawal et al. (2022) demonstrate LLMs' potential for clinical NLU tasks through few-shot in-context learning (ICL). Hu et al. (2023) evaluate ChatGPT on two clinical NER datasets, representing a subset of NLU tasks. Wang et al. (2023b) propose a novel prompting strategy for multiple clinical NLU tasks using proprietary LLMs such as ChatGPT (Cha, 2022) and GPT-4 (Achiam et al., 2023). However, they only evaluate the LLMs on a few samples from each task within the BLUE benchmark. Similarly, Chen et al. (2023) and Feng et al. (2024) systematically evaluate multiple LLMs using the BLURB benchmark (Gu et al., 2021). Although ChatGPT and GPT-4 outperform other LLMs, they considerably underperform the in-domain fine-tuned systems. This performance gap highlights the need for the development of more generalized systems for medical NLU.

2.2 Instruction tuning for Medical NLU

Instruction tuning involves fine-tuning a pre-trained LM on a diverse collection of instruction-following tasks and thus enables the LM to understand and follow natural language instructions, and generalize to previously unseen tasks in zero-shot and few-shot settings (Chung et al., 2022; Ouyang et al., 2022). Instruction-tuning datasets typically encompass a wide range of natural language processing (NLP) tasks presented in an instructional format, including reasoning, question-answering, dialogue, and summarization (Zhang et al., 2023b). Utilizing instruction tuning, previous research has developed systems focused on generalizing to a limited subset of NLU tasks in the general domain, such as IE tasks (Wang et al., 2023a; Jiao et al., 2023; Sainz et al., 2023; Wang et al., 2022; Lu et al., 2022) and more specific Named Entity Recognition (NER) (Zhou et al., 2023; Zhao et al., 2024).

Several previous studies aim to adapt instruction-tuning to the medical domain, with a major focus on dialogue-based chatbots, such as ChatDoctor (Yunxiang et al., 2023) and MedAlpaca (Han et al., 2023). Other medical foundation LLMs, like MedGemini (Saab et al., 2024) and Taiyi (Luo et al., 2024), show potential for diverse NLU tasks but lack comprehensive evaluation. Previous system development has often focused on a limited subset of medical NLU tasks. For example, Luo et al. (2022b) explore Table QA; Zhao et al. (2024) focused on NER; Rohanian et al. (2023) focused on QA, IE, and text generation; However, the application of these models to other NLU tasks, such as sentence similarity and natural language inference, has not yet been explored. To the best of our knowledge, there is no comprehensive system development and evaluation across all medical NLU

tasks for their generalizability. Therefore, in this work, we aim to bridge this gap by evaluating our proposed system in a zero-shot setting using two widely adopted benchmarks, encompassing 7 important medical NLU tasks.

3 Methods

In this section, we will introduce the task formulation, and outline the three-step approach to creating our generalized LLM across medical NLU tasks.

3.1 Task formulation

We reformulate the NLU problem as text generation tasks. Our learning objective M for the medical NLU system is defined by the function $M : (I, X, T) \rightarrow O$. Specifically, given a user instruction I, associated medical text X, and NLU task labels T, the model M is instructed to output the system output O, where I, X, T, O correspond to sequences of tokens.

We reference the NLU task definitions by Gu et al. (2021) in the BLURB benchmark and group the most common NLU tasks into three categories: (1) token classification, (2) sequence classification, and (3) sequence regression.

3.2 Unified Medical NLU format

Building on prior research outlined in Section 2.1, we develop our unified NLU format that focuses on seven critical NLU tasks. This unified format simplifies evaluation across diverse NLU task outputs, and potentially facilitates knowledge transfer when the system is fine-tuned for a wider range of NLU tasks. Six of these NLU tasks are directly adapted from the BLUE and BLURB benchmarks, including named entity recognition (NER), document classification (DC), relation extraction (RE), multi-choice question-answering (QA), natural language inference (NLI), and semantic text similarity (STS). We also incorporate event extraction (EE), which is extensively researched in the medical domain (Frisoni et al., 2021). In EE, each event consists of a trigger and multiple arguments that characterize the event. The event trigger extraction (ETE) and event argument extraction (EAE) can be considered as NER. The event argument classification (EAC) classifies the event argument into a subtype, and can be considered as sequence classification. Table 1 demonstrates the example input-output format for each medical NLU task.

NER, ETE, and EAE are **token classification tasks**, which assign a class label to each token in the input sequence [2]. In token classification, the input includes the user instruction I with pre-defined token labels, and the target text T. In the output O, each line includes all the token annotations associated with a specific label. Each line starts with a class label, followed by the corresponding positive tokens in the order they appear in X. Continuous positive tokens are grouped into text spans (entities), separated by "...". If no tokens are classified as entities, the O is "None". More specifically, NER classifies each token as a possible named entity.

EAC, DC, RE, QA, and NLI are **sequence classification tasks**, which assign a class label to the entire input token sequence. In sequence classification, the user instruction I specifies pre-defined class labels as multiple choices, which is a commonly adopted format in instruction-tuning (Chung et al., 2022). The system output O is always one or more multi-choice options. In DC, the medical text X is the document. In RE, X is the corresponding medical text snippet with labeled named entities. In NLI, X is a pair of a premise and a hypothesis. In QA, user instruction I involves the task question, and X is the corresponding medical text.

STS is a **sequence regression task**, which assigns a numeric score to the entire input. In this study, we explore the widely researched task of sequence regression: calculating the semantic text similarity (STS) score between two sentences. Due to the inherent ability of LLMs to generate text, we approach this regression task as an ordinal classification task through a similar multi-QA format as sequence classification. In the user instruction I of STS, the STS scores correspond to the scoring criteria from the original publication, and are presented as multi-choice options. The STS example can be found in Table 1.

3.3 MNLU-Instruct dataset

Focusing on the 7 medical NLU tasks outlined in Table 1, we construct the instruction-tuning dataset, MNLU-Instruct, through intensively searching for publicly available clinical and biomedical NLU datasets outside of BLUE and BLURB. To better assess the generalizability of our proposed system, we intentionally avoid adding any QA datasets to the MNLU-Instruct dataset, using QA tasks as

[2] Tasks such as NER are often treated as sequence labeling tasks in the NLP field (He et al., 2020). In this work, we refer to them as Token classification tasks for consistency with the BLURB (Gu et al., 2021).

Task	Input prompt	Example output
NER/ ETE	Extract all relevant medical named entities from the medical text below. Focus on identifying following entities: {$type_1$}, {$type_2$}, ... {*text*}	Chemical: None Disease: Azotemia ... infection
EAE	What is the {$type$} attribute of the {$trigger$} '{$span$}' in the medical text below? {*text*}	Disease - Anatomy: neck...hand
EAC	What is the {$type$} attribute of the {$trigger$} '{$span$}' in the medical text below? {*text*} {*options*}	Disease - Assertion: (A) present
DC	Which options best describe cancer hallmark from the medical text below? {*text*} {*options*}	(A) Cellular energetics
RE	What is the relation between the{$type_1$} entity '{$span_1$}' and the {$type_2$} entity '{$span_2$}' from the medical text below? {*text*} {*options*}	(C) 'stress' causes 'headache'.
QA	{*question*} {*text*} {*options*}	(B) LPS is a microbial product.
NLI	What is the relation between the premise and hypothesis? Premise: {$premise$}. Hypothesis: {$hypothesis$} {*options*}	(C) Contradicts
STS	How similar are the two sentences below? Sentence 1: {$sentence_1$}. Sentence 2: {$sentence_2$}. {*options*}	(A) The two sentences are on different topics (score 0).

Table 1: The task-agnostic prompt format for 7 medical NLU tasks: named entity recognition (NER), event extraction (EE), document classification (DC), relation extraction (RE), multi-choice question-answering (QA), natural language inference (NLI), and semantic text similarity (STS). Event trigger extraction (ETE), event argument extraction (EAE), and event argument classification (EAC) are all components of the EE task. *Variables* inside {} are derived from each dataset instance.

novel tasks specifically for assessment purposes. Instead, beyond NLU tasks, we additionally incorporate three medical summarization tasks, which require similar text summarization and understanding abilities as the QA tasks. Meanwhile, Given the limited availability of public medical datasets for NLI and STS, we incorporate datasets from the general domain, including SNLI, Multi-NLI, and SIS-B. As a result, we derive the MNLU-Instruct dataset with the train splits from 33 publicly available datasets shown in Table 2.

We construct the NLU input-output pairs in MNLU-Instruct through the task-agnostic prompting strategy shown in Table 1, which directly adapts pre-defined label names from the original publications. We additionally expand abbreviated label names, i.e., from 'GENERIF' to 'Gene reference into a function (function of a gene)'. To increase the variability of MNLU-Instruct, for every NLU input-output pair, we randomly shuffle the order of task labels. Specifically, token labels in token classification tasks and multi-choice options in sequence classification and regression tasks are randomly shuffled. When train splits are unavailable or datasets have very few input-output pairs, we utilize the entire datasets for training. The complete dataset labels, prompts, and statistics can be found in Appendix A.1.

3.4 BioMistral-NLU system development

We hypothesize that instruction-tuning on a diverse, yet relevant set of tasks improves the generalizability of LLMs on medical NLU tasks. To verify this hypothesis, we fine-tune a high-performing medical LLM on MNLU-Instruct and evaluate it in a zero-shot setting.

We chose BioMistral-7B-DARE as our baseline system, which is the state-of-the-art open-source LLM on multiple medical QA tasks. For simplicity, we refer to BioMistral-7B-DARE as BioMistral in this work. We fine-tune BioMistral with full parameters on MNLU-Instruct, resulting in BioMistral-NLU-FT. However, fine-tuning LLMs in specialized domains can potentially degrade their original generalization ability across broader tasks (Ainsworth et al., 2022). To mitigate this risk and preserve the versatility of the original BioMistral, we utilize DARE (Yu et al., 2023), as suggested by Labrak et al. (2024). This approach integrates model parameters from BioMistral-NLU-FT and BioMistral, without additional training, and creates the merged system **BioMistral-NLU**.

The experiment is conducted using the alignment-handbook[3] package. Based on the engineering judgment recommended by the alignment-handbook GitHub discussion, we set the number of epochs to 3, the batch size to 16, and configured the learning rate to 2e-04 with a warmup ratio of 0.1, using 4 A100 GPUs. The rest hyperparameters are the same as the default configurations by the alignment-handbook. For inference, we use the vllm package[4] and set the temperature to 0.

[3]https://github.com/huggingface/alignment-handbook
[4]https://github.com/vllm-project/vllm

Task	Datasets used for instruction-tuning
NER	i2b2 2006DeID (Uzuner et al., 2007), i2b2 2011Coreference (Uzuner et al., 2012), i2b2 2012Temporal (Sun et al., 2013), i2b2 2014 DeID (Stubbs and Uzuner, 2015), GENIA (Yu et al., 2020), linnaeus (Kocaman and Talby, 2021), tmVar (Wei et al., 2018), DrugProt (Miranda-Escalada et al., 2023), BioRed (Luo et al., 2022a), GNorm (Morgan et al., 2008), NLM-Gene (Islamaj et al., 2021), ClinicalIE (Agrawal et al., 2022), BC4CHEMD (Kocaman and Talby, 2021), PubMed PICO (Jin and Szolovits, 2018), PICO-Data (Nguyen et al., 2017)
EE	i2b2 2009Medication (Uzuner et al., 2010), i2b2 2018ADE (Henry et al., 2020), n2c2 2022SDoH (Lybarger et al., 2023),
DC	i2b2 2006Smoking (Uzuner et al., 2008), i2b2 2008Obesity (Uzuner, 2009), n2c2 2018 (Stubbs et al., 2019), 2024 SemEval Task 2 (Jullien et al., 2024), TrialStop (Razuvayevskaya et al., 2023), MTSamples (MTS, 2023)
RE	i2b2 2011Coreference (Uzuner et al., 2012), i2b2 2012Temporal (Sun et al., 2013), EUADR (van Mulligen et al., 2012), DrugProt (Miranda-Escalada et al., 2023), BioRed (Luo et al., 2022a)
NLI	BioNLI (Bastan et al., 2022), SNLI (Bowman et al., 2015), Multi-NLI (Williams et al., 2018)
STS	SIS-B (Wang et al., 2018)
Summ	PubMedSum (Cohan et al., 2018), CDSR (Guo et al., 2021), AciDemo (Yim et al., 2023)

Table 2: The MNLU-Instruct dataset, which is used for fine-tuning: NLU and summarization datasets and tasks curated from existing open-source medical corpora.

4 Experiment setup

In this section, we will introduce our evaluation datasets, evaluation metrics, and comparative systems.

4.1 Evaluation datasets

We evaluate BioMistral-NLU in a zero-shot setting using BLURB and BLUE. Due to the sensitivity in deploying clinical-note-based corpora, we exclude the two inaccessible datasets from BLUE, ShARe/CLEF (Suominen et al., 2013) and MedSTS (Wang et al., 2020). Some datasets are included in both benchmarks evaluated, resulting in a total of 7 tasks and 15 unique datasets evaluated. We developed the evaluation datasets by utilizing the unified prompt format outlined in Table 1; the entity types and multi-choice options for those datasets are shown in Table 3 and 4. The example prompts can be found in the Appendix A.1.

Dataset	Named entities
BC2GM	Gene
BC5-chemical	Chemical
BC5-disease	Disease
NCBI-disease	Disease
JNLPBA	Protein, Cell type, RNA, Cell line, DNA
EBM PICO	Interventions, Participants, Outcomes

Table 3: NER datasets used in the evaluation.

4.2 Evaluation metrics

For consistency with prior studies, we utilize the same evaluation criteria from BLUE (Peng et al., 2019) and BLURB (Gu et al., 2021). Token classification tasks are evaluated using F1 scores, either at the token or entity level. When class labels are balanced like in NLI and QA, sequence classification tasks are evaluated using accuracy. When class labels are imbalanced, like in RE, sequence classification tasks are evaluated using F1. For the sequence regression task, STS, system outputs are converted to numerical integer scores and evaluated based on Pearson correlation.

Task	Dataset	Multi-choice options
DC	HoC	10 cancer hallmarks
QA	PubMedQA	yes / maybe / no
	BioASQ	yes / no
RE	GAD	2 gene-disease relations
	DDI	4 drug-drug interactions
	ChemProt	5 chemical-protein relations
	i2b2-2010	8 medical problem relations
NLI	MedNLI	entails / neutral / contradicts
STS	BioSSES	5 similarity score definitions

Table 4: Sequence classification and regression datasets used in the evaluation.

4.3 Comparative systems

We compare our proposed system, BioMistral-NLU, with our baseline, BioMistral, as well as other high-performing systems.

Open-source LLMs: **BioMistral** and **Llama-3-8B** (at Meta, 2024). In our controlled experiments, we evaluate open-source LLMs using our proposed unified prompting formats, shown in Table 1. The

evaluation is conducted in a zero-shot setting, except for NER datasets. Because our desired token classification output prompt format is less common during those open-source LLMs' instruction tuning phase, we additionally incorporate an explanation for the output formats and two random few-shot examples from the corresponding training set in each task. More details about the prompts and few-shot sample selection can be found in the Appendix A.2.

Proprietary LLMs: **ChatGPT** (Cha, 2022) and **GPT-4** (Achiam et al., 2023). We reference prior research that evaluates these proprietary LLMs on BLURB (Chen et al., 2023; Feng et al., 2024). Note that ChatGPT's performance is reported under one-shot ICL, while GPT-4's performance is based on randomly selected few-shot examples for NER tasks and zero-shot for other tasks. Additionally, their prompts are strategically optimized for each dataset, resulting in competitive systems.

Task- and dataset-specific fine-tuned LM: **BERT-FT**. To better understand the gap between generalized foundation LLMs and in-domain fine-tuned systems, we refer to the reported performance of BERT-based systems by the BLUE (Peng et al., 2019) and BLURB (Gu et al., 2021) benchmarks. For each dataset, a BERT-FT system is fine-tuned on its corresponding train split.

5 Results

Following the practice in BLURB (Gu et al., 2021), we average system performance across datasets for an overview. As shown in Table 5, BioMistral-NLU outperforms the baseline BioMistral with an increase in the macro average score of 19.7 for BLURB and 16.7 for BLUE. Meanwhile, BioMistral-NLU outperforms the proprietary models, achieving an increase in the macro average score of 9.0 over ChatGPT, and 2.7 over GPT-4 for BLURB.Our results demonstrate that instruction-tuning on diverse medical NLU tasks using our unified format effectively improves the LLMs' generalizability to unseen NLU datasets. In this section, we will analyze the results and characterize the gaps between the systems.

5.1 Comparison across systems

Comparing BioMistral-NLU with the baseline BioMistral, we observe an average performance increase of 33.7 for NER tasks and 8.2 for other tasks. This difference may originate from the instruct-tuning phase of BioMistral. While the NER task might be less frequent during BioMistral's instruction-tuning phase, the other tasks utilize a QA prompting strategy and are likely similar to some of BioMistral's instruction-tuning tasks. This necessitates instruction-tuning on a wider variety of NLU tasks to improve the LLM's generalizability.

Comparing BioMistral-NLU with proprietary LLMs in the BLURB benchmark, we observe that BioMistral-NLU has an average F1 score of 9.7 higher than GPT-4 across NER tasks. However, for other BLURB tasks, BioMistral-NLU has an average score of 2.0 higher than ChatGPT and 5.4 lower than GPT-4. Given that GPT-4 is significantly larger in terms of parameter size and has been instruction-tuned on much more diverse corpora, its superior generalization ability for other tasks involving more complex reasoning is consistent with the empirical scaling law (Kaplan et al., 2020; Chung et al., 2022).

Compared with the dataset-specific BERT-FT systems, we observe that BioMistral-NLU has an average performance gap of 20.3 in BLURB and 26.3 in BLUE. This disparity might be due to the ambiguity in medical NLU tasks, where disagreements are common even among human annotators following the same instructions (**?**Oortwijn et al., 2021). To tackle such ambiguity, for each dataset, the BERT-FT system requires finetuning on the corresponding train split using extensive annotated data. In contrast, BioMistral-NLU uses simplified task definitions from input prompts. It is challenging for generalized LLMs using ICL to match BERT-FT's performance.

5.2 Error analysis

We observe that for NER tasks, a major source of error for BioMistral-NLU is the nuanced task of accurately identifying exact named entity boundaries. For example, in the BC2GM gene NER dataset, the predicted named entity is 'Id - 1', whereas the gold named entity is 'mouse Id - 1'. To better understand the prevalence of this discrepancy, we evaluate the 5 NER datasets using a relaxed criterion, where two named entities are considered equivalent if their spans overlap. Using this relaxed criterion, we observe an average improvement of 15.5 in F1 across the 5 NER datasets from the original entity-level F1.

In all RE tasks, BioMistral-NLU demonstrates recall rates that are 10 to 70 points higher than its

Task	Evaluation Metric	Dataset	# test instances	In-domain	Generalized LLMs with zero- or few-shot ICL				
				BERT-FT (Peng et al., 2019) (Gu et al., 2021)	Chat-GPT (Chen et al., 2023)	GPT-4 (Feng et al., 2024)	Llama-3-8B	BioMistral	
								Baseline	Ours
NER	Entity-level F1	BC2GM†	6,322	84.5	37.5	54.6	12.6	34.1	**61.5**
		BC5-chemical†*	5,385	93.3	60.3	78.2	52.5	45.0	**89.9**
		BC5-disease†*	4,424	85.6	51.8	63.9	38.7	33.7	**67.0**
		NCBI-disease†	955	89.1	50.5	66.0	33.5	39.9	**61.8**
		JNLPBA†	8,657	79.1	41.3	45.4	33.3	25.6	**64.4**
	Token-level F1	EBM PICO†	24,474	73.4	55.6	33.5	20.2	19.6	**55.3**
DC	F1	HoC†*	315	81.5	51.2	62.5	23.1	47.3	**63.8**
QA	Acc	PubMedQA†	500	60.2	76.5	70.6	71.0	**72.0**	70.2
		BioASQ†	263	94.8	88.6	85.7	78.7	74.9	**86.7**
RE	F1	GAD†	534	84.0	52.4	51.5	55.6	55.0	**58.5**
		DDI†*	5,761	82.4	51.6	37.7	**13.2**	10.0	13.0
		ChemProt†*	14,744	77.2	34.2	37.6	35.2	28.6	**38.1**
		i2b2-2010*	6,292	76.4	-	-	38.9	30.9	**41.8**
NLI	Acc	MedNLI*	1,422	73.5	-	-	49.1	49.3	**57.5**
STS	Pearson Corr	BioSSES†*	20	92.3	42.8	89.3	67.9	69.1	**80.8**
Overall	Macro average	BLURB†	-	82.9	53.4	59.7	41.2	42.7	**62.4**
		BLUE*	-	82.8	-	-	39.8	39.2	**56.5**

Table 5: Our proposed system, BioMistral-NLU's zero-shot performance on 15 unseen medical NLU datasets from 2 benchmarks: BLURB (labeled by †) and BLUE (labeled by *). **Bold** indicates superior performance over the BioMistra-7B and Llama-3-8B, which utilize the same, dataset-agnostic prompts as BioMistral-NLU. Underline indicates better performance over the ChatGPT and GPT-4 ICL, which utilize dataset-specific prompts.

precision, suggesting a tendency to identify many false positive relationships. One major source of these false positives is the occurrence of interactions between entities, which do not fit into any of the pre-defined relation categories of interest. As a result, BioMistral-NLU assigns a wrong relation label instead of recognizing no relation.

In the sequence regression dataset, BioSSES, BioMistral-NLU tends to predict intermediate similarity scores (such as scores of 2 or 3) rather than extreme scores (0, 1, 4, or 5).

6 Discussion

We have demonstrated that instruction-tuning on diverse medical NLU tasks can enhance LLMs' downstream generalization to unseen medical NLU datasets in a zero-shot setting. In this section, we will evaluate the impact of instruction dataset composition, focusing on two components: instruction-tuning tasks and domains.

6.1 Impact of instruction-tuning tasks

We aim to assess the impact of instruction-tuning task selection from two perspectives: (1) its relevance to downstream tasks and (2) its task diversity. Focusing on these two perspectives, we fine-tune the baseline system, BioMistral, with different subsets of tasks used to build BioMistral-NLU. We evaluate the fine-tuned system on the 4 RE datasets from Table 5 in a zero-shot setting, and compare the macro-average F1 scores across the 4 RE datasets.

To study the impact of task relevance, we first construct two instruction-tuning setups: (1) with the RE task (**w/ RE**) and (2) with the DC task (**w/o RE**). We chose the DC task because DC employs a similar QA prompting format to RE and it contains 6 diverse datasets from Table 2. To study the impact of task diversity, besides DC and RE, we additionally include 2 and 4 more randomly selected tasks from Table 2. More specifically, our experiment settings are:

1. **w/ RE**:
   (a) 1 task: RE
   (b) 3 tasks: RE, NLI, NER
   (c) 5 tasks: RE, NLI, NER, EE, STS
2. **w/o RE**:
   (a) 1 task: DC
   (b) 3 tasks: DC, NLI, NER
   (c) 5 tasks: DC, NLI, NER, EE, STS

All fine-tuning experiments are controlled by using a fixed number of 50,000 data instances and running for three epochs. We maintain an equal number of instances for each task (i.e., 50,000/k instances per task when fine-tuning with k tasks),

and randomly sample fine-tuning instances from all datasets within the same task.

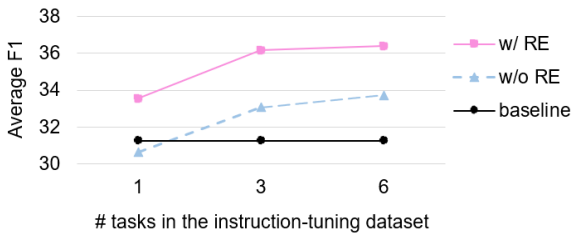

Figure 2: Average zero-shot performance on the 4 RE datasets, after instruction-tuning on 50k instances.

After BioMistral is fine-tuned with the same number of instances, we observe the following from Figure 2: (1) Overall, setting 1 (with RE) consistently outperforms setting 2 (without RE), due to its relevance to the RE datasets used in downstream evaluation; (2) In both settings, system performance increases with the number of fine-tuning tasks, demonstrating the benefits of fine-tuning with multiple tasks; (3) When fine-tuning on a single task, whether fine-tuning improves system performance on downstream tasks depends on the similarity between fine-tuning task and the downstream task.

6.2 Impact of instruction-tuning domain

After demonstrating the benefits of diverse instruction-tuning tasks, we now examine individual tasks. Note that the BLUE benchmark includes both biomedical and clinical datasets: biomedical data comes from scientific publications, while clinical data consists of semi-structured clinical notes from patients (Wu and Liu, 2011). In this section, we assess how domain selection affects downstream generalizability.

We follow a similar experimental setup as described in Section 6.1, fine-tuning BioMistral for three epochs over 25,000 data instances. The fine-tuned system is evaluated on six biomedical NER datasets from Table 5 in a zero-shot setting, using macro average F1 scores. The instruction-tuning NER datasets from MNLU-Instruct [5] are divided into biomedical and clinical splits. Our experiments include fine-tuning on a single split (**BioMed / Clinical**) and both splits (**Both**). We additionally combine single splits or include additional instances, creating a similar experiment setting with 50k instances. We use the 2-shot BioMistral described in Section 4.3 as the baseline system.

[5]We also include event triggers as named entities.

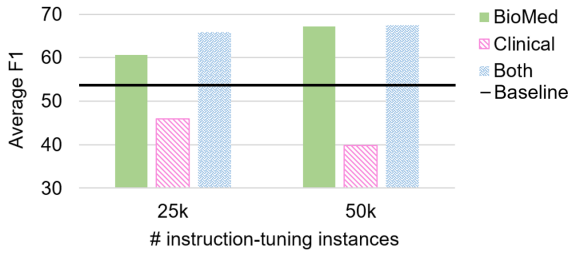

Figure 3: Average zero-shot performance on 6 biomedical NER datasets, when finetuned on different domains.

From Figure 3, we observe the following: (1) Instruction-tuning on the BioMed domain alone consistently outperforms tuning on the Clinical domain alone when using the same number of instances. (2) Compared to the baseline, instruction-tuning on the Clinical domain negatively impacts downstream performance on the BioMed domain. (3) Combining instances from both domains improves downstream generalizability to the BioMed domain, even with the same total number of instances. (4) Increasing the number of instances from the BioMed or Both domains improves performance, whereas more instances from the Clinical domain alone decrease performance.

7 Conclusion

In this work, we introduce a unified prompting format for 7 important medical NLU tasks, and develop an instruction-tuning dataset based on publicly available clinical and biomedical corpora. Our experiment demonstrates that fine-tuning across diverse medical NLU datasets improves the system's generalizability in a zero-shot setting with dataset-agnostic prompt tuning. Our ablation study underscores the necessity for instruction tuning across diverse medical NLU tasks, including domain-specific lexicon and common biomedical tasks.

Our future work will focus on further improving the generalized LLM's zero-shot performance on medical NLU tasks and narrowing its gap to in-domain fine-tuned systems. Because LLMs often struggle to adhere to in-context annotation guidelines (Zhang et al., 2023a), our future work will focus on integrating nuanced task descriptions from annotation guidelines into both the fine-tuning and inference stages (Sainz et al., 2023). Future work could also involve a self-verification step (Gero et al., 2023) or using a knowledge base as augmentation (Lewis et al., 2020) to reduce false positives in the sequence classification tasks.

Limitation

Our experiments demonstrate the effectiveness of our proposed unified and dataset-agnostic prompting strategy for medical NLU tasks. However, we acknowledge that there may be other alternative unified prompting strategies that could also be effective. We plan to evaluate the impact of different prompting formats in instruction tuning for medical NLU tasks.

In the medical field, the term “medical domain" typically encompasses both biomedical and clinical domains. Our work is primarily evaluated on biomedical datasets due to the sensitivity and inaccessibility of clinical datasets. We plan to collaborate with our home institution to gain access to real-world clinical datasets, and further evaluate and validate our proposed system in more diverse and realistic clinical settings.

References

2022. Openai: Introducing chatgpt. https://openai.com/blog/chatgpt. Accessed: 2024-04-12.

2023. Welcome to mtsamples. https://mtsamples.com/. Accessed: 2024-6-8.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. 2022. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*.

AI at Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/. Accessed: 2024-04-18.

Mohaddeseh Bastan, Mihai Surdeanu, and Niranjan Balasubramanian. 2022. Bionli: Generating a biomedical nli dataset using lexico-semantic constraints for adversarial examples. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5093–5104.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Qijie Chen, Haotong Sun, Haoyang Liu, Yinghui Jiang, Ting Ran, Xurui Jin, Xianglu Xiao, Zhimin Lin, Hongming Chen, and Zhangmin Niu. 2023. An extensive benchmark study on biomedical text generation and mining with chatgpt. *Bioinformatics*, 39(9):btad557.

Shan Chen, Yingya Li, Sheng Lu, Hoang Van, Hugo JWL Aerts, Guergana K Savova, and Danielle S Bitterman. 2024. Evaluating the chatgpt family of models for biomedical reasoning and classification. *Journal of the American Medical Informatics Association*, 31(4):940–948.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv e-prints*, pages arXiv–2210.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Hui Feng, Francesco Ronzano, Jude LaFleur, Matthew Garber, Rodrigo de Oliveira, Kathryn Rough, Katharine Roth, Jay Nanavati, Khaldoun Zine El Abidine, and Christina Mack. 2024. Evaluation of large language model performance on the biomedical language understanding and reasoning benchmark. *medRxiv*, pages 2024–05.

Giacomo Frisoni, Gianluca Moro, and Antonella Carbonaro. 2021. A survey on event extraction for natural language understanding: Riding the biomedical literature wave. *IEEE Access*, 9:160721–160757.

Zelalem Gero, Chandan Singh, Hao Cheng, Tristan Naumann, Michel Galley, Jianfeng Gao, and Hoifung Poon. 2023. Self-verification improves few-shot clinical information extraction. *arXiv preprint arXiv:2306.00024*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Zhiyong He, Zanbo Wang, Wei Wei, Shanshan Feng, Xianling Mao, and Sheng Jiang. 2020. A survey on recent advances in sequence labeling from deep learning models. *arXiv preprint arXiv:2011.06727*.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023. Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:2303.16416*.

Rezarta Islamaj, Chih-Hsuan Wei, David Cissel, Nicholas Miliaras, Olga Printseva, Oleg Rodionov, Keiko Sekiya, Janice Ward, and Zhiyong Lu. 2021. Nlm-gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. *Journal of biomedical informatics*, 118:103779.

Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji, and Jiawei Han. 2023. Instruct and extract: Instruction tuning for on-demand information extraction. *arXiv preprint arXiv:2310.16040*.

Di Jin and Peter Szolovits. 2018. Pico element detection in medical text via long short-term memory neural networks. In *Proceedings of the BioNLP 2018 workshop*, pages 67–75.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. 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)*, pages 2567–2577.

Maël Jullien, Marco Valentino, and André Freitas. 2024. Semeval-2024 task 2: Safe biomedical natural language inference for clinical trials. *arXiv preprint arXiv:2404.04963*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Veysel Kocaman and David Talby. 2021. Biomedical named entity recognition at scale. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I*, pages 635–646. Springer.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. *arXiv preprint arXiv:2203.12277*.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022a. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Ling Luo, Jinzhong Ning, Yingwen Zhao, Zhijun Wang, Zeyuan Ding, Peng Chen, Weiru Fu, Qinyu Han, Guangtao Xu, Yunzhi Qiu, et al. 2024. Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. *Journal of the American Medical Informatics Association*, page ocae037.

Man Luo, Sharad Saxena, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2022b. Biotabqa: Instruction learning for biomedical table question answering. *arXiv preprint arXiv:2207.02419*.

Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. 2023. The 2022 n2c2/uw shared task on extracting social determinants of health. *Journal of the American Medical Informatics Association*, 30(8):1367–1378.

Antonio Miranda-Escalada, Farrokh Mehryary, Jouni Luoma, Darryl Estrada-Zavala, Luis Gasco, Sampo Pyysalo, Alfonso Valencia, and Martin Krallinger. 2023. Overview of drugprot task at biocreative vii: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical–protein relations. *Database*, 2023:baad080.

Alexander A Morgan, Zhiyong Lu, Xinglong Wang, Aaron M Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jörg Hakenberg, et al. 2008. Overview of biocreative ii gene normalization. *Genome biology*, 9:1–19.

An T Nguyen, Byron C Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. 2017. Aggregating and predicting sequence labels from crowd annotations. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2017, page 299. NIH Public Access.

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*.

Yvette Oortwijn, Thijs Ossenkoppele, and Arianna Betti. 2021. Interrater disagreement resolution: A systematic procedure to reach consensus in annotation tasks. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 131–141.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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. *BioNLP 2019*, page 58.

Olesya Razuvayevskaya, Irene Lopez, Ian Dunham, and David Ochoa. 2023. Why clinical trials stop: the role of genetics. *medRxiv*, pages 2023–02.

Omid Rohanian, Mohammadmahdi Nouriborji, and David A Clifton. 2023. Exploring the effectiveness of instruction tuning in biomedical language processing. *arXiv preprint arXiv:2401.00579*.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. Capabilities of gemini models in medicine. *Preprint*, arXiv:2404.18416.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2023. Gollie: Annotation guidelines improve zero-shot information-extraction. *arXiv preprint arXiv:2310.03668*.

Amber Stubbs, Michele Filannino, Ergin Soysal, Samuel Henry, and Özlem Uzuner. 2019. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, 26(11):1163–1171.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58:S20–S29.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. 2013. Overview of the share/clef ehealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization: 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings 4*, pages 212–231. Springer.

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*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16:1–28.

Özlem Uzuner. 2009. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(4):561–570.

Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5):786–791.

Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic deidentification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Erik M. van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A. Kors, and Laura I. Furlong. 2012. The eu-adr corpus: Annotated drugs, diseases, targets, and their relationships. *Journal of Biomedical Informatics*, 45(5):879–884. Text Mining and Natural Language Processing in Pharmacogenomics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022. Deepstruct: Pretraining of language models for structure prediction. *arXiv preprint arXiv:2205.10475*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023a. Instructuie: multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2020. Medsts: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54:57–72.

Yuqing Wang, Yun Zhao, and Linda Petzold. 2023b. Are large language models ready for healthcare? a comparative study on clinical language understanding. In *Machine Learning for Healthcare Conference*, pages 804–823. PMLR.

Chih-Hsuan Wei, Lon Phan, Juliana Feltz, Rama Maiti, Tim Hefferon, and Zhiyong Lu. 2018. tmvar 2.0: integrating genomic variant information from literature with dbsnp and clinvar for precision medicine. *Bioinformatics*, 34(1):80–87.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Stephen Wu and Hongfang Liu. 2011. Semantic characteristics of nlp-extracted concepts in clinical notes vs. biomedical literature. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1550. American Medical Informatics Association.

Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, et al. 2020. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, et al. 2024. Me llama: Foundation large language models for medical applications. *arXiv preprint arXiv:2402.12749*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. *arXiv preprint arXiv:2005.07150*.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2023. Language models are super mario: Absorbing abilities from homologous models as a free lunch. *arXiv preprint arXiv:2311.03099*.

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Mozhi Zhang, Hang Yan, Yaqian Zhou, and Xipeng Qiu. 2023a. Promptner: A prompting method for few-shot named entity recognition via k nearest neighbor search. *arXiv preprint arXiv:2305.12217*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023b. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Yunkun Zhang, Jin Gao, Zheling Tan, Lingfeng Zhou, Kexin Ding, Mu Zhou, Shaoting Zhang, and Dequan Wang. 2024. Data-centric foundation models in computational healthcare: A survey. *arXiv preprint arXiv:2401.02458*.

Jin Zhao, Chao Liu, Jiaqing Liang, Zhixu Li, and Yanghua Xiao. 2024. A novel cascade instruction tuning method for biomedical ner. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11701–11705. IEEE.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition.

A Appendices

A.1 Unified Prompt Format

Utilizing the unified prompt format outlined in Table 1, we developed (1) the MNLU-Instruct dataset based on the collection of datasets detailed in Table 6; and (2) the evaluation dataset from BLUE and BLURB utilizing the labels from Table 3 and 4. In this section, we provide detailed information on dataset creation and examples of the input-output format for each task type.

A.1.1 Named entity recognition (NER)

We conduct NER at the sentence level, because most NER datasets comprise pre-split sentences.

For NER datasets where the medical text is an entire document, we use the SpaCy tokenizer[6] to split the document into sentences.

Below is an example of the NER input-output pair. The example is from the n2c2 2022 dataset (Lybarger et al., 2023), a shared task focused on extracting social determinants of health from clinical notes.

- NER Input -

Extract all relevant medical named entities faithfully from the medical text below. Focus on identifying the following entities: Living status, Tobacco, Drug, Employment, Alcohol.

Medical text: Denies any IV drug use or any recreational drug use.

- NER Output -

Living status: None
Tobacco: None
Drug: IV drug use ... recreational drug use
Employment: None
Alcohol: None

A.1.2 Event extraction (EE)

The EE task is composed of event trigger extraction (ETE), event argument extraction (EAE), and event argument classification (EAC). ETE uses the same prompting formats as NER. In EAE and EAC, we additionally include two adjacent sentences to provide more context information. Below are examples of the EAE and EAC input-output pairs from the n2c2 2022 dataset (Lybarger et al., 2023)

- EAE Input -

According to the medical text, what is the Method attribute of the Drug event 'IV drug use' in the medical text below? Extract the attribute faithfully from the medical text.

Medical text: ... Currently admits to five drinks of alcohol per week. Denies any IV drug use or any recreational drug use. Divorced with no children. ...

- EAE Output -

Drug - Method: IV

- EAC Input -

According to the medical text, what is the Status time attribute of the Drug event 'IV drug use' in the medical text below? Choose from the following options.

Medical text: ... Currently admits to five drinks of alcohol per week. Denies any IV drug use or any recreational drug use. Divorced with no children. ...

Options: (A) none (B) past (C) future (D) current

- EAC Output -

Drug - Status time: (A) none

A.1.3 Document classification (DC)

Our document classification task involves classifying a document or sentence into one or multiple pre-defined categories.

In the i2b2 2006Smoke (Uzuner et al., 2008) and i2b2 2008 (Uzuner, 2009) dataset, where the input document is a lengthy clinical note, we first deploy BioMistral to summarize the document. We use the prompt format, 'Summarize the {*type*} from the following clinical note.', where *type* is the corresponding DC type label, such as smoking status or asthma status.

The MTSamples dataset aims to classify a medical report into one of 48 medical specialties or domains (MTS, 2023). The large number of possible categories results in lengthy prompts. Instead, in each instance, we include the correct category along with 12 randomly selected negative categories in our prompts for more efficient training.

Below is an example of the DC input-output pair from the TrialStop dataset (Razuvayevskaya et al., 2023).

- DC Input -

According to the medical text below, which options best describe reason to stop the study? Choose from the following options. Multiple options can be true.

Medical text: 13 of 15 patients recruited.Study patients responded with no safety signals. Recruitment's slow, timely end of study necessary to keep development timelines.

Options: (A) Insufficient enrollment (B) Logistics resources (C) Business administrative (D) Insufficient data (E) Endpoint met (F) Negative (G) Study success (H) Regulatory (I) Interim analysis (J) Ethical reason (K) Invalid reason (L) Study design (M) No context (N) Another study (O) Covid19

- DC output -

(A) Insufficient enrollment (C) Business administrative

A.1.4 Relation extraction (RE)

The RE task focuses on classifying the relation between any possible entity pairs within the same sentence. We adapt the relation labels from the original publications into descriptive language. We additionally include two adjacent sentences to pro-

[6] https://spacy.io/api/sentencizer

vide more context information. Below is an example from the i2b2 2011 for coreference resolution on clinical named entities (Uzuner et al., 2012):

- RE Input -

According to the Medical text below, what is the co-reference relationship between the Person entity 'Mr. Andersen' and the Person entity 'who'? Choose from the following options.

Medical text: ... History of Present Illness: Mr. Andersen is a 71-year-old male with worsening anginal symptoms who underwent catheterization that showed severe three-vessel disease. He is presenting for revascularization . ... Options: (A) 'Mr. Andersen' refers to 'who' (B) None of the above.

- RE Output -

(A) 'Mr. Andersen' refers to 'who'

A.1.5 Multi-choice Question-answering (QA)

The QA task aims to answer a research question regarding the medical text within a pre-defined answer set. The PubMedQA dataset consists of research questions about PubMed abstracts, with answers categorized as yes, no, or maybe (Jin et al., 2019). The BioASQ includes biomedical questions with answers classified as yes or no (Tsatsaronis et al., 2015).

Directly applying our sequence classification prompt format for the QA task results in single-word multi-choice answers like *yes* or *no*. Instead, we transform the single-word options into descriptive sentences so that the QA output format is more straight-forward. We utilize one-shot learning with BioMistral to combine the question and each answer into a single statement. The one-shot example is randomly chosen from the PubMedQA train split, and the example output is written by human.

Below is an example of the QA input-output pair from the PubMedQA dataset, with descriptive multi-choice options.

- QA Input -

According to the medical literature below, Is there a connection between sublingual varices and hypertension? Choose from the following options. Only one option can be true.

Medical literature: BACKGROUND: Sublingual varices have earlier been related to ageing, smoking and cardiovascular disease. The aim of this study was to investigate whether sublingual varices are related to presence of ...

Options: (A) The answer is not mentioned in the text (*maybe*). (B) There is a connection between sublingual varices and hypertension (*yes*). (C) There is not a connection between sublingual varices and hypertension (*no*).

- QA Output -

(B) There is a connection between sublingual varices and hypertension (*yes*).

A.1.6 Natural language inference (NLI)

The NLI task utilizes a similar multi-choice prompt format to other sequence classification tasks. Below is an example from the BioNLI dataset (Bastan et al., 2022)

- NLI Input -

What is the relationship of the hypothesis with respect to the premise? Choose from the following options.

Premise: The administration of heparin with or without ACTH significantly decreased hepatic cholesterol content in catfish. In serum, heparin alone produced first hypercholesterolemia which was followed by hypocholesterolemia whereas it potentiated hypercholesterolemic action of ACTH three hours after administration.

Hypothesis: It is concluded that heparin inhibits the cholesterol-lowering action of ACTH in catfish.

Options: (A) neutral (B) entailment (C) contradiction

- NLI Output -

(C) contradiction

A.1.7 Semantic text similarity (STS)

We adapt the scoring criteria from the original publications and translate the numerical similarity scores into a descriptive sentences. Below is an example from the STS-B dataset (Wang et al., 2018)

- STS Input -

How similar are the two sentences below? Choose from the following options.

Sentence 1: A plane is taking off.

Sentence 2: An air plane is taking off.

Options: (A) The two sentences are completely dissimilar. (B) The two sentences are not equivalent, but are on the same topic. (C) The two sentences are not equivalent, but share some details (D) The two sentences are roughly equivalent, but some important information differs / missing. (E) The two sentences are mostly equivalent, but some unimportant details differ. (F) The two sentences are completely or mostly equivalent, as they mean the same thing.

- STS Output -

(F) The two sentences are completely or mostly equivalent, as they mean the same thing.

A.1.8 Natural language generation (NLG)

We focus on the abstractive summarization task from NLG. Although abstractive summarization is different from our focus on NLU, it also requires in-depth document understanding, and thus we believe it can help improve NLU performance. We include three NLG datasets in the MNLU-Instruct: PubMedSum (Cohan et al., 2018), CDSR (Guo et al., 2021), and AciDemo (Yim et al., 2023). PubMedSum has the input as the complete PubMed articles and the output as their abstracts. CDSR is a text simplification task that translates domain-specialized summaries into lay-user summaries. AciDemo is a task that summarizes doctor-patient dialogues into clinical note sections. Because the PubMedSum and AciDemo documents can be very lengthy, we only include instances with less than 800 words. Additionally, we restrict the output in PubMedSum to be at most half of its corresponding input word count to ensure that the PubMedSum splits contain high-quality summaries.

Below is an example from the AciDemo dataset.

- NLG Input -

Summarize the relevant medical information from a dialogue between a doctor and a patient. The summary should be the objective exam section from the clinical note. Output None if no relevant information is found.

Dialogue:[doctor] hi alan , how are you ?

[patient] hi , good to see you .

[doctor] good to see you as well . are you ready to get started ?

[patient] ...

- NLG Output -

PHYSICAL EXAMINATION

Neck: Supple.

No jugular venous distension.

Respiratory: Slight expiratory wheezing bilaterally.

Cardiovascular: Regular rate and rhythm. No murmurs.

Musculoskeletal: Trace edema in the bilateral lower extremities.

A.2 Baseline system with ICL for NER tasks

Generalized LLMs do not automatically extract named entities in a unified format. To avoid confounding factors from different output formats and simplify NER evaluation, we utilize the same NER input-output format as described in Appendix A.1.1. Additionally, we include a descriptive paragraph at the beginning of the input prompt to specify the output format: "Your answer should use the following format, with one entity type per line. The span refers to the original text span from the Medical text. Output None if there is no such span. Use '...' to separate multiple spans."

We also include two in-context examples to ensure the baseline system adheres to the desired output format. For each inference query, the 2-shot examples are randomly selected from the training split of each dataset. We ensure the outputs from the 2-shot examples are different from each other, to prevent bias towards a specific extraction response.

Task	dataset	# instances	Labels
NER	i2b2 2006DeID	5,608	Location, ID, Date, Hospital, Doctor, Contact, Name, Age
	i2b2 2011	25,689	Person, Treatment, Test, Problem
	i2b2 2012	7,446	Test, Problem, Frequency, Time, Date, Occurrence, Treatment, Duration, Clinical department
	i2b2 2014	52,462	ID, Contact, Age, Name, Location, Profession, Date
	GENIA	15,023	RNA, DNA, Cell type, Protein, Cell line
	linnaeus	11,935	Species
	tmVar	5,351	Cell Line, SNP, Gene, Protein Mutation, Protein Allele, Species DNA Allele, DNA Mutation, Other Mutation, Acid Change,
	DrugProt	17,274	Organism Taxon, Disease Or Phenotypic Feature, Cell Line, Gene Or Gene Product, Sequence Variant, Chemical
	BioRed	13,706	Chemical, Gene
	GNorm	4,006	Family Name, Domain Motif, Gene
	NLM-Gene	5,048	Gene, Gene reference into function (function of a gene), Domain, Steroidogenic acute regulatory protein (a protein coding gene)
	ClinicalIE_Med	105	Route, Duration, Reason, Dosage, Frequency, Medication
	ClinicalIE_Status	105	Neither medications, Discontinued medications, Active medications
	BC4CHEMD	30,682	Chemical
	PubMed PICO	1,961	Species, Comparator, Outcome, Intervention, Strain, Induction
	PICO-Data	36,224	Participants, Intervention, Outcome
EE	i2b2 2009	117,446	Medication (Dosage, Route, Frequency, Duration, Reason, Context)
	i2b2 2018	155,716	Drug, ADE (Strength, Frequency, Reason, Form, Route, Dosage)
	n2c2 2022	36,359	Alcohol, Drug, Tobacco, Employment, Living (time, duration, history, type, amount, frequency)
DC	i2b2 2006Smoke	398	Current smoker/Past smoker/Non-smoker/Unknown
	i2b2 2008	17,242	10 obesity commodities (Asthma, Depression, ...)
	n2c2 2018	2,626	Different selection criteria for 13 cohorts (Abdominal, English, ...)
	2024 SemEval2	1,700	Adverse Events, Eligibility, Results, Intervention
	TrialStop	3,747	17 reasons to stop a study (Study staff moved, Another study, ...)
	MTSamples	3,206	48 medical specialties or domains (Bariatrics, Nephrology, ...)
RE	i2b2 2011	25,689	Refers to
	i2b2 2012	7,446	Ends by, Happens during, Happens before and overlap, Begins by, Happens before, Happens simultaneously with, Happens after, Overlaps with,
	EUADR	318	Gene-disease association
	DrugProt	35,624	Antagonist, Agonist, Indirect upregulator, Part of, Agonist activator, Substrate, Activator, Inhibitor, Direct regulator, Agonist inhibitor, Product of, Substrate product of, Indirect downregulator
	BioRed	4,328	Drug interaction, Positive correlation, Cotreatment, Comparison, Bind, Conversion, Association, Negative correlation
NLI	Multi-NLI	785,404	Entailment, Contradiction, Neutral
	SNLI	1,098,734	Entailment, Contradiction, Neutral
	BioNLI	23,704	Entailment, Contradiction, Neutral
STS	SIS-B	11,018	6 similarity scales
NLG	PubMedSum	1,407	Article summarization
	CDSR	436	Article simplication
	AciDemo	204	Dialogue to note summarization

Table 6: Task labels and number of instances in the MNLU-Instruct datasets. For EE tasks, labels inside () refer to event arguments.