MedINST: Meta Dataset of Biomedical Instructions

Wenhan Han[1], Meng Fang[2,1], Zihan Zhang[3], Yu Yin[2],
Zirui Song[3], Ling Chen[3], Mykola Pechenizkiy[1], Qingyu Chen[4]
[1]Eindhoven University of Technology [2]University of Liverpool
[3]University of Technology Sydney [4]Yale University
w.han@tue.nl, Meng.Fang@liverpool.ac.uk, qingyu.chen@yale.edu

Abstract

The integration of large language model (LLM) techniques in the field of medical analysis has brought about significant advancements, yet the scarcity of large, diverse, and well-annotated datasets remains a major challenge. Medical data and tasks, which vary in format, size, and other parameters, require extensive preprocessing and standardization for effective use in training LLMs. To address these challenges, we introduce MEDINST, the Meta Dataset of Biomedical Instructions, a novel multi-domain, multi-task instructional meta-dataset. MEDINST comprises 133 biomedical NLP tasks and over 7 million training samples, making it the most comprehensive biomedical instruction dataset to date. Using MEDINST as the meta dataset, we curate MEDINST32, a challenging benchmark with different task difficulties aiming to evaluate LLMs' generalization ability. We fine-tune several LLMs on MEDINST and evaluate on MEDINST32, showcasing enhanced cross-task generalization.

1 Introduction

Recent advancements in large language models (LLMs), such as GPT-4 (OpenAI, 2024), LLaMA-3 (Meta, 2024) and Mistral (Jiang et al., 2023) have demonstrated impressive performance across various open-domain NLP tasks. Rather than developing specialized, task-specific systems, there is an increasing focus on rapidly adapting LLMs to specific tasks through simple prompting techniques. Studies have demonstrated that such prompted LLMs can achieve and even outperforms the capabilities of specialized models in a variety of NLP tasks (Radford et al.; Brown et al., 2020; Wei et al., 2022; Sanh et al., 2022). Due to the high cost of pre-training LLMs, instruction finetuning has become the standard method for adapting base LLMs to specific domains. Therefore, training domain-specific LLMs has largely shifted to a data-centric approach.

In recent years, the field of medical analysis has experienced a transformative shift with the integration of large language model (LLM) techniques, fundamentally expanding the landscape of diagnostic and therapeutic strategies. The advancement of this field relies heavily on the availability of large, diverse, and well-annotated datasets, which are crucial for training robust and effective machine learning (ML) models. Although specialized biomedical models such as BioBERT (Lee et al., 2020), ClinicalXLNET (Huang et al., 2019), BioM-Transformers (Alrowili and Shanker, 2021) and SciFive (Phan et al., 2021) have achieved a success, they rely on task-specific modules and follow a pre-train then fine-tune paradigm for specified tasks (Liu et al., 2021; Wang et al., 2023a). In this context, generalizing to unseen tasks is computationally expensive and time-consuming. Attempts exist such as In-BoXBART (Parmar et al., 2022) and BioMistral (Labrak et al., 2024) are finetuned with biomedical instructions. However, the data involved in training and evaluation are limited. Collecting raw medical data and converting it into a format suitable for LLM applications is often complex and challenging. Medical data and tasks vary significantly in format, size, and other parameters, necessitating extensive preprocessing and standardization. This task becomes even more intricate when integrating multiple datasets from various domains into a cohesive, standardized format. This raises the necessity of a comprehensive biomedical instruction meta-dataset.

To address the problem, we release MEDINST[1], an instruction dataset collection includes 133 biomedical NLP tasks in 12 categories such as Named Entity Recognition (NER), Question-

[1]The code, models and data are available at https://github.com/aialt/MedINST.

Findings of the Association for Computational Linguistics: EACL 2024, pages 8221–8240

Resource	MEDINST (this work)	SUP-NATINST (Wang et al., 2022) (Biomedicine)	BoX (Parmar et al., 2022)	BLURB (Gu et al., 2021)
Has task instructions?	✓	✓	✓	×
Has multi-task datasets?	✓	×	×	×
Has examples?	✓	✓	✓	×
Is public?	✓	✓	✓	✓
Number of tasks	133	30	32	13
Number of instructions	133	30	32	-
Number of annotated task types	12	-	9	6
Avg. task definition length (words)	45.98	56.6	-	-

Table 1: Comparison of MEDINST to several datasets in biomedical field.

Answering (QA), Relation Extraction (RE), etc. A benchmark is set by curate a test set from the entire collected dataset. In the experiment, multiple scales of LLMs are finetuned on our training data to demonstrate the generalization performance enhancement. Table 1 presents the comparison of MEDINST to relevant datasets in biomedical field. In summary, our contributions are:

- We release a novel dataset MEDINST, a biomedical instruction meta-dataset that involves 7M samples spanning 133 tasks among 12 categories.
- Using the meta dataset, we curate MEDINST32, a challenging benchmark for evaluating the cross-task generalization ability of LLMs in the biomedical domain.
- We introduce instruction fine-tuned LLMs on MEDINST based on LLaMA-3 and conduct comprehensive evaluation and analysis across multiple baselines.

2 Related Work

2.1 Instruction Finetuning

Instruction finetuning involves training models to follow specific instructions, often resulting in improved generalization and the ability to perform a wider range of tasks (Wei et al., 2022). There are already numerous open-domain instruction datasets and finetuned models. NATURAL INSTRUCTIONS is curated from samples of different NLP datasets and the crowdsourcing instructions used to annotate them. FLAN 2021 (Wei et al., 2022) and 2022 (Longpre et al., 2023) provide extensive publicly available set of tasks and methods for instruction tuning. FLAN models are trained on the collection and exhibits strong generalization performance on a variety tasks. The InstructGPT (Ouyang et al., 2022) model benefits in part from a substantial dataset of prompts gathered through various synthetic data augmentation methods. However, this dataset is not publicly accessible. SUPER-NATURAL INSTRUCTIONS (Wang et al., 2022) is established as a benchmark of 1,616 diverse NLP tasks along with expert-written instructions. The collection covers 76 distinct task types, providing a rigorous benchmarking of generalization performance of LLMs. The corresponding trained model Tk-INSTRUCT outperforms InstructGPT despite being an order of magnitude smaller. Self-Instruct (Wang et al., 2023b) provides a new approach for instruction fine-tuning. It involves bootstrapping off the generations of pre-trained language models to improve the instruction-following performance of themselves. After the great success of ChatGPT (OpenAI, 2022), many efforts have been made to use data generated by ChatGPT to train their own large language models (LLMs). Alpaca (Taori et al., 2023) is finetuned from LLaMA (Touvron et al., 2023) on 52k instruction-following instances generated by Text-davinci-003. Compared to open-domain instruction datasets, instruction datasets in the biomedical field are relatively scarce. MedAlpaca (Han et al., 2023) utilizes a data collection of 160k entries from a reformatted medical NLP task and a crawl of internet resources. ChatDoctor (Li et al., 2023) is trained using 100k patient-doctor dialogues from an online medical consultation platform. Similar to Alpaca, AlpaCare (Zhang et al., 2024) uses medical related instruction demonstrations generated by ChatGPT to train on LLaMA. By prompting ChatGPT to conduct self-chat, Baize (Xu et al., 2023) collect the dialogues to train a specialized model for healthcare. Additionally, BioMistral (Labrak et al., 2024) and PMC-LLaMA (Wu et al., 2023) use medical-related corpora to pretrain their respective base models, followed by finetuning with an instruction dataset. All these models are only finetuned on a limited number of tasks, making them prone to failure when confronted with new tasks. Our dataset focuses on biomedical domain, offering comprehensive instruction-following demonstrates spanning 133 tasks in 12

task categories, facilitating LLMs generalizing to unseen tasks.

2.2 Biomedical Benchmarks

Biomedical workshops, such as BioNLP (Kim et al., 2009) and BioCreative (Hirschman et al., 2005), often employ task-specific benchmark datasets. With the rise of LLMs, there are higher expectations for the comprehensive capabilities of medical models. As a result, evaluating them on a single task is no longer sufficient. BLUE (Biomedical Language Understanding Evaluation) (Peng et al., 2019) took the first step by constructing a benchmark that includes 10 datasets covering 5 different task types. Building on this foundation, BLURB (Biomedical Language Understanding and Reasoning Benchmark) (Gu et al., 2021) expanded the dataset to 13, encompassing 7 different types. Instruction datasets exist for few and zero-shot evaluations. Agrawal et al. (2022) introduce 3 datasets for clinical information extraction by reannotating the CASI datset. SUPER-NATURAL INSTRUCTION (Wang et al., 2022) delivers 1600+ open-domain NLP tasks, among which 30 tasks are related to medicine and healthcare. Tailored for biomedicine, BoX (Parmar et al., 2022) provides 32 tasks in the scope of 9 categories. BigBIO (Fries et al., 2022) focuses on the process of constructing meta-datasets, providing unified schema for 126 existing datasets across various tasks and offering tools for building new datasets. However, it does not contain instructions and the datasets are not in text generation format. Our dataset offers an extensive instruction benchmark including 32 tasks representing a comprehensive evaluation of LLM performance in biomedical fields.

3 MEDINST: Meta Dataset of Biomedical Instructions

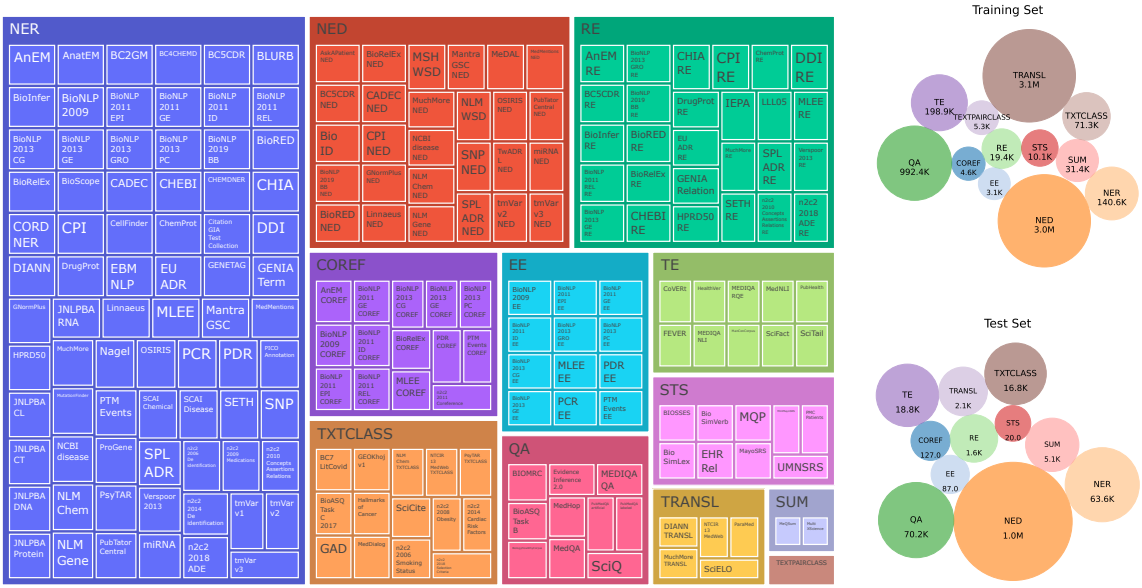
We curate MEDINST by collecting 98 well-adopted biomedical datasets from 12 task categories and reformulating them into 133 tasks. All tasks are regarded as text generation task and the data are formatted to instruction-following samples. The instructions are human annotated and tailored for each dataset/task. Figure 1 (a) depicts a visualization of the dataset composition of MEDINST.

3.1 Tasks

Figure 1 (b) shows the number of samples included in each task categories. We adopted 12 categories of tasks, where each may have several sub-categories. The categories are as follows:

Named Entity Recognition (NER) NER is a task in natural language processing that involves identifying and classifying key information entities. In the biomedical field, NER involves detecting and extract key entities such as diseases, drugs, genes, and other relevant biological terms within biomedical texts. In MEDINST, 56 NER datasets are collected, including the most commonly used BC5CDR (Li et al., 2016), JNLPBA (Collier et al., 2004), LINNAEUS (Gerner et al., 2010), etc. We have created a unified instruction template for the NER task and made variations based on the specific requirements of each dataset. We have divided the NER task into two sub-categories, differing in output format. Sub-category 1 requires labeling each word in the input text using the BIO format, while Sub-category 2 requires directly outputting all detected entities that meet the criteria. In Sub-category 1, the input for each instance is a single sentence, whereas in Sub-category 2, the input is an entire passage. This adds diversity to the NER task and creates different levels of difficulty, thereby enhancing the model's stability in handling various output requirements and understanding longer texts.

Named Entity Disambiguation (NED) The NED task involves determining the correct identity of named entities in a text by linking them to a specific entry in a knowledge base. Most of the NER datasets contain annotations for entity disambiguation. The NED task has also been repurposed into two difficulty levels. The AskAPatient and TwADR datasets (Limsopatham and Collier, 2016) are used to create simpler tasks, where the input includes a specified biomedical entity and its context, and the requirement is to output its identifiers in the corresponding database. Other dataset such as BioRelEx (Khachatrian et al., 2019), CPI (Döring et al., 2020), MedMentions (Mohan and Li, 2019), etc. have been reformatted into more challenging tasks, requiring the extraction of relevant biomedical entities from the given text and providing the corresponding identifiers for each entity. In additional, MeDAL dataset (Wen et al., 2020) has also been included in the NED task, which is a medical text dataset curated for abbreviation disambiguation. We include a total of 23 datasets in the NED task category.



(a) Treemap. (b) Number of samples.

Figure 1: MEDINST overview.

			NER	RE	NED	QA	COREF	EE	TE	STS	TXTCLASS	TRANSL	SUM	TEXTPAIRCLASS	ALL
# Dataset	MEDINST	Train	56	24	21	13	13	10	8	7	5	3	2	1	163
		Dev	30	11	10	8	10	7	5	1	4	1	1	-	88
		Test	37	9	12	10	2	1	8	1	5	1	1	-	87
	MEDINST32	Train	43	21	19	10	11	9	5	6	3	2	1	1	131
		Dev	19	9	9	6	8	6	5	-	2	-	-	-	64
		Test	13	3	2	3	2	1	3	1	2	1	1	-	32
# Instruction/Task			49	23	19	9	7	9	3	3	5	3	2	1	133

Table 2: Dataset statistics across various categories.

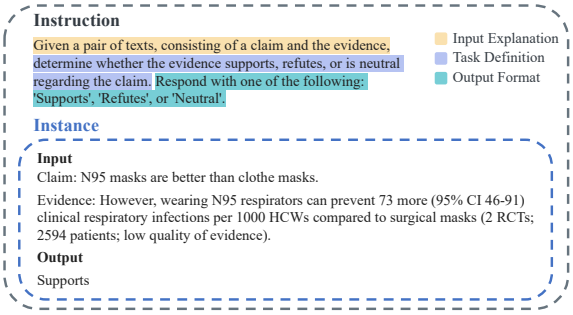


Figure 2: Instruction and instance example.

Relation Extraction (RE) RE involves identifying and categorizing the relationships between entities within a given text. We utilize 24 datasets for RE task, including AnEM (Ohta et al., 2012), BioNLP 2011 REL (Pyysalo et al., 2011), etc. We simplified the task by listing all the possible relation types in the instruction for each dataset. The language model is prompted to extract all possible triples from the input text.

Coreference Resolution (COREF) COREF is the task of determining which words or phrases in a text refer to the same entity. We used 13 datasets for this task category, most of which come from the BioNLP Shared Task. In addition, the MLEE (Pyysalo et al., 2012) and PDR (Kim et al., 2019) datasets have also been included.

Question-Answering (QA) Multiple types of QA are collected, including yes/no, yes/no/maybe, factoid, multi-choice, etc. In this category, 10 datasets are employed and reformatted. For multiple-choice QA, we write out the full options in the output rather than assigning letters or numbers to each option.

Textual Entailment (TE) Determining whether two texts contradict each other and whether a statement aligns with the facts is crucial in the medical field. In this category, we re-format 6 fact-checking datasets, FEVER (Thorne et al., 2018), HealthVer (Sarrouti et al., 2021), SciFact (Wadden

et al., 2020), PubHealth (Kotonya and Toni, 2020), etc., into claim-evidence pairs. These datasets range from the general scientific domain to specific medical domains, such as COVID-19. Moreover, MEDIQA-RQE (Ben Abacha et al., 2019) is incorporated as a question entailment task, i.e. determine whether the meaning of one question can be inferred from another question. As a classic task in the TE category, the premise-hypothesis entailment task is represented by the SciTail dataset (Khot et al., 2018).

Text Classification (TXTCLASS) The text classification task involves assigning predefined categories or labels to a given piece of text based on its content. Although the input and output formats for text classification tasks are relatively fixed, the definitions and objectives of each task are highly diverse. Therefore, it is challenging to use a template to standardize this type of instruction. To ensure the quality of the instructions, each task within this category is entirely crafted manually. We collect 5 datasets, SciCite (Cohan et al., 2019), Hallmarks-of-Cancer (Baker et al., 2016), BC7-LitCovid (Chen et al., 2022), MedDialog (Zeng et al., 2020) and GEOKhoj-v1[2], for this category.

Semantic Similarity (STS) The Semantic Similarity task aims at measuring how similar the meanings of two pieces of text are to each other. Originally a regression task with similarity scores as outputs, we have redefined it as a classification task by categorizing the similarity scores of all datasets into six integer levels from 0 to 5, where 0 indicates completely unrelated and 5 indicates highly similar. This task category includes 7 datasets, e.g. Bio-SimVerb, Bio-SimLex (Chiu et al., 2018) BIOSSES (Soğancıoğlu et al., 2017), MQP (McCreery et al., 2020), etc.

Event Extraction (EE) EE task requires identifying and categorizing events, such as biological processes or interactions, within biomedical texts. The Event Extraction (EE) task is typically complex, with events in documents often containing nested structures. To format the EE task as a text generation task, we simplify it according to the BioNLP 2009 Core Event Detection subtask (Kim et al., 2009). We only detect events within a given range of types and their primary arguments. Note that primary arguments must be a biomedical entity within the text; we do not consider cases where primary arguments refer to another event.

Translation (TRANSL) We have included the MuchMore (Buitelaar et al., 2003), ParaMed (Liu and Huang, 2021) and SciELO (Soares et al., 2018) datasets translated from German, Chinese, and Spanish into English.

Text Pair Classification (TEXTPAIRCLASS) For this category, we employ a sentiment analysis dataset, the Medical-Data[3] , which analyzing the sentiment in a text where a drug is mentioned to determine whether the sentiment towards the drug is positive, negative, or neutral.

Summarization (SUM) Summarization is also crucial for the application of LLMs in the biomedical field. In this category, we use the MeQSum (Ben Abacha and Demner-Fushman, 2019) and Multi-XScience (Lu et al., 2020) datasets. MeQSum presents patient questions, often in the form of lengthy texts, and the task requires capturing the main concern of these questions and providing a concise rewrite. Multi-XScience is a multi-document summarization task, which requires generating a related work section for a given article based on its abstract and the abstracts of some cited references.

The complete dataset collection details are listed in the Appendix E.

3.2 Instruction Construction

All instructions are written according to a unified schema to ensure their quality. An instruction includes the following elements:

Input Explanation The instruction first specifies the structure of the input. For example, for NER, the given input is typically a sentence or a passage; for QA tasks, the input can be a question alone, a question with context, or a question with context and options. We describe the elements included in the input for each dataset's task individually, avoiding the use of generalized descriptions.

Task Definition The instructions include an explanation of the task and the specific actions the model needs to perform. The task definition is tailored to the content of each dataset and specifies any optional parameters. For example, the definition for the SciCite (Cohan et al., 2019) task is

[2] https://github.com/ElucidataInc/GEOKhoj-datasets/tree/main/geokhoj_v1

[3] https://www.kaggle.com/datasets/arbazkhan971/analyticvidhyadatasetsentiment

"Classify the intent of the citation within this context. Intents are: [background, method, result]." avoiding vague instructions like "Classify the text into [background, method, result]."

Output Format Here we specify the format of the output. In MedINST, we adopt formats corresponding to the complexity of the output content. For open text generation, the output is generally plain text; for classification tasks, multiple labels are separated by commas; for tasks like NER, where the output biomedical entities may contain various special characters, we enclose them in square brackets. For complex outputs, the instruction will provide a template example of the output format.

After drafting instructions according to the abovementioned elements, we further proofread them to make them more concise and aligned with natural human instructions, avoiding rigid, structured descriptions. Appendix B presents the examples of instructions.

3.3 MEDINST32 Benchmark Construction

Using the MEDINST as a meta dataset, we carefully curate MEDINST32, a challenging benchmark that covers 32 tasks with different difficulties to evaluate LLMs' performance across various medical-related tasks comprehensively. Unlike previous works, the tasks selected for MEDINST32 encompass different difficulty levels, including *knowledge difficulty* and *instruction difficulty*. Specifically, knowledge difficulty assesses the model's amount of biomedical knowledge, such as understanding levels of biomedical terms and their relationships, while instruction difficulty evaluates the model's understanding and adherence to instructions. We divide difficulty into four categories and choose tasks from simplest (e.g., acronym completion) to hardest (e.g., RE, EE). Moreover, two positive examples are offered for each tasks. See more details in Appendix A.

4 Experiments

4.1 Setup

Problem Formulation. We combine the training sets to train multi-task biomedical models. Given an instruction $Inst_t$ for a task t, and the dataset (X_t, Y_t), multi-task models learns a map $M_t : (Inst_t, x) \rightarrow y$, where $(x, y) \in (X_t, Y_t)$. After learning a set of maps $M_1, M_2, ..., M_T$, the multi-task models can generalize to unseen tasks $i \in \{T+1, T+2, ..., T+N\}$ and approximate the maps M_i, where $M_i : (Inst_i, x) \rightarrow y$, $(x, y) \in (X_i, Y_i)$.

Training Data. Our goal is to test the generalization ability of LLMs on unseen tasks after instruction tuning multiple biomedical tasks. Once we have selected the 32 tasks in MEDINST32 (Sec. 3.3), we use the training set of the *remaining* tasks from MEDINST for multi-task fine-tuning. Since the MEDINST training set is too large and a large number of training instances per task do not help generalization in instruction finetuning (Wang et al., 2022), we sample 100K samples to train our multi-task biomedical LLMs, denoted as **MI32**. We select an equal number of samples from each task category to ensure balance across all tasks.

Evaluation setup. Following Wang et al. (2022), we limit the test set for large size datasets aiming at efficient evaluation. We observe that models not fine-tuned on MedINST sometimes struggled to output according to the instructions, posing challenges for post-processing and metric calculation. To ensure a fair comparison, we use few-shot prompts for baseline models during evaluation. Each test task is provided with two examples to help zero-shot models output in the standard format. Appendix C details the implementation of training and evaluation.

Model. We fine-tune the instruction-tuned LLaMA-3 (8B; Meta, 2024) and MMed-LLaMA-3 (8B; Qiu et al., 2024) on the aforementioned **MI32** training set and derive **LLaMA3-MI32** and **MMedL3-MI32**, respectively. Additionally, we fine-tune LLaMA-3 on the 100K samples from MEDINST, where the training sets of the datasets in **MI32** are exposed to the model, to produce **LLaMA3-MI**, as an oracle model.

Baselines. As a direct comparison, we compare our **LLaMA3-MI32** fine-tuned on **MI32** with its base version, **LLaMA3**. Since MMed-LLaMA-3 is a foundation model that has not been instruction fine-tuned, to make a fair comparison, we use MMed-LLaMA-3-EnIns (Qiu et al., 2024), which is fine-tuned on the English medical instruction dataset from PMC-LLaMA (Wu et al., 2023). We denote it as **MMedL3-EnIns**. In addition, we compare **BioMistral**, an open-source LLM further pretrained on PubMed Central utilizing the instruction fine-tuned version of Mistral-7B (Jiang et al.,

2023) and **GPT-4o**, an advanced variant of GPT-4, excels in the biomedical domain with enhanced capabilities for understanding and generating complex medical and scientific texts.

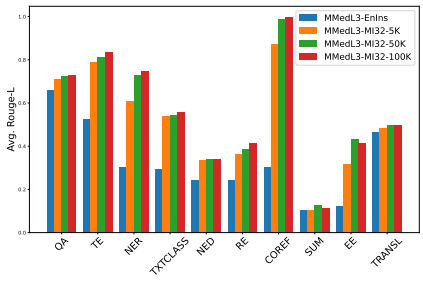
Metrics. Inspired by BLURB (Gu et al., 2021), we select appropriate metrics for each task in MEDINST32, including **Rouge-L**, **Entity F1** (Entity-level F1), **Label F1** (Label-level F1), **MSE** (Mean Squared Error) and **EM** (Exact Match). Entity-level F1 measures the overlap between the entities detected by the models and the ground truth, which is calculated by each data sample. Label-level F1 is calculated from the entire dataset to measure the similarity between the model's predictions and the labels.

4.2 Results

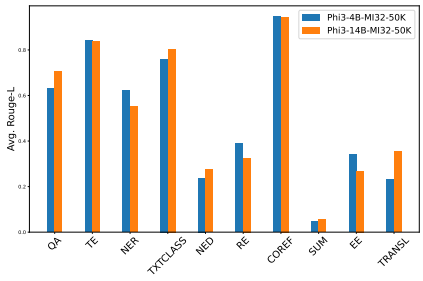
Table 3 presents the evaluation results of our models and baselines on MEDINST32. As an oracle model, **MMedL3-MI** demonstrated excellent performance across various difficulty levels, outperforming **GPT-4o** in 25 tasks. This highlights the significant impact of the MEDINST dataset in enhancing the overall performance of models on biomedical tasks.

The two zero-shot models, **LLaMA3-MI32** and **MMedL3-MI32**, showed significant generalization improvements over their base models in most unseen tasks. They respectively outperformed GPT-4o in 15 and 13 tasks. However, surprisingly, **MMedL3-MI32**, which used MMed-LLaMA-3 (further pretrained on biomedical corpora) as its base model, lagged behind LLaMA3-MI32 in 22 tasks. This indicates that using further pretraining to specialize a general LLM to the biomedical domain may not be as effective as instruction finetuning, especially considering the substantial computational resources required for pretraining. This also underscores the necessity of building a comprehensive biomedical instruction meta-dataset.

MMedL3-EnIns was fine-tuned on 500K medical question-answering data, which includes training data from MedQA and PubMedQA that appeared in MEDINST32. Despite using few-shot prompting, its performance on MEDINST32 was still unsatisfactory. It even significantly lagged behind in QA tasks, especially in MedQA, achieving only a 15.40 accuracy. This highlights the necessity of reformulating tasks to improve model generalization capabilities: training models to output in a single format alone increases the risk of overfitting.



(a) Performance with varying training data sizes.



(b) Performance with varying model parameter sizes.

Figure 3: Training sample and model parameter scale analysis.

4.3 Ablation Analysis

We design experiments to explore the impact of the number of training samples and model parameters on finetuning performance. We employ the same strategy to sample 5K and 50K instances from the MI32 training set for training two additional MMedL3-MI32 models for comparison. Additionally, we trained both 4B and 14B versions of Phi-3 using the 50K dataset.

In Figure 3, we calculate the average Rouge-L score for each task category to measure the performance of the models. In (a), it can be seen that as the number of training samples increases, the model's overall performance improves. However, performance deteriorates with increased sample size in tasks such as summarization (SUM) and event extraction (EE). This is because as sampling expands, the proportion of smaller datasets decreases, leading to data imbalance, which causes uneven learning progress across different tasks. Part (b) demonstrates unexpected results regarding the scale of model parameters. Phi-3-14B performs less than the 4B version in three core tasks for the biomedical field: NER, RE, and EE. A possible reason is that larger models require more data to be fully optimized and achieve generalization performance on unseen biomedical tasks. Specialized

Category	Dataset	Difficulty Level	Metric	Model						
				LLaMA3	BioMistral	MMEDL3-EnIns	GPT-4o	LLaMA3-MI32	MMEDL3-MI32	LLaMA3-MI†
				(Few Shot)				Ours (Zero Shot)		
NER	NCBI-disease	2	Label-F1	51.67	24.00	30.59	47.57	78.55	78.20	**84.61**
	BC5CDR	2	Label-F1	58.68	33.86	28.77	75.11	81.28	73.57	**87.39**
	AnEM	3	Entity-F1	8.20	3.66	1.72	37.44	32.03	31.38	**49.44**
	BioNLP-2009	2	Entity-F1	30.33	22.24	19.71	57.83	76.06	78.61	**80.74**
	BioNLP-2011-GE	2	Entity-F1	29.60	20.97	14.40	57.43	76.29	79.89	**80.39**
	BioNLP-2011-ID	3	Entity-F1	32.83	18.45	21.19	**68.59**	51.80	50.80	**76.26**
	BioNLP-2011-REL	2	Entity-F1	30.14	22.73	20.26	59.01	75.93	78.66	**80.41**
	BioNLP-2013-CG	3	Entity-F1	24.46	10.49	8.63	57.59	56.36	51.61	**72.32**
	BioNLP-2013-GE	2	Entity-F1	16.98	15.49	13.29	43.74	**71.59**	71.25	71.32
	BioNLP-2013-GRO	4	Entity-F1	10.34	4.14	2.91	**37.79**	12.48	12.86	35.13
	BioNLP-2013-PC	3	Entity-F1	31.96	19.45	19.57	68.75	62.94	61.38	**82.05**
	BioRED	3	Entity-F1	29.38	16.45	16.33	60.73	74.01	72.45	**78.76**
	tmVar-v3	3	Entity-F1	16.34	8.96	0.39	42.08	58.46	56.77	**63.22**
QA	BioASQ-Task-B-yesno	1	Label-F1	91.62	67.57	91.82	93.52	93.10	86.19	**93.87**
	PubMedQA-labeled	2	Label-F1	50.85	23.73	48.28	56.11	53.81	53.65	**59.94**
	MedQA	2	EM	49.25	24.51	15.40	**81.93**	47.68	45.72	53.26
TE	SciFact	2	Label-F1	42.09	36.33	33.69	92.61	85.85	84.14	**95.06**
	ManConCorpus	2	Label-F1	66.66	29.92	51.83	60.09	68.20	68.57	**69.14**
	CoVERt	1	Label-F1	82.24	47.77	55.87	93.76	91.15	93.49	**96.93**
TXTCLASS	Hallmarks-of-Cancer	2	Entity-F1	45.33	**54.77**	11.93	42.40	44.01	32.65	45.84
	MedDialog	1	Label-F1	91.34	86.02	56.52	98.77	96.72	77.67	**100.00**
NED	MeDAL	2	EM	21.6	15.90	17.00	**59.40**	28.90	30.00	36.60
	tmVar-v3-NED	4	Entity-F1	0.18	0.05	0.00	**7.45**	2.84	0.78	1.10
RE	AnEM-RE	4	Entity-F1	2.56	0.00	5.13	**25.64**	0.20	1.54	16.24
	BC5CDR-RE	4	Entity-F1	4.28	6.27	3.34	9.46	14.21	13.69	**27.93**
	BioInfer-RE	4	Entity-F1	18.74	9.73	8.86	17.49	28.06	26.23	**32.83**
COREF	AnEM-COREF	1	Entity-F1	34.52	14.29	21.43	82.20	**100.00**	**100.00**	**100.00**
	MLEE-COREF	1	Entity-F1	54.17	26.55	25.66	79.97	**99.12**	98.23	95.72
SUM	Multi-XScience	2	Rouge-L	13.28	11.61	10.36	12.78	11.61	11.57	**14.51**
EE	MLEE-EE	4	Entity-F1	0.96	0.19	0.09	9.88	**30.47**	28.61	27.48
STS	BIOSSES	1	MSE↓	2.05	4.15	4.15	**0.6**	1.05	2.15	1.20
TRANSL	ParaMed	2	Rouge-L	47.51	50.49	46.49	**63.08**	49.01	49.65	59.32

Table 3: Test results of various models on MEDINST32. † indicates that the training sets of LLaMA3-MI includes the corresponding training sets of the datasets used by MEDINST32, whereas other models have not seen the MEDINST32 dataset. ↓ represents that a lower score is better, while for other metrics, a higher score is better. The best and second-best results for each row are highlighted in bold and underlined, respectively. For the baselines, we use a few-shot prompt, providing two examples in the instruction. For the fine-tuned models, we use a zero-shot prompt.

Method	MMLU						
	An	CK	CB	CM	MG	PM	Avg.
BioMistral	48.89	66.42	63.19	58.38	70.00	58.46	60.88
MMedL3	65.19	70.19	72.22	55.49	74.00	66.91	67.03
MMedL3-EnIns	68.15	64.91	71.52	59.53	76.00	72.79	68.32
LLaMA3	67.41	**76.60**	**80.56**	**67.63**	82.00	72.06	73.92
MMedL3-MI (Ours)	64.44	67.92	71.53	58.96	74.00	66.54	66.76
LLaMA3-MI (Ours)	**68.15**	75.47	75.00	**67.63**	**83.00**	**77.21**	**74.38**

Table 4: Multiple-choice accuracy evaluation on MMLU-Medicine, a subset of MMLU benchmark. The subjects used are anatomy (An), clinical knowledge (CK), college biology (CB), college medicine (CM), medical genetics (MG) and professional medicine (PM).

tasks such as Named Entity Recognition (NER), Relation Extraction (RE), and Event Extraction (EE) in the biomedical field are more susceptible to overfitting on small data sets compared to more general tasks like summarization (SUM).

4.4 Evaluation on Public English Benchmarks

The Massive Multitask Language Understanding (MMLU; Hendrycks et al., 2021) is a benchmark that evaluates language models across various QA tasks and subjects. We train **MMedL3-MI** using the same 100K dataset that was used to train **LLaMA3-MI**. The models are tested on 6 medical-related subtasks of MMLU. Table 4 exhibits the result. As seen, **LLaMA3-MI** and **MMedL3-MI** perform similarly to the baseline model on MMLU-Medicine. Additionally, note that **LLaMA3-MI** and **MMedL3-MI** are multitask models in the biomedical field, capable of handling various other, more challenging biomedical tasks.

5 Conclusion

In this paper, we introduce an instruction meta-dataset MEDINST comprising 133 biomedical tasks across 12 task categories and a challenging benchmark MEDINST32 for evaluating multitask biomedical models. Through various experiments, we train multiple biomedical models and demonstrate their strong generalization performance on biomedical tasks using our dataset. Due to resource constraints, we trained only on a small subset and 8B models. Using the full dataset and larger models may lead to further improvements, which is left for future work. Our work lays the foundation for developing better-performing biomedical LLMs.

Limitations

We identify our limitations as follows.

First, due to computational resource constraints, we conducted our experiments with limited data and model sizes. We used the LoRA technique to finetune our model, which might limit the learning outcomes. Full-parameter finetuning could potentially yield better results. In future work, we will continue to explore ways to further enhance the performance of LLMs on biomedical-related tasks.

Currently, the MedINST dataset only includes single-turn dialogues, which may limit the model's ability to generalize to multi-turn dialogue tasks. Therefore, in the future, we plan to incorporate multi-turn instruction samples.

Additionally, the current dataset is primarily in English, with other languages featured in the TRANSL tasks, so another direction for future work is to continue expanding the multilingual data.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large Language Models are Few-Shot Clinical Information Extractors. *Preprint*, arxiv:2205.12689.

Sultan Alrowili and Vijay Shanker. 2021. BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 221–227, Online. Association for Computational Linguistics.

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2016. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics (Oxford, England)*, 32(3):432–440.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the Summarization of Consumer Health Questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy. Association for Computational Linguistics.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 Shared Task on Textual Inference, Question Entailment and Question Answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379, Florence, Italy. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Preprint*, arxiv:2005.14165.

Paul Buitelaar, Thierry Declerck, Bogdan Sacaleanu, Špela Vintar, Diana Raileanu, and Claudia Crispi. 2003. A Multi-Layered, XML-Based Approach to the Integration of Linguistic and Semantic Annotations.

Qingyu Chen, Alexis Allot, Robert Leaman, Rezarta Islamaj, Jingcheng Du, Li Fang, Kai Wang, Shuo Xu, Yuefu Zhang, Parsa Bagherzadeh, Sabine Bergler, Aakash Bhatnagar, Nidhir Bhavsar, Yung-Chun Chang, Sheng-Jie Lin, Wentai Tang, Hongtong Zhang, Ilija Tavchioski, Senja Pollak, Shubo Tian, Jinfeng Zhang, Yulia Otmakhova, Antonio Jimeno Yepes, Hang Dong, Honghan Wu, Richard Dufour, Yanis Labrak, Niladri Chatterjee, Kushagri Tandon, Fréjus A. A. Laleye, Loïc Rakotoson, Emmanuele Chersoni, Jinghang Gu, Annemarie Friedrich, Subhash Chandra Pujari, Mariia Chizhikova, Naveen Sivadasan, Saipradeep Vg, and Zhiyong Lu. 2022. Multi-label classification for biomedical literature: An overview of the BioCreative VII LitCovid Track for COVID-19 literature topic annotations. *Database: The Journal of Biological Databases and Curation*, 2022:baac069.

Billy Chiu, Sampo Pyysalo, Ivan Vulić, and Anna Korhonen. 2018. Bio-SimVerb and Bio-SimLex: Wide-coverage evaluation sets of word similarity in biomedicine. *BMC Bioinformatics*, 19(1):33.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural Scaffolds for Citation Intent Classification in Scientific Publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Nigel Collier, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Jin-Dong Kim. 2004. Introduction to the Bio-entity Recognition Task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

Kersten Döring, Ammar Qaseem, Michael Becer, Jianyu Li, Pankaj Mishra, Mingjie Gao, Pascal Kirchner, Florian Sauter, Kiran K. Telukunta, Aurélien F. A. Moumbock, Philippe Thomas, and Stefan Günther. 2020. Automated recognition of functional compound-protein relationships in literature. *PloS One*, 15(3):e0220925.

Jason Alan Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Myungsun Kang, Ruisi Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen Bach, Stella Biderman, Mario Sänger, Bo Wang, Alison Callahan, Daniel León Periñán, Théo Gigant, Patrick Haller, Jenny Chim, Jose David Posada, John Michael Giorgi, Karthik Rangasai Sivaraman, Marc Pàmies, Marianna Nezhurina, Robert Martin, Michael Cullan, Moritz Freidank, Nathan Dahlberg, Shubhanshu Mishra, Shamik Bose, Nicholas Michio Broad, Yanis Labrak, Shlok S. Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, and Benjamin Beilharz. 2022. BigBIO: A Framework for Data-Centric Biomedical Natural Language Processing. *Preprint*, arxiv:2206.15076.

Martin Gerner, Goran Nenadic, and Casey M. Bergman. 2010. LINNAEUS: A species name identification system for biomedical literature. *BMC Bioinformatics*, 11(1):85.

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. *Preprint*, arxiv:2007.15779.

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. *Preprint*, arxiv:2304.08247.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. 2005. Overview of BioCreAtIvE: Critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(1):S1.

Kexin Huang, Abhishek Singh, Sitong Chen, Edward T. Moseley, Chih-ying Deng, Naomi George, and Charlotta Lindvall. 2019. Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation. *Preprint*, arxiv:1912.11975.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arxiv:2310.06825.

Hrant Khachatrian, Lilit Nersisyan, Karen Hambardzumyan, Tigran Galstyan, Anna Hakobyan, Arsen Arakelyan, Andrey Rzhetsky, and Aram Galstyan. 2019. BioRelEx 1.0: Biological Relation Extraction Benchmark. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 176–190, Florence, Italy. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTaiL: A Textual Entailment Dataset from Science Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Baeksoo Kim, Wonjun Choi, and Hyunju Lee. 2019. A corpus of plant–disease relations in the biomedical domain. *PLoS ONE*, 14(8):e0221582.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, Colorado. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking for Public Health Claims. *Preprint*, arxiv:2010.09926.

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. *Preprint*, arxiv:2402.10373.

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.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: A resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016:baw068.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge.

Nut Limsopatham and Nigel Collier. 2016. Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1014–1023, Berlin, Germany. Association for Computational Linguistics.

Boxiang Liu and Liang Huang. 2021. ParaMed: A parallel corpus for English–Chinese translation in the biomedical domain. *BMC Medical Informatics and Decision Making*, 21(1):258.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *Preprint*, arxiv:2107.13586.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. *Preprint*, arxiv:2301.13688.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A Large-scale Dataset for Extreme Multi-document Summarization of Scientific Articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online. Association for Computational Linguistics.

Clara H. McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. 2020. Effective Transfer Learning for Identifying Similar Questions: Matching User Questions to COVID-19 FAQs. *Preprint*, arxiv:2008.13546.

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date.

Sunil Mohan and Donghui Li. 2019. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. *Preprint*, arxiv:1902.09476.

Tomoko Ohta, Sampo Pyysalo, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Open-domain Anatomical Entity Mention Detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 27–36, Jeju Island, Korea. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2024. GPT-4 Technical Report. *Preprint*, arxiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arxiv:2203.02155.

Mihir Parmar, Swaroop Mishra, Mirali Purohit, Man Luo, M. Hassan Murad, and Chitta Baral. 2022. In-BoXBART: Get Instructions into Biomedical Multi-Task Learning. *Preprint*, arxiv:2204.07600.

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. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. SciFive: A text-to-text transformer model for biomedical literature. *Preprint*, arxiv:2106.03598.

Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Event extraction across multiple levels of biological organization. *Bioinformatics*, 28(18):i575–i581.

Sampo Pyysalo, Tomoko Ohta, and Jun'ichi Tsujii. 2011. Overview of the Entity Relations (REL) supporting task of BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 83–88, Portland, Oregon, USA. Association for Computational Linguistics.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Preprint*, arXiv:2402.13963.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. *Preprint*, arxiv:2110.08207.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based Fact-Checking of Health-related Claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Felipe Soares, Viviane Moreira, and Karin Becker. 2018. A Large Parallel Corpus of Full-Text Scientific Articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. 2017. BIOSSES: A semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for Fact Extraction and VERification. https://arxiv.org/abs/1803.05355v3.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arxiv:2302.13971.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie fu. 2023a. Pre-trained Language Models in Biomedical Domain: A Systematic Survey. *Preprint*, arxiv:2110.05006.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-Instruct: Aligning Language Models with Self-Generated Instructions. *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, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. *Preprint*, arxiv:2204.07705.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models Are Zero-Shot Learners. *Preprint*, arxiv:2109.01652.

Zhi Wen, Xing Han Lu, and Siva Reddy. 2020. MeDAL: Medical Abbreviation Disambiguation Dataset for Natural Language Understanding Pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 130–135, Online. Association for Computational Linguistics.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-LLaMA: Towards Building Open-source Language Models for Medicine. *Preprint*, arxiv:2304.14454.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data. *Preprint*, arxiv:2304.01196.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale Medical Dialogue Datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2024. AlpaCare:Instruction-tuned Large Language Models for Medical Application. *Preprint*, arxiv:2310.14558.

A Instruction Benchmark Construction

To comprehensively evaluate the model's performance across various medical-related tasks, we selected 32 tasks from each task category in the MEDINST dataset to establish a new biomedical benchmark, MEDINST32. The tasks selected for benchmarking encompass different levels of difficulty. This includes two aspects: knowledge difficulty and instruction difficulty. Knowledge difficulty assesses the amount of biomedical knowledge the model possesses, such as understanding categories of biomedical terms and their relationships. For basic-level assessment, we chose tasks like acronym completion (MeDAL). Intermediate-level tasks include various NER, QA, TE, and TXTCLASS tasks. Finally, we included more challenging tasks like RE, EE, and tasks in NED that involve annotating identifiers. Instruction difficulty evaluates the model's understanding and adherence to instructions. This dimension was not considered in previous benchmark datasets. For example, in multichoice QA tasks, previous works often labeled each option as A, B, C, etc., and the model only needed to respond with the corresponding label. In our QA task construction, we require the model to output the selected option as it is, which increases the task difficulty and reduces the chance of the model bypassing with simple letter responses. Additionally, we construct different instructions for similar tasks. For instance, in NER tasks, we developed two types of instructions: one requiring the model to repeat each word in the text in a BIO format and label them one by one, and the other asking the model to directly extract all biomedical entity mentions and annotate their categories.

For each task in MEDINST32, we provide two positive examples. For tasks that have a training set, we select two examples from their training set. If a task does not have a training set, we find the most similar task from all the test set tasks in MEDINST and select from there. During selection, we strive to ensure that the two examples are diverse in content. For instance, in classification tasks, we choose examples with different labels.

We remove all the datasets used in MEDINST32 from the MEDINST training set to create the training set for MEDINST32.

We performed random sampling on a portion of tasks with abundant data resources to control the number of test data in each category to be roughly consistent. This helps to reduce the computational resource consumption for evaluation. The sample sizes are shown in Table 5. For other datasets, we use the entire test set data.

Dataset Name	Sample Size
NCBI-disease	100
BC5CDR	100
BioNLP-2011-GE	100
tmVar-v3	100
MeDAL	1000
ParaMed	200
Multi-XScience	200

Table 5: Sampling sizes for evaluation.

Overall, we provide a more comprehensive and challenging biomedical instruction benchmark compared to previous works.

B Instruction Examples

Table B presents the instruction examples for each task categories. Each instruction contains three parts: input explanation, task definition, and output format, which clearly tell the LLM how to complete the task. For each task within a category, the instruction can vary, thus requiring manual composition. However, for categories such as NED, RE, and EE tasks, the main body of the instruction is generic. We can efficiently edit the instruction by modifying some variable fields based on the metadata of each dataset, and these variable fields are highlighted in blue.

C Implementation Details

Training For the baseline models, we used the LLaMA-3-8B-Instruct [4] and MMed-LLaMA-3-8B [5] models available on Hugging Face. Due to limited computational resources, we employed Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning (PEFT). The LoRA rank was set to 8, targeting all linear layers, including q_proj, k_proj, v_proj, o_proj, $gate\_proj$, up_proj, and $down\_proj$. The learning rate was set at 1.0e-4, with a batch size of 4 and gradient accumulation steps of 4. We used a cosine learning rate scheduler with a 0.1 ratio of warmup. For training, we ran 5 epochs with 5K data, and 3 epochs for 50K and

[4] https://huggingface.co/unsloth/llama-3-8b-Instruct

[5] https://huggingface.co/Henrychur/MMed-Llama-3-8B

QA Given a question and context, select the correct answer from the provided options.
TE Given a pair of texts, consisting of a claim and the evidence, determine whether the evidence supports, refutes, or is neutral regarding the claim. Respond with one of the following: 'Supports', 'Refutes', or 'Neutral'.
NER Given a sentence, label each disease, disease class and symptom entity using the BIO format. In BIO format, 'B' indicates the beginning of an entity, 'I' indicates the inside of an entity, and 'O' indicates a token not part of any entity. Label each word in the format: 'word [LABEL]'.
TXTCLASS You are provided with a citation context. Classify the intent of the citation within this context. Intents are: [background, method, result].
NED You are provided with a text. Your objective is to identify and extract all chemical and disease entities mentioned in the text, maintaining the order in which they appear. For each entity, provide its corresponding database identifier from MESH. The entities should be presented in the format: [entity1 <db_name/db_id>].
RE Given a text, identify and extract specified relations between anatomical entities mentioned within it. The specified relation types are [frag, Part-of]. Relation explanation: frag: Frag relation marking coordination with ellipsis; Part-of: Part-of relation marking entity mention spanning a prepositional phrase. Present each relation in format as follows: [<entity1> <relation> <entity2>].
COREF Given a text and a specified anatomical entity, identify and extract all co-references to that entity within the text. Present each co-reference entity in the following format: [co-reference entity].
STS Given two texts, evaluate their similarity and provide an integer score ranging from 0 to 5, where 0 indicates no similarity and 5 indicates high similarity.
EE Given a text, identify and extract the epecified types of bio-molecular events along with their primary arguments. The event type can be [Binding, Positive_regulation, Phosphorylation, Regulation, Transcription, Localization, Gene_expression, Protein_catabolism, Negative_regulation]. Present each event in the format as follows: [<type> <trigger> <theme entity>].
TRANSL Translate the text from Chinese to English.
TEXTPAIRCLASS You are given a drug name and a piece of text. Analyze the sentiment in the text and determine whether the sentiment towards the drug is positive, negative, or neutral. Answer with 'Positive', 'Negative', or 'Neutral'.
SUM Writing the related-work section of a paper based on its abstract and the articles it references.

Table 6: Instruction examples for each task category.

100K datasets. The training was conducted on a single 40GB A100 GPU.

Query Template For the training and evaluation of all LLaMA-3 series models, we used the standard LLaMA-3 chat template. Table 7 shows an example. When constructing few-shot prompts, each example is treated as a round of dialogue and added before the query that needs an answer. Unlike the approach where instructions are only given in the first round of dialogue, we included instructions in each example. This is because for some tasks without a training set, we selected examples from the training sets of similar tasks, so the instructions in the examples may not completely match the instructions of the query. Table 8 demonstrates a query of a NER task.

D Extra Metrics for SUM and TRANSL Tasks

We add additional metrics, BERT score and METEOR score, to evaluate the generated text on summarization and translation tasks. The evaluation results are presented in Table 9 and Table 10.

E Dataset Collection

Table 11 lists all the dataset employed in MEDINST. Because a single dataset might be reformulated into multiple tasks, we added suffixes to the names in the multi-task dataset. For example, BC5CDR appears in the NER, NED, and RE tasks. For the primary task, NER, we use the dataset's original name, and for the other two tasks, we append the respective suffixes to the dataset name.

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are a helpful assistant.<|eot_id|><|start_header_id|>user<|end_header_id|>

Given an utterance, determine if it is from a doctor or a patient. Do i have covid 19?<|eot_id|><|start_header_id|>assistant<|end_header_id|>

patient<|eot_id|>

Table 7: LLaMA-3 prompt template.

Example 1	**Instrcution:**	You are provided with a text. Your objective is to identify, extract and classify all gene and protein entities mentioned in the text, maintaining the order in which they appear. Types are [Gene, DomainMotif, FamilyName]. The entities should be presented in the following format: [entity <type>].
	Input:	Cloning, expression and localization of an RNA helicase gene from a human lymphoid cell ... ... cell line from a diffuse large B-cell lymphoma.
	Output:	[RNA helicase <FamilyName>] [RNA helicase <FamilyName>] [p54 <Gene>] [RNA helicase <FamilyName>] [ME31B <Gene>] [ME31B <Gene>]
Example 2	**Instrcution:**	You are provided with a text. Your objective is to identify, extract and classify all gene variant entities mentioned in the text, maintaining the order in which they appear. Types are [DNAMutation, SNP, ProteinMutation]. The entities should be presented in the following format: [entity <type>].
	Input:	A novel multidrug-resistance protein 2 gene mutation identifies a ... ... heterozygous mutation was significantly associated with the presence of pruritus.
	Output:	[V1188E <ProteinMutation>]
Query	**Instrcution:**	You are provided with a text. Your objective is to identify, extract and classify all gene variant entities mentioned in the text, maintaining the order in which they appear. Types are [OtherMutation, Species, DNAAllele, DNAMutation, CellLine, SNP, ProteinMutation, ProteinAllele, Gene, AcidChange]. The entities should be presented in the following format: [entity <type>].
	Input:	A novel single-nucleotide substitution, Glu 4 Lys ... ... Thus, our results suggest that Glu 4 Lys in the LTC4S might be associated with allergic diseases.

Table 8: Query example.

Model	BERTScore	METEOR Score
LLaMA3	0.7467	0.1758
BioMistral	0.7253	0.1152
MMEDL3-EnIns	0.7314	0.1185
GPT-4o	0.8317	0.2333
LLaMA3-MI32 (ours)	0.7951	0.1566
MMEDL3-MI32 (ours)	0.7963	0.1220
LLaMA3-MI (ours)	0.8203	0.1592

Table 9: SUM task: Multi-XScience results.

Model	BERTScore	METEOR Score
LLaMA3	0.9000	0.3776
BioMistral	0.9101	0.3670
MMEDL3-EnIns	0.8888	0.3625
GPT-4o	0.9291	0.4661
LLaMA3-MI32 (ours)	0.9115	0.3933
MMEDL3-MI32 (ours)	0.9080	0.3781
LLaMA3-MI (ours)	0.9379	0.6126

Table 10: TRANSL task: ParaMed results.

Table 11: Dataset collection.

Dataset	Task	Train	Dev	Test
BioASQ-Task-B-yesno	QA	15,568	0	813
BioASQ-Task-B-list	QA	11,687	0	1,000
BioASQ-Task-B-factoid	QA	16,389	0	724
BioASQ-Task-B-summary	QA	13,151	0	824
BiologyHowWhyCorpus	QA	1,269	0	0
BIOMRC	QA	700,000	50,000	62,707
Evidence-Inference-2.0	QA	10,056	1,233	1,222
MedQA	QA	10,178	1,273	1,272
MedHop	QA	1,620	342	0
MEDIQA-QA	QA	312	25	150
PubMedQA-artificial	QA	200,000	11,269	0
PubMedQA-labeled	QA	450	50	500
SciQ	QA	11,679	1,000	1,000
FEVER	TE	145,449	9,999	9,999
HealthVer	TE	10,590	1,917	1,823
PubHealth	TE	9,804	1,214	1,233
SciFact	TE	868	0	1,189
ManConCorpus	TE	0	0	2,775
CoVERt	TE	0	0	212
MEDIQA-RQE	TE	8,588	302	230
SciTail	TE	23,596	2,126	1,304
NCBI-disease	NER	5,432	923	942
BC2GM	NER	12,632	2,531	5,065
CHEMDNER-BIO	NER	30,884	30,841	26,561
BC5CDR	NER	4,560	4,581	4,797
Linnaeus	NER	12,004	4,086	7,181
JNLPBA-DNA	NER	4,699	552	622
JNLPBA-RNA	NER	721	89	102
JNLPBA-CT	NER	4,792	420	1,422
JNLPBA-CL	NER	2,596	284	377
AnatEM	NER	5,861	2,118	3,830
AnEM	NER	164	137	30
BioInfer	NER	894	0	206
BioNLP-2009	NER	756	260	150
BioNLP-2011-EPI	NER	600	200	0
BioNLP-2011-GE	NER	856	0	338
BioNLP-2011-ID	NER	151	46	117
BioNLP-2011-REL	NER	756	150	260
BioNLP-2013-CG	NER	300	100	200
BioNLP-2013-GE	NER	194	212	256
BioNLP-2013-GRO	NER	150	50	100
BioNLP-2013-PC	NER	260	90	175
BioNLP-2019-BB	NER	132	66	0
BioRED	NER	400	100	100
BioRelEx	NER	1,402	201	0
CellFinder	NER	5	0	5
CHEBI	NER	476	0	0
CHEMDNER	NER	2,915	2,906	2,477

Continued on next page

Table 11 – *Continued from previous page*

Dataset	Task	Train	Dev	Test
ChemProt	NER	1,020	612	800
CHIA	NER	1,932	0	0
CPI	NER	1,808	0	0
DDI	NER	673	0	279
DrugProt	NER	3,500	750	0
EBM-NLP	NER	4,735	0	187
EU-ADR	NER	299	0	0
GENETAG	NER	3,875	1,311	2,567
PTM-Events	NER	112	0	0
GENIA-Term	NER	2,000	0	0
GNormPlus	NER	418	0	261
HPRD50	NER	34	0	9
MedMentions	NER	2,635	878	879
miRNA	NER	201	0	100
MLEE	NER	130	44	87
NLM-Gene	NER	450	0	100
NLM-Chem	NER	80	20	50
OSIRIS	NER	105	0	0
PDR	NER	179	0	0
PICO-Annotation	NER	361	0	0
ProGene	NER	20,055	1,109	2,414
SCAI-Chemical	NER	67	0	0
SCAI-Disease	NER	330	0	0
SETH	NER	433	0	0
SPL-ADR	NER	101	0	0
tmVar-v1	NER	213	0	101
tmVar-v2	NER	158	0	0
tmVar-v3	NER	0	0	493
Verspoor-2013	NER	117	0	0
MedDialog	TXTCLASS	981	126	122
SciCite	TXTCLASS	8,243	916	1,861
Hallmarks-of-Cancer	TXTCLASS	12,119	1,798	3,547
GEOKhoj-v1	TXTCLASS	25,000	0	5,000
BC7-LitCovid	TXTCLASS	24,960	2,500	6,239
AskAPatient-NED	NED	15,612	845	867
BC5CDR-NED	NED	500	500	500
Bio-ID	NED	11,366	0	0
BioNLP-2019-BB-NED	NED	132	66	0
BioRED-NED	NED	400	100	100
BioRelEx-NED	NED	1,402	201	0
CPI-NED	NED	1,808	0	0
GNormPlus-NED	NED	418	0	261
Linnaeus-NED	NED	95	0	0
MeDAL	NED	3,000,000	1,000,000	1,000,000
MedMentions-NED	NED	2,635	878	879
miRNA-NED	NED	201	0	100
MuchMore-NED	NED	7,820	0	0
NCBI-disease-NED	NED	592	100	100
NLM-Gene-NED	NED	450	0	100

Continued on next page

Table 11 – *Continued from previous page*

Dataset	Task	Train	Dev	Test
NLM-Chem-NED	NED	80	20	50
OSIRIS-NED	NED	105	0	0
SPL-ADR-NED	NED	101	0	0
tmVar-v2-NED	NED	158	0	0
tmVar-v3-NED	NED	0	0	493
TwADR-L-NED	NED	4,816	115	143
AnEM-RE	RE	22	5	13
BC5CDR-RE	RE	500	500	500
BioInfer-RE	RE	642	0	142
BioNLP-2011-REL-RE	RE	378	92	0
BioNLP-2013-GE-RE	RE	40	41	0
BioNLP-2013-GRO-RE	RE	149	48	0
BioNLP-2019-BB-RE	RE	121	59	0
BioRED-RE	RE	395	97	100
BioRelEx-RE	RE	1,263	178	0
CHEBI-RE	RE	415	0	0
ChemProt-RE	RE	767	443	620
CHIA-RE	RE	1,876	0	0
CPI-RE	RE	1,246	0	0
DDI-RE	RE	510	0	191
DrugProt-RE	RE	2,433	542	0
EU-ADR-RE	RE	253	0	0
HPRD50-RE	RE	28	0	8
IEPA	RE	114	0	26
LLL05	RE	77	0	0
MLEE-RE	RE	32	11	16
MuchMore-RE	RE	7,734	0	0
SETH-RE	RE	212	0	0
SPL-ADR-RE	RE	96	0	0
Verspoor-2013-RE	RE	114	0	0
AnEM-COREF	COREF	10	2	14
BioNLP-2009-COREF	COREF	536	110	0
BioNLP-2011-EPI-COREF	COREF	440	168	0
BioNLP-2011-GE-COREF	COREF	571	0	0
BioNLP-2011-ID-COREF	COREF	170	31	0
BioNLP-2011-REL-COREF	COREF	535	110	0
BioNLP-2013-CG-COREF	COREF	466	176	0
BioNLP-2013-GE-COREF	COREF	53	41	0
BioNLP-2013-PC-COREF	COREF	455	128	0
BioRelEx-COREF	COREF	1,143	167	0
PTM-Events-COREF	COREF	25	0	0
MLEE-COREF	COREF	198	57	113
PDR-COREF	COREF	19	0	0
Bio-SimVerb	STS	1,000	0	0
Bio-SimLex	STS	988	0	0
BIOSSES	STS	64	16	20
EHR-Rel	STS	3,741	0	0
MayoSRS	STS	101	0	0
MQP	STS	3,048	0	0

Continued on next page

Table 11 – *Continued from previous page*

Dataset	Task	Train	Dev	Test
UMNSRS	STS	1,153	0	0
BioNLP-2009-EE	EE	695	150	0
BioNLP-2011-EPI-EE	EE	383	121	0
BioNLP-2011-GE-EE	EE	765	0	0
BioNLP-2011-ID-EE	EE	110	30	0
BioNLP-2013-CG-EE	EE	299	100	0
BioNLP-2013-GE-EE	EE	149	157	0
BioNLP-2013-PC-EE	EE	257	90	0
PTM-Events-EE	EE	111	0	0
MLEE-EE	EE	127	44	87
PDR-EE	EE	167	0	0
MuchMore-TRANSL	TRANSL	6,374	0	0
ParaMed	TRANSL	62,127	2,036	2,102
SciELO	TRANSL	3,006,699	0	0
Medical-Data	TEXTPAIRCLASS	5,279	0	0
MeQSum	SUM	1,000	0	0
Multi-XScience	SUM	30,369	5,066	5,093