Towards Better Multi-task Learning: A Framework for Optimizing Dataset Combinations in Large Language Models

Zaifu Zhan, Rui Zhang
University of Minnesota Twin Cities
{zhan8023,ruizhang}@umn.edu

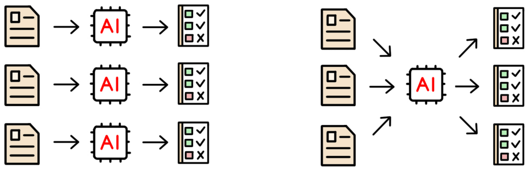
Abstract

To efficiently select optimal dataset combinations for enhancing multi-task learning (MTL) performance in large language models, we proposed a novel framework that leverages a neural network to predict the best dataset combinations. The framework iteratively refines the selection, greatly improving efficiency, while being model-, dataset-, and domain-independent. Through experiments on 12 biomedical datasets across four tasks—named entity recognition, relation extraction, event extraction, and text classification—we demonstrate that our approach effectively identifies better combinations, even for tasks that may seem unpromising from a human perspective. This verifies that our framework provides a promising solution for maximizing MTL potential.

1 Introduction

Natural Language Processing (NLP) has made significant strides in recent years (Liu et al., 2023), evolving from *fully supervised learning* (Kotsiantis et al., 2007), to *feature engineering* (Patil et al., 2023), *architecture innovations* like Transformer (Vaswani et al., 2017), and the dominance of *pre-trained large models* such as BERT and GPT (Devlin et al., 2018; Radford et al., 2018, 2019). More recently, the *instruction-tuning* (Zhang et al., 2023) and *prompting engineering* (Liu et al., 2023) have emerged, allowing Large Language Models (LLMs) to handle tasks effectively through prompting (Wei et al., 2022).

With the rapid advancements in NLP, Multi-Task Learning (MTL) has emerged as a powerful technique to boost model performance by jointly training on multiple related tasks (Zhang and Yang, 2018, 2021; Zhan et al., 2025), as illustrated in Fig. 1. By sharing knowledge across tasks, MTL enhances model generalization (Wang et al., 2021) and efficiently captures the complementary relationships between tasks (Ma et al., 2018). For instance, the Named Entity Recognition (NER) task and the Relation Extraction (RE) task are closely linked—accurate entity recognition can provide a critical context for extracting relationships, while relation extraction, in turn, can refine entity recognition.



(a) Single-task learning (b) Multi-task learning

Figure 1: Comparison of (a) single-task learning and (b) multi-task learning in large language models.

As models become larger, the amount of data used for training has also significantly increased (Chen et al., 2024), giving rise to models capable of understanding and solving problems across various domains. For instance, ChatGPT can generate realistic and creative outputs across various domains (Yenduri et al., 2024; Zhou et al., 2024). Previously, achieving MTL required modifying the model architecture and adjusting the output layers for different tasks (Misra et al., 2016). Now, by simply modifying the prompt, MTL has found greater utility and flexibility without redesigning the architecture (Liu et al., 2023; Li et al., 2024). This indicates that large models have truly become tools accessible to everyone. As long as users know how to utilize frameworks like Hugging Face (Wolf et al., 2020), they can fine-tune models with their own prompts without needing any knowledge of Attention mechanisms (Vaswani et al., 2017) or model architecture.

To take advantage of the natural compatibility of MTL and LLMs, many recent large models have improved performance by incorporating multi-task training. Models like DeepStruct (Wang et al.,

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 5388–5401
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

2022a), InstructUIE (Wang et al., 2023), and Code4Struct (Wang et al., 2022b) have demonstrated the power of multi-task learning by training on diverse datasets across tasks like NER, RE, Event Extraction (EE), and Slot Filling (SF), etc. In addition, GoLLIE (Sainz et al., 2023) used datasets spanning both biomedical and news domains, achieving state-of-the-art results in these fields. ADELIE (Qi et al., 2024) further emphasized that aligning LLMs with multiple Information Extraction (IE) tasks significantly improves task performance in the biomedical domain.

Despite the encouraging results achieved by many multi-task LLMs, we still lack a clear understanding of how to effectively select datasets for training. Most successful outcomes rely heavily on trial and error, subjective judgment, or experience. This reveals a significant question: how to effectively select dataset combinations to enhance performance? If we have one dataset and four related auxiliary datasets, there are $2^4 = 16$ possible combinations, which is feasible to enumerate. However, when the number of auxiliary datasets increases to 10 or more, it becomes practically impossible to train and evaluate all possible combinations to find the optimal one.

To take a step toward filling the gap, this paper proposes a framework to identify a good combination of datasets for improving model performance. The framework is dataset-independent and model-independent, so the framework could be applied to any domain, datasets, and to any LLMs. The full description of the proposed framework is in Section 3. The main contributions of this paper are as follows:

- We propose a new MTL framework designed specifically to effectively optimize the selection of datasets to release the potential of MTL-LLMs.
- Based on this framework, we conduct a comprehensive evaluation of 12 biomedical datasets across four tasks: NER, EE, RE, and TC. The results show that the performance of LLMs on datasets could be increased by finding better combinations using our framework.

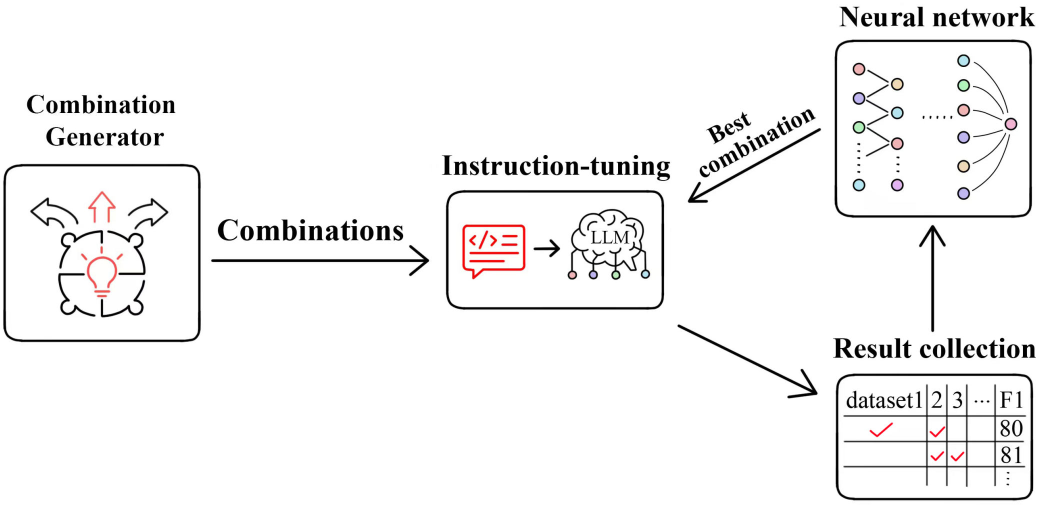
2 Related Work

There have been several attempts to explore how to select dataset combinations to improve MTL (Zhang and Yang, 2018, 2021; Thung and Wee, 2018; Sener and Koltun, 2018; Crawshaw, 2020). For instance, Bingel and Søgaard (2017) systematically investigated and found the MTL gains are related to the characteristics and features of datasets. However, what kind of characteristics and features we should consider for different domains are unknown, so the analysis is mostly based on experience. Also, Standley et al. (2020) tried to figure out which tasks should be learned together by considering the space of all possible task subsets, training networks for each subset, and then using each network's performance to choose the best combination. It is straightforward, but training all possible combinations is impossible when we consider many tasks. Pruksachatkun et al. (2020) perform a large-scale study on the pre-trained RoBERTa model with 110 intermediate–target task combinations, and then evaluate all trained models with 25 probing tasks. However, they failed to observe more granular correlations between probing and target task performance. Guo et al. (2019) proposed AUTOSEM which uses a multi-armed bandit controller to find appropriate auxiliary tasks but each arm only considers the relation between 2 tasks and doesn't consider the interaction of two or more auxiliary tasks. Fifty et al. (2021) proposed to measure inter-task affinity by training all tasks together in a single multi-task network and to quantify the effect to which one task gradient update would affect another task loss. It effectively computes task groupings from only a single training run but we know the gradient update depends on the loss function and the initial point. For non-convex loss functions, there may be many local optima, so the grouping results from this method are various.

By contrast, our proposed framework avoids unreliable human experience and low effective brute force, using a simple neural network to find good dataset combinations based on some combination-score pair data.

3 Framework

The most straightforward approach to evaluate the effect of all possible dataset combinations on a given LLM would be to directly do experiments for each combination via fine-tuning and inference of the corresponding LLM. However, this naive method is computationally expensive and requires substantial resources. By contrast, multi-layer neural networks are much faster to compute. If a fast



dataset1	2	3	···	F1
✓	✓			80
	✓	✓		81
				⋮

Figure 2: **Overview of the proposed framework.** The combination generator produces dataset combinations to be instruction-tuned on an LLM. In each iteration, the neural network identifies and refines the best combination until no further improvements can be made.

neural network could effectively filter out combinations that are unlikely to yield good results, or directly predict the best combination, it would save significant time and computational power. Driven by this idea, we propose a new framework to optimize dataset selection for MTL.

As shown in Fig. 2, our proposed framework consists of four parts. First, we generate sufficient combinations of datasets to at least cover all datasets. Next, fine-tuning LLMs on these dataset combinations. After inference for each combination, we record all combinations and their corresponding performance scores in a table. Using this table data, we then train a neural network to predict the best dataset combination by enumerating all possible combinations through the neural network, which could predict the best combination in a super-efficient way. The framework iteratively trains the large model with the predicted combinations and tests the results. The process continues until the neural network predicts an optimal combination that has already been tested, at which point the loop terminates.

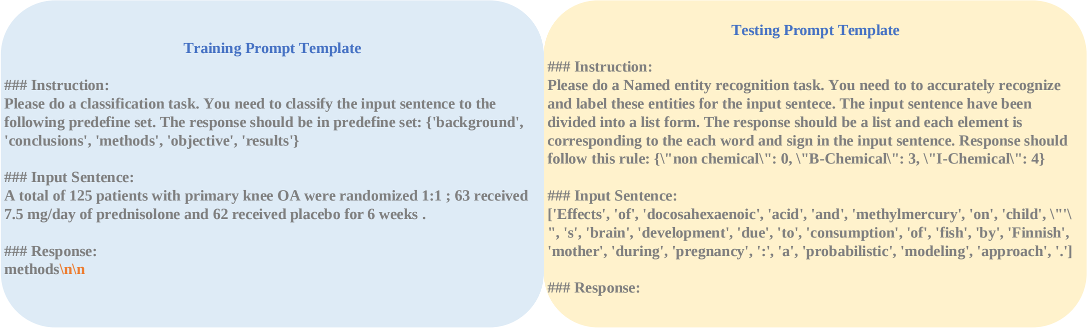
The proposed framework offers several advantages. First, it saves time by using a neural network to infer relationships between datasets, allowing us to focus on testing the most promising combinations based on existing data. Second, it is highly flexible, applicable across different models and dataset pools, and adaptable to various neural network architectures for predicting the optimal combination. Lastly, it is robust. Even if the initial selections are limited or random, and the neural network struggles to predict good combinations due to insufficient data, the iterative process ensures that as more data is gathered, the system will eventually find a good (or even optimal) combination.

The inspiration for this framework comes from the concept of feedback in control systems (Doyle et al., 2013). In a data-driven system, finding the optimal controller requires first accumulating a certain amount of data and then trying to control the systems using the controller trained by the accumulated data (Kiumarsi et al., 2014). With each attempt, more feedback is gathered, allowing the system to refine the controller. The better the controller becomes, the closer it gets to finding the optimal solution. Similarly, in our framework, we begin by generating various dataset combinations and get the corresponding performance score for each combination by inference. The neural network, much like a feedback-driven controller, gives lower scores to poor combinations and higher scores to promising ones. As the process continues, the neural network encourages the exploration of good combinations and discourages poor ones. With more iterations, the system converges towards identifying the optimal dataset combination.

4 Experiment setup

4.1 Dataset

In this paper, we focus on four critical NLP tasks: EE, RE, NER, and TC. For each of these tasks, we



Training Prompt Template

```
### Instruction:
Please do a classification task. You need to classify the input sentence to the
following predefine set. The response should be in predefine set: {'background',
'conclusions', 'methods', 'objective', 'results'}

### Input Sentence:
A total of 125 patients with primary knee OA were randomized 1:1 ; 63 received
7.5 mg/day of prednisolone and 62 received placebo for 6 weeks .

### Response:
methods\n\n
```

Testing Prompt Template

```
### Instruction:
Please do a Named entity recognition task. You need to to accurately recognize
and label these entities for the input sentece. The input sentence have been
divided into a list form. The response should be a list and each element is
corresponding to the each word and sign in the input sentence. Response should
follow this rule: {\"non chemical\": 0, \"B-Chemical\": 3, \"I-Chemical\": 4}

### Input Sentence:
['Effects', 'of', 'docosahexaenoic', 'acid', 'and', 'methylmercury', 'on', 'child', \""\
", 's', 'brain', 'development', 'due', 'to', 'consumption', 'of', 'fish', 'by', 'Finnish',
'mother', 'during', 'pregnancy', ':', 'a', 'probabilistic', 'modeling', 'approach', '.']

### Response:
```

Figure 3: Training and testing prompt template.

selected three datasets to evaluate the potential of our framework comprehensively.

For **Event Extraction** task, we utilize the following datasets:

- **PHEE** (Sun et al., 2022): A comprehensive dataset providing high-accuracy annotations for 2 events: 'Adverse_event', 'Potential_therapeutic_event'.

- **GENIA2011** (Kim et al., 2011): A benchmark EE dataset from the biomedical domain. The event types include 'Regulation', 'Localization', 'Transcription', 'Binding', 'Gene_expression', 'Positive_regulation', 'Protein_catabolism', 'Negative_regulation', 'Phosphorylation' and 'NA', totally 10 event types.

- **GENIA2013** (Kim et al., 2013): An updated version of the genia2011 dataset, with additional and refined annotations reflecting the latest biomedical research. The event type include 'Regulation', 'Negative_regulation', 'Protein_modification', 'Positive_regulation', 'Localization', 'Phosphorylation', 'Ubiquitination', 'Protein_catabolism', 'Gene_expression', 'Binding', 'Transcription' and 'NA, totally 12 event types.

For **Relation Extraction** task, we employ:

- **DDI** (Segura-Bedmar et al., 2013): The Drug-Drug Interaction dataset, crucial for identifying interactions between different medications. The relation types include 'advise', 'effect', 'int', 'mechanism', and 'NA', totally 5 relation types.

- **GIT** (Li et al., 2023): A dataset focused on general interactions between entities, aiding in the development of versatile relation extraction models. There are 22 relation types: 'PREVENTS', 'TREATS', 'DOES_NOT_TREAT', 'ASSOCIATED_WITH', 'CAUSES', 'DIAGNOSES', 'MANIFESTATION_OF', 'USES', 'STIMULATES', 'INHIBITS', 'DISRUPTS', 'INTERACTS_WITH', 'PRODUCES', 'ADMINISTERED_TO', 'COEXISTS_WITH', 'AFFECTS', 'PROCESS_OF', 'COMPLICATES', 'AUGMENTS', 'PRECEDES', 'SYMPTOM_OF', 'PREDISPOSES'

- **BioRED** (Luo et al., 2022): The Biological Relation Extraction Dataset, which includes detailed annotations of biological interactions within scientific texts. The relation types include 'Positive_Correlation', 'Negative_Correlation', 'Conversion', 'Drug_Interaction', 'Cotreatment', 'Comparison', 'Association', 'Bind', totally 8 relation types.

For **Named Entity Recognition** task, our chosen datasets are:

- **BC5CDR** (Li et al., 2016): This dataset contains annotated mentions of chemicals and diseases in biomedical literature. Named entities include 'B-Chemical', 'B-Disease', 'I-Disease', 'I-Chemical' and 'else', which are mapped to 1,2,3,4,0 respectively in response.

- **BC2GM** (Smith et al., 2008): The BioCreative II Gene Mention dataset, used for recognizing gene names. In response, gene names are mapped to 1 and others to 0.

- **BC4CHEMD** (Krallinger et al., 2015): The BioCreative IV Chemical and Drug dataset, providing comprehensive annotations for chemical and drug names. The model need to map ’B-chemical’ to 3, ’I-chemical’ to 4, ’others’ to 0.

For **Text Classification** task, we utilize:

- **ADE** (Gurulingappa et al., 2012): The Adverse Drug Events dataset could be used for binary classification task. The goal is to determine if the input sentence is ADE-related or not.

- **PubMed20krct** (Dernoncourt and Lee, 2017): A large dataset of biomedical literature from PubMed. The task requires the model to classify the sentence into 5 categories: ’background’, ’conclusions’, ’methods’, ’objective’, ’results’.

- **HealthAdvice** (Yu et al., 2019): A dataset containing health-related information and advice. The model need to classify the sentence to three categories: ’no advice’, ’week advice’, ’strong advice’.

4.2 Model

The specific model used in this paper is not the focus of our paper, as the proposed framework is designed to be applicable to any LLMs. Therefore, we opted for a representative open-source model: LLama3-8B [1]. It was selected due to its widespread use and strong performance across various NLP tasks, making it an ideal candidate for demonstrating the effectiveness of our dataset selection framework. Additionally, its open-source nature ensures reproducibility and allows for further experimentation by the research community.

4.3 Data preparation

To begin, it is necessary to generate several dataset combinations as basic data (including the single task learning as baseline), which will later allow the neural network to learn the relationships between datasets. For the 12 datasets we selected, we run combinations both within the same task and across different tasks. These combinations ensure that all datasets are involved, providing a diverse and comprehensive basis for understanding how different datasets interact and contribute to overall performance.

4.4 Neural Network

The choice of neural network architecture is flexible. In our experiment, we used a two-layer neural network with 12 inputs and 1 output to perform a regression task. The 12 inputs correspond to the 12 datasets: if a dataset is used, the corresponding input is set to 1; if it is not used, the input is set to 0. The output of the neural network is the score we aim to predict, such as the F1 score in our experiments. This design directly establishes a relationship between the use of specific datasets and the resulting F1 score.

Each time a new combination of datasets is trained and tested, the neural network is trained again. After training, we input all possible combinations and use the predicted F1 scores to determine the best combination for the next iteration and keep optimizing model performance.

4.5 Instruction-tuning

The prompt examples we used are shown in Fig. 3. The training prompt example includes instruction, an input sentence, and its corresponding response. By contrast, the testing prompt is empty in response, which allows the model to continue to generate the next tokens. \n\n is the ending mark for completing the response generation. Then in the evaluation stage, we will extract output tokens before the ending token.

Since the size of each dataset varies, to be fair, we instruction-tuned each model for 5000 steps, no matter how many tasks we included. We saved the model every 1000 steps and used the best model for the latter generation. When using multiple datasets, we sampled evenly from each one to make sure they were equally weighted.

4.6 Metrics

In the evaluation stage, we used Micro Precision, Recall, and F1-score. When using these metrics, a prediction is considered correct only if the entire predicted output exactly matches the ground truth.

5 Results

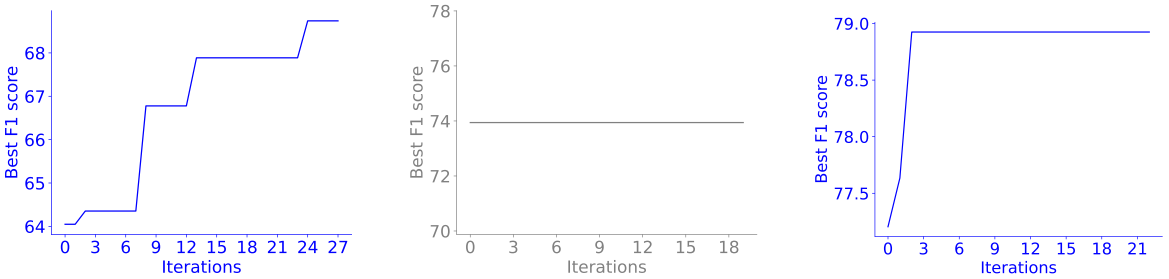
5.1 Data preparation for RE task

We included all attempts related to the RE task in Table 1. As shown, the combinations involving BioRED, DDI demonstrate performance improvements for some specific combinations, while for the remaining GIT datasets, the baseline achieved the highest F1 score. Specifically, for the BioRED

[1] https://github.com/meta-llama/llama3

Table 1: Data preparation for RE task, including BioRED, DDI, and GIT datasets. ★ indicates the best F1 score and best combination in the data preparation phase. Each line represents one experiment and √ means the dataset is selected for this run.

Training sets																
RE task			NER task			EE task			TC task					Metrics		
BioRED	DDI	GIT	BC2GM	BC4CHEMD	BC5CDR	GENIA2011	GENIA2013	PHEE	ADE	HealthAdvice	PubMed20krct	Task	Test Set	Precision	Recall	F1 Score
√												Baseline	BioRED	37.11	95.95	53.52
√		√										RE	BioRED	36.23	97.75	52.86 ↓
√	√											RE	BioRED	37.77	95.95	54.20 ↑
√									√			RE	BioRED	37.45	91.44	53.14 ↓
√						√						RE	BioRED	38.57	97.30	55.24 ↑
√			√									RE	BioRED	40.65	95.95	57.10 ↑
√	√	√										RE	BioRED	40.22	97.30	56.92 ↑
√			√			√						RE	BioRED	37.99	96.85	54.57 ↑
√						√			√			RE	BioRED	46.84	90.09	61.63 ↑
√			√						√			RE	BioRED	41.62	97.30	58.30 ↑
√			√			√			√			RE	BioRED	42.69	97.30	59.34 ↑
√			√			√				√		RE	BioRED	43.93	94.59	60.00 ↑
√			√			√					√	RE	BioRED	41.73	95.50	58.08 ↑
√				√		√			√			RE	BioRED	42.15	95.50	58.48 ↑
√				√		√				√		RE	BioRED	44.69	92.79	60.32 ↑
√				√		√					√	RE	BioRED	41.15	96.40	57.68 ↑
√			√					√	√			RE	BioRED	40.34	95.95	56.80 ↑
√					√	√			√			RE	BioRED	42.97	97.75	59.70 ↑
√					√	√					√	RE	BioRED	38.92	97.30	55.60 ↑
√					√	√				√		RE	BioRED	48.18	95.50	64.05 ★
√			√					√		√		RE	BioRED	40.26	96.85	56.88 ↑
√				√				√	√			RE	BioRED	39.85	96.40	56.39 ↑
√				√				√		√		RE	BioRED	35.16	98.20	51.78 ↓
√				√				√			√	RE	BioRED	42.48	95.50	58.81 ↑
√					√			√	√			RE	BioRED	39.63	96.40	56.17 ↑
√			√					√			√	RE	BioRED	44.61	95.05	60.72 ↑
√					√			√		√		RE	BioRED	40.80	96.85	57.41 ↑
√					√			√			√	RE	BioRED	46.67	94.59	62.50 ↑
√			√				√		√			RE	BioRED	44.57	86.94	58.93 ↑
√			√				√			√		RE	BioRED	39.23	96.85	55.84 ↑
√				√			√		√			RE	BioRED	37.16	91.89	52.92 ↓
√				√			√			√		RE	BioRED	33.65	96.40	49.88 ↓
√				√			√				√	RE	BioRED	39.17	93.69	55.25 ↑
√					√		√		√			RE	BioRED	38.64	97.30	55.31 ↑
√			√				√				√	RE	BioRED	38.60	96.85	55.20 ↑
√					√		√			√		RE	BioRED	34.76	98.65	51.41 ↓
√					√		√				√	RE	BioRED	39.17	97.75	55.93 ↑
	√											Baseline	DDI	71.10	71.10	71.10
	√	√										RE	DDI	72.60	72.60	72.60 ↑
√	√											RE	DDI	38.98	67.90	49.53 ↓
	√									√		RE	DDI	73.20	73.20	73.20 ↑
	√							√				RE	DDI	48.30	73.90	58.42 ↓
	√			√								RE	DDI	72.40	72.40	72.40 ↑
√	√	√										RE	DDI	64.21	65.30	64.75 ↓
	√			√				√				RE	DDI	61.70	67.50	64.47 ↓
	√							√		√		RE	DDI	61.52	68.10	64.64 ↓
	√			√						√		RE	DDI	73.68	74.20	73.94 ★
	√			√				√		√		RE	DDI	68.92	69.20	69.06 ↓
	√		√			√			√			RE	DDI	31.31	76.30	44.40 ↓
	√		√			√				√		RE	DDI	37.86	71.90	49.60 ↓
	√		√			√					√	RE	DDI	54.99	68.90	61.16 ↓
	√			√		√			√			RE	DDI	30.20	66.60	41.56 ↓
	√			√		√				√		RE	DDI	37.97	73.70	50.12 ↓
	√			√		√					√	RE	DDI	60.56	68.80	64.42 ↓
	√				√	√			√			RE	DDI	34.35	80.20	48.10 ↓
	√				√	√					√	RE	DDI	67.07	67.20	67.13 ↓
	√				√	√				√		RE	DDI	53.15	65.00	58.48 ↓
	√		√					√	√			RE	DDI	65.68	70.80	68.14 ↓
	√		√					√		√		RE	DDI	69.80	69.80	69.80 ↓
	√		√					√			√	RE	DDI	67.80	67.80	67.80 ↓
	√				√			√	√			RE	DDI	61.88	72.40	66.73 ↓
	√			√				√			√	RE	DDI	64.29	64.80	64.54 ↓
	√				√			√		√		RE	DDI	61.95	65.30	63.58 ↓
	√				√			√			√	RE	DDI	62.41	68.90	65.49 ↓
	√		√				√		√			RE	DDI	29.27	53.30	37.79 ↓
	√		√				√			√		RE	DDI	29.93	78.20	43.29 ↓
	√		√				√				√	RE	DDI	62.13	67.60	64.75 ↓
	√			√			√		√			RE	DDI	25.19	63.80	36.12 ↓
	√			√			√			√		RE	DDI	21.57	60.30	31.78 ↓
	√			√			√				√	RE	DDI	62.65	66.10	64.33 ↓
	√				√		√		√			RE	DDI	60.96	68.70	64.60 ↓
	√				√		√			√		RE	DDI	36.25	70.80	47.95 ↓
	√				√		√				√	RE	DDI	66.32	70.50	68.35 ↓
		√										Baseline	GIT	77.20	77.20	77.20 ★
√		√										RE	GIT	17.55	80.65	28.82 ↓
	√	√										RE	GIT	64.52	64.52	64.52 ↓
		√			√							RE	GIT	67.31	67.31	67.31 ↓
		√									√	RE	GIT	66.67	66.67	66.67 ↓
		√					√					RE	GIT	43.65	67.31	52.96 ↓
√	√	√										RE	GIT	15.97	70.54	26.04 ↓
		√			√		√					RE	GIT	50.75	58.49	54.35 ↓
		√					√				√	RE	GIT	20.42	71.83	31.79 ↓
		√			√						√	RE	GIT	57.20	57.20	57.20 ↓
		√			√		√				√	RE	GIT	46.21	58.92	51.80 ↓
		√			√			√		√		RE	GIT	41.87	47.10	44.33 ↓
		√			√			√			√	RE	GIT	28.45	43.01	34.25 ↓
		√	√				√		√			RE	GIT	25.73	66.67	37.13 ↓
		√	√				√			√		RE	GIT	17.96	64.30	28.08 ↓
		√	√				√				√	RE	GIT	43.04	58.49	49.59 ↓
		√		√			√		√			RE	GIT	18.54	73.12	29.58 ↓
		√		√			√			√		RE	GIT	26.88	60.86	37.29 ↓
		√		√			√				√	RE	GIT	31.86	54.41	40.19 ↓
		√			√		√		√			RE	GIT	17.37	70.97	27.91 ↓
		√			√		√			√		RE	GIT	27.68	64.95	38.82 ↓
		√	√			√					√	RE	GIT	41.55	59.78	49.03 ↓
		√		√		√				√		RE	GIT	16.04	76.56	26.52 ↓
		√		√		√			√			RE	GIT	16.73	73.76	27.28 ↓
		√		√		√					√	RE	GIT	18.71	68.39	29.38 ↓
		√			√	√			√			RE	GIT	35.96	62.80	45.73 ↓
		√			√	√				√		RE	GIT	20.28	67.96	31.24 ↓
		√			√	√					√	RE	GIT	33.10	61.51	43.04 ↓
		√	√					√	√			RE	GIT	47.53	47.53	47.53 ↓
		√	√					√		√		RE	GIT	34.60	54.62	42.37 ↓
		√	√					√			√	RE	GIT	23.22	60.43	33.55 ↓
		√		√				√	√			RE	GIT	17.65	31.40	22.60 ↓
		√		√				√		√		RE	GIT	29.33	60.86	39.58 ↓
		√		√				√			√	RE	GIT	19.28	46.02	27.17 ↓
		√			√			√	√			RE	GIT	17.93	64.52	28.06 ↓
		√	√			√			√			RE	GIT	18.51	66.45	28.96 ↓
		√	√			√				√		RE	GIT	28.73	54.19	37.56 ↓



(a) BioRED

(b) DDI

(c) GIT

Figure 4: Framework results for datasets in the RE task.

dataset, 30/36 combinations show improvement, 4 combinations for DDI dataset demonstrate improvement, and all combinations we tried for the GIT dataset failed to bring in MTL gain.

As shown in Table. 1, based on the results, the human experience might conclude that MTL works particularly well for the BioRED dataset, where the knowledge learned from other datasets aids the LLM in extracting relations from BioRED. For the DDI dataset, MTL offers slight improvements, but it does not provide any benefits for the GIT dataset.

Another important observation from our data preparation experiments (also considering the results from the ablation study) is that for tasks with poor baseline performance (e.g., F1 scores below 60 in this paper), incorporating multiple related training datasets within our framework significantly improves the results. This suggests that for tasks that are more challenging for LLMs, MTL can effectively leverage diverse data sources to enhance learning and improve performance. In contrast, for tasks with already high baseline performance, the benefits of MTL are minimal, indicating that MTL may not provide substantial improvements for tasks that are already well-optimized. This finding highlights that the impact of MTL is influenced by the LLM's capability to handle the task, showcasing its potential in improving low-performing tasks, while offering limited gains for tasks that are already performing well.

5.2 Find better combination for datasets in the RE task iteratively using the proposed framework

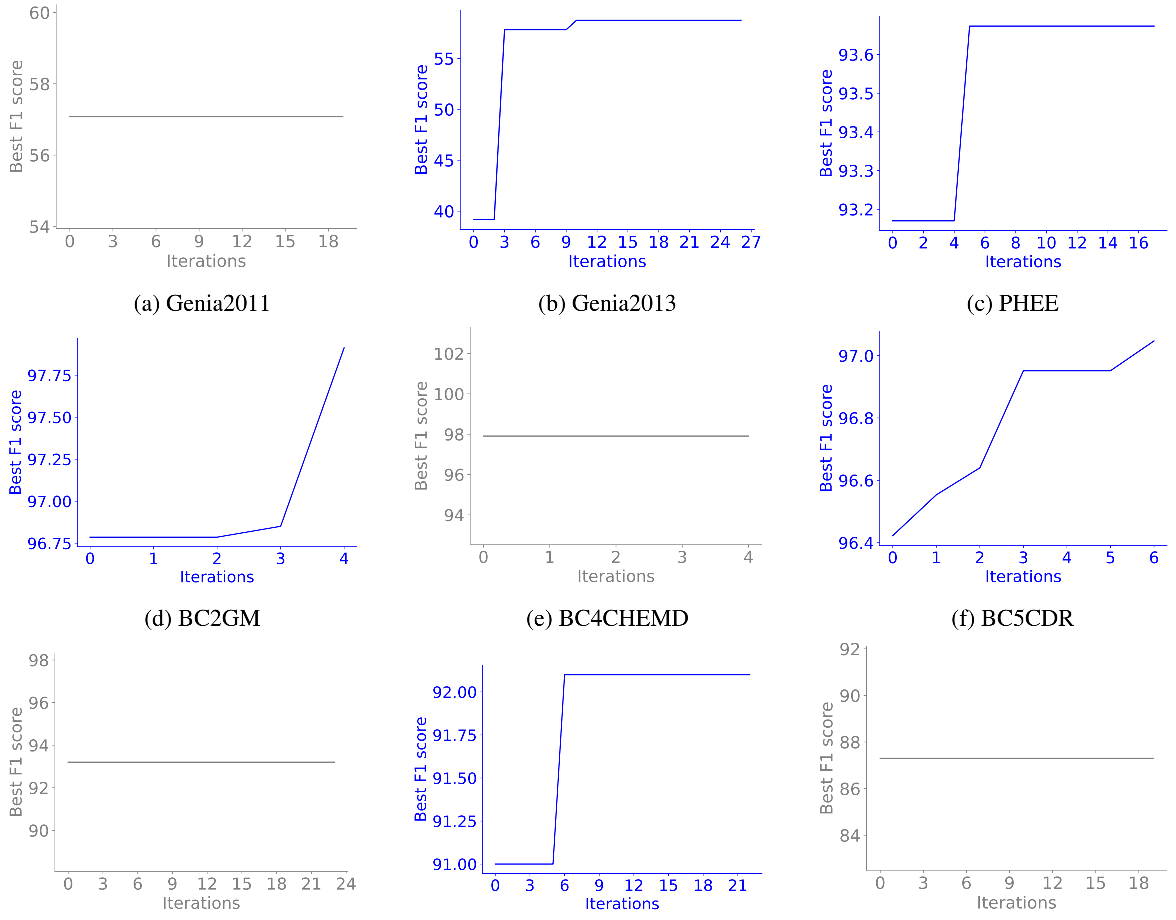
We applied the framework to three datasets in the RE task separately. The framework was set to automatically run for 48 hours, exploring better dataset combinations and stopping once sufficient exploration was achieved, keeping only the best model and the highest F1 score.

When using the framework, for each dataset, after each iteration, we enumerate $2^{11} = 2048$ kinds of combinations for the neural network and find the best combination. The best F1 score curves for three datasets are shown in Fig. 4. For the BioRED dataset in Fig. 4(a), our framework could effectively find better combinations for each of several iterations. In addition, the framework also helps improve the performance of the GIT dataset, and then it stops after some exploration, which is out of expectation and demonstrates the robustness of our framework because it can find better combinations even if our initial combinations are bad. For the DDI dataset, the framework cannot find better combinations and stops soon, which matches our expectations because most of the cases show performance degradation.

5.3 Ablation study for other tasks

Due to space limitation, the different combinations involving the other 9 datasets for NER, EE, and TC tasks are presented in the appendix, in Tables 2, 3, and 4, respectively.

For the EE task, during the data preparation phase, we observed that several dataset combinations could improve LLM performance on GENIA2011 and GENIA2013, further demonstrating the potential of MTL. After applying our framework, the performance of GENIA2013 improved significantly from below 40 to around 58, a substantial gain. This result is corroborated by the improvements in GENIA2011, where an effective combination was tested early in the data preparation stage, boosting performance from a baseline of 34 to 57. Given the high similarity between GENIA2011 and GENIA2013, this suggests that if performance can be enhanced in GENIA2011, similar improvements should likely be achieved for GENIA2013. Although the initial combinations differed, our framework was able to find similar



(a) Genia2011 (b) Genia2013 (c) PHEE

(d) BC2GM (e) BC4CHEMD (f) BC5CDR

(g) ADE (h) HealthAdvice (i) PubMed20krct

Figure 5: Finding best F1 results using proposed framework for EE (GENIA2011, GENIA2013, PHEE), NER (BC2GM, BC4CHEMD, BC5CDR), TC tasks (ADE, HealthAdvice, PubMed20krct).

combinations for both datasets, ultimately achieving comparable results.

For the NER and TC tasks, the six datasets involved all achieved their best F1 scores with the baseline settings, and every combination we attempted resulted in a performance decline. From an expert perspective, it appeared that MTL could not leverage the other datasets in this paper to improve these six datasets. However, we continued to use our framework to investigate whether a specific combination could improve these seemingly unpromising datasets. Surprisingly, as shown in Figures 5(d), 5(f), and 5(h), our framework identified promising combinations that led to significant improvements for three of the datasets, despite initial expectations to the contrary. For the remaining three datasets (BC4CHEMD, ADE, PubMed20krct), we were unable to find better combinations, which aligns with our initial expectations.

Summarizing all the experiments, each dataset had 2,048 possible auxiliary combinations. However, in most cases, we were able to find relatively optimal combinations within just a dozen iterations, and the framework predicted that no better combinations existed. Compared to a brute-force approach, our method significantly improved efficiency.

6 Conclusion

We proposed a novel yet simple framework to address the challenge of selecting the optimal dataset combination for multi-task learning. By iteratively refining these combinations within a feedback loop, we take a significant step toward fully unlocking the potential of MTL in the future.

7 Limitations

The experiments in this paper were conducted using only a single LLM. Although the authors intended to experiment with multiple LLMs to explore broader performance variations, the computational resources and time required were prohibitive. As a result, the findings may not fully represent the potential of our framework.

Secondly, we did not perform a grid search to find the optimal hyperparameters for model training. Instead, we ensured that all experiments were conducted with the same set of parameters to maintain fairness.

Acknowledgements

This work was supported by the National Institutes of Health's National Center for Complementary and Integrative Health under grant numbers R01AT009457 and U01AT012871, the National Institute on Aging under grant number R01AG078154, the National Cancer Institute under grant number R01CA287413, the National Institute of Diabetes and Digestive and Kidney Diseases under grant number R01DK115629, and the National Institute on Minority Health and Health Disparities under grant number 1R21MD019134-01. Many thanks to Yuqian Chen for her help with the drawing.

References

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. *arXiv preprint arXiv:1702.08303*.

Shijie Chen, Yu Zhang, and Qiang Yang. 2024. Multi-task learning in natural language processing: An overview. *ACM Computing Surveys*, 56(12):1–32.

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*.

Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

John C Doyle, Bruce A Francis, and Allen R Tannenbaum. 2013. *Feedback control theory*. Courier Corporation.

Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. 2021. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2019. Autosem: Automatic task selection and mixing in multi-task learning. *arXiv preprint arXiv:1904.04153*.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885 – 892. Text Mining and Natural Language Processing in Pharmacogenomics.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011. Overview of genia event task in bionlp shared task 2011. In *Proceedings of BioNLP shared task 2011 workshop*, pages 7–15.

Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. 2013. The genia event extraction shared task, 2013 edition-overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15.

Bahare Kiumarsi, Frank L Lewis, Hamidreza Modares, Ali Karimpour, and Mohammad-Bagher Naghibi-Sistani. 2014. Reinforcement q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics. *Automatica*, 50(4):1167–1175.

Sotiris B Kotsiantis, Ioannis Zaharakis, P Pintelas, et al. 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7:1–17.

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*, 2016.

Mingchen Li, Ming Chen, Huixue Zhou, and Rui Zhang. 2023. Petailor: Improving large language model by tailored chunk scorer in biomedical triple extraction. *arXiv preprint arXiv:2310.18463*.

Mingchen Li, Zaifu Zhan, Han Yang, Yongkang Xiao, Jiatan Huang, and Rui Zhang. 2024. Benchmarking retrieval-augmented large language models in biomedical nlp: Application, robustness, and self-awareness. *arXiv preprint arXiv:2405.08151*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1930–1939.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. 2016. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3994–4003.

Rajvardhan Patil, Sorio Boit, Venkat Gudivada, and Jagadeesh Nandigam. 2023. A survey of text representation and embedding techniques in nlp. *IEEE Access*, 11:36120–36146.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R Bowman. 2020. Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? *arXiv preprint arXiv:2005.00628*.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2024. Adelie: Aligning large language models on information extraction. *arXiv preprint arXiv:2405.05008*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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*.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350.

Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9:1–19.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. 2020. Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, pages 9120–9132. PMLR.

Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron C Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. 2022. Phee: A dataset for pharmacovigilance event extraction from text. *arXiv preprint arXiv:2210.12560*.

Kim-Han Thung and Chong-Yaw Wee. 2018. A brief review on multi-task learning. *Multimedia Tools and Applications*, 77(22):29705–29725.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Haoxiang Wang, Han Zhao, and Bo Li. 2021. Bridging multi-task learning and meta-learning: Towards efficient training and effective adaptation. In *International conference on machine learning*, pages 10991–11002. PMLR.

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

Xingyao Wang, Sha Li, and Heng Ji. 2022b. Code4struct: Code generation for few-shot event structure prediction. *arXiv preprint arXiv:2210.12810*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System*

Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Gokul Yenduri, M Ramalingam, G Chemmalar Selvi, Y Supriya, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, G Deepti Raj, Rutvij H Jhaveri, B Prabadevi, Weizheng Wang, et al. 2024. Gpt (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*.

Bei Yu, Yingya Li, and Jun Wang. 2019. Detecting causal language use in science findings. 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 4664–4674, Hong Kong, China. Association for Computational Linguistics.

Zaifu Zhan, Shuang Zhou, Mingchen Li, and Rui Zhang. 2025. Ramie: retrieval-augmented multi-task information extraction with large language models on dietary supplements. *Journal of the American Medical Informatics Association*, page ocaf002.

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

Yu Zhang and Qiang Yang. 2018. An overview of multi-task learning. *National Science Review*, 5(1):30–43.

Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609.

Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Sirui Ding, Jiashuo Wang, Kaishuai Xu, Yi Fang, et al. 2024. Large language models for disease diagnosis: A scoping review. *arXiv preprint arXiv:2409.00097*.

A Appendix

Table 2: Data preparation for the NER task, including BC2GM, BC4CHEMD, and BC5CDR datasets. ★ indicates the best F1 score and best combination in the data preparation phase. Each line represents one experiment and √ means the dataset is selected for this run.

Training sets																
RE task			NER task			EE task			TC task					Metrics		
BioRED	DDI	GIT	BC2GM	BC4CHEMD	BC5CDR	GENIA2011	GENIA2013	PHEE	ADE	HealthAdvice	PubMed20krct	Task	Test Set	Precision	Recall	F1 Score
			√									Baseline	BC2GM	98.01	95.60	96.78 ★
			√	√								NER	BC2GM	96.85	95.07	95.95 ↓
			√		√							NER	BC2GM	97.42	94.36	95.87 ↓
			√						√			NER	BC2GM	97.50	91.92	94.63 ↓
√			√									NER	BC2GM	96.30	93.32	94.79 ↓
			√			√						NER	BC2GM	97.51	91.61	94.47 ↓
			√	√	√							NER	BC2GM	96.76	92.77	94.72 ↓
√			√			√						NER	BC2GM	97.01	90.37	93.57 ↓
√			√						√			NER	BC2GM	96.78	91.05	93.83 ↓
			√			√			√			NER	BC2GM	96.62	92.53	94.54 ↓
√			√			√			√			NER	BC2GM	96.95	89.49	93.07 ↓
√			√			√					√	NER	BC2GM	95.59	89.06	92.21 ↓
√			√			√				√		NER	BC2GM	95.31	91.61	93.42 ↓
√			√					√	√			NER	BC2GM	96.35	90.16	93.15 ↓
√			√					√		√		NER	BC2GM	96.21	87.73	91.77 ↓
√			√					√			√	NER	BC2GM	97.28	87.84	92.32 ↓
√			√				√		√			NER	BC2GM	96.40	88.17	92.10 ↓
√			√				√			√		NER	BC2GM	96.68	87.17	91.68 ↓
√			√				√				√	NER	BC2GM	96.11	88.34	92.06 ↓
	√		√			√					√	NER	BC2GM	95.60	85.91	90.49 ↓
	√		√			√				√		NER	BC2GM	95.55	85.83	90.43 ↓
	√		√			√			√			NER	BC2GM	96.10	89.45	92.66 ↓
	√		√					√	√			NER	BC2GM	96.73	87.41	91.83 ↓
		√	√				√			√		NER	BC2GM	96.76	89.36	92.91 ↓
		√	√				√		√			NER	BC2GM	96.52	90.03	93.16 ↓
		√	√				√				√	NER	BC2GM	96.12	89.27	92.57 ↓
	√		√					√		√		NER	BC2GM	96.71	87.45	91.85 ↓
	√		√					√			√	NER	BC2GM	96.66	87.42	91.81 ↓
	√		√				√				√	NER	BC2GM	96.26	89.05	92.51 ↓
	√		√				√		√			NER	BC2GM	95.45	89.64	92.45 ↓
	√		√				√			√		NER	BC2GM	96.52	87.05	91.54 ↓
		√	√			√					√	NER	BC2GM	96.86	87.59	91.99 ↓
		√	√					√		√		NER	BC2GM	95.95	92.47	94.18 ↓
		√	√					√			√	NER	BC2GM	96.79	90.17	93.36 ↓
		√	√					√	√			NER	BC2GM	97.27	88.38	92.61 ↓
		√	√			√				√		NER	BC2GM	95.23	92.07	93.63 ↓
		√	√			√			√			NER	BC2GM	95.02	90.71	92.81 ↓
				√								Baseline	BC4CHEMD	98.62	97.20	97.91 ★
				√	√							NER	BC4CHEMD	97.36	95.64	96.49 ↓
			√	√								NER	BC4CHEMD	98.20	93.44	95.76 ↓
				√						√		NER	BC4CHEMD	97.74	91.80	94.67 ↓
	√			√								NER	BC4CHEMD	98.02	89.97	93.82 ↓
				√				√				NER	BC4CHEMD	97.94	92.68	95.24 ↓
	√			√				√				NER	BC4CHEMD	98.00	89.49	93.55 ↓
	√			√						√		NER	BC4CHEMD	97.60	88.67	92.92 ↓
				√				√		√		NER	BC4CHEMD	97.37	90.14	93.61 ↓
			√	√	√							NER	BC4CHEMD	97.99	93.11	95.49 ↓
	√			√				√		√		NER	BC4CHEMD	97.49	88.01	92.51 ↓
√				√		√			√			NER	BC4CHEMD	98.11	87.13	92.29 ↓
√				√		√				√		NER	BC4CHEMD	96.02	89.68	92.74 ↓
√				√		√					√	NER	BC4CHEMD	97.94	85.96	91.56 ↓
√				√				√		√		NER	BC4CHEMD	97.74	88.30	92.78 ↓
√				√				√	√			NER	BC4CHEMD	97.85	86.98	92.10 ↓
√				√				√			√	NER	BC4CHEMD	97.96	87.50	92.43 ↓
√				√			√				√	NER	BC4CHEMD	96.85	86.85	91.57 ↓
√				√			√			√		NER	BC4CHEMD	97.59	87.59	92.32 ↓
√				√			√		√			NER	BC4CHEMD	98.12	86.31	91.84 ↓
	√			√		√				√		NER	BC4CHEMD	96.94	88.16	92.35 ↓
	√			√		√					√	NER	BC4CHEMD	97.49	87.61	92.29 ↓
	√			√		√			√			NER	BC4CHEMD	96.90	88.23	92.36 ↓
	√			√				√			√	NER	BC4CHEMD	97.87	86.72	91.96 ↓
		√		√			√		√			NER	BC4CHEMD	97.24	89.02	92.95 ↓
		√		√			√			√		NER	BC4CHEMD	96.51	88.46	92.31 ↓
	√			√			√				√	NER	BC4CHEMD	98.00	86.66	91.98 ↓
		√		√			√				√	NER	BC4CHEMD	97.19	88.13	92.44 ↓
	√			√			√		√			NER	BC4CHEMD	97.86	86.43	91.79 ↓
	√			√			√			√		NER	BC4CHEMD	97.29	87.06	91.89 ↓
		√		√		√				√		NER	BC4CHEMD	97.98	86.98	92.15 ↓
		√		√		√			√			NER	BC4CHEMD	97.39	90.75	93.96 ↓
		√		√		√					√	NER	BC4CHEMD	97.54	88.35	92.72 ↓
		√		√				√		√		NER	BC4CHEMD	97.45	89.08	93.07 ↓
		√		√				√	√			NER	BC4CHEMD	97.59	88.92	93.06 ↓
		√		√				√			√	NER	BC4CHEMD	98.33	87.71	92.72 ↓
					√							Baseline	BC5CDR	97.40	95.46	96.42 ★
				√	√							NER	BC5CDR	95.98	95.06	95.52 ↓
			√		√							NER	BC5CDR	96.35	95.28	95.81 ↓
		√			√							NER	BC5CDR	96.60	95.18	95.89 ↓
					√		√					NER	BC5CDR	95.80	93.96	94.87 ↓
					√						√	NER	BC5CDR	96.40	94.02	95.20 ↓
		√			√		√					NER	BC5CDR	96.10	93.54	94.80 ↓
		√			√						√	NER	BC5CDR	96.65	93.24	94.91 ↓
					√		√				√	NER	BC5CDR	96.08	91.41	93.68 ↓
			√	√	√							NER	BC5CDR	96.74	94.21	95.46 ↓
		√			√		√				√	NER	BC5CDR	94.68	89.67	92.11 ↓
√					√	√			√			NER	BC5CDR	95.11	91.20	93.11 ↓
√					√	√					√	NER	BC5CDR	95.46	91.06	93.21 ↓
√					√	√				√		NER	BC5CDR	95.37	91.38	93.33 ↓
√					√			√	√			NER	BC5CDR	96.63	90.16	93.28 ↓
√					√			√		√		NER	BC5CDR	94.83	92.23	93.51 ↓
√					√			√			√	NER	BC5CDR	95.87	90.12	92.90 ↓
√					√		√		√			NER	BC5CDR	96.00	90.96	93.41 ↓
√					√		√				√	NER	BC5CDR	96.44	90.55	93.40 ↓
√					√		√			√		NER	BC5CDR	94.25	92.09	93.15 ↓
	√				√	√					√	NER	BC5CDR	95.66	90.59	93.06 ↓
	√				√	√				√		NER	BC5CDR	94.48	90.76	92.58 ↓
	√				√	√			√			NER	BC5CDR	94.09	91.93	93.00 ↓
		√			√			√		√		NER	BC5CDR	95.74	90.61	93.10 ↓
		√			√			√			√	NER	BC5CDR	95.30	90.90	93.05 ↓
	√				√			√	√			NER	BC5CDR	94.81	89.67	92.17 ↓
	√				√			√		√		NER	BC5CDR	94.68	91.31	92.97 ↓
	√				√			√			√	NER	BC5CDR	96.04	88.27	91.99 ↓
	√				√		√		√			NER	BC5CDR	94.67	90.13	92.34 ↓
		√			√		√		√			NER	BC5CDR	95.42	92.04	93.70 ↓
		√			√		√			√		NER	BC5CDR	94.01	91.52	92.75 ↓
	√				√		√			√		NER	BC5CDR	94.66	89.73	92.13 ↓
	√				√		√				√	NER	BC5CDR	95.65	88.83	92.11 ↓
		√			√	√				√		NER	BC5CDR	95.71	90.97	93.28 ↓
		√			√	√					√	NER	BC5CDR	95.77	90.18	92.89 ↓
		√			√	√			√			NER	BC5CDR	95.62	92.18	93.87 ↓
		√			√			√	√			NER	BC5CDR	95.76	91.35	93.50 ↓

Table 3: Data preparation for the EE task, including GENIA2011, GENIA2013, and PHEE datasets. ★ indicates the best F1 score and best combination in the data preparation phase. Each line represents one experiment and √ means the dataset is selected for this run.

Training sets																
RE task			NER task			EE task			TC task					Metrics		
BioRED	DDI	GIT	BC2GM	BC4CHEMD	BC5CDR	GENIA2011	GENIA2013	PHEE	ADE	HealthAdvice	PubMed20krct	Task	Test Set	Precision	Recall	F1 Score
						√						Baseline	GENIA2011	21.56	86.56	34.52
						√		√				EE	GENIA2011	27.65	81.70	41.32 ↑
						√	√					EE	GENIA2011	45.09	67.82	54.17 ↑
						√			√			EE	GENIA2011	29.71	84.15	43.92 ↑
			√			√						EE	GENIA2011	34.97	80.08	48.68 ↑
√						√						EE	GENIA2011	36.70	73.52	48.96 ↑
						√	√	√				EE	GENIA2011	27.45	81.06	41.01 ↑
√			√			√						EE	GENIA2011	28.78	82.62	42.68 ↑
			√			√			√			EE	GENIA2011	40.65	69.71	51.35 ↑
√						√			√			EE	GENIA2011	29.95	83.68	44.11 ↑
√			√			√			√			EE	GENIA2011	31.47	80.34	45.22 ↑
√			√			√				√		EE	GENIA2011	29.35	80.10	42.96 ↑
√			√			√					√	EE	GENIA2011	31.83	77.86	45.19 ↑
√				√		√			√			EE	GENIA2011	29.75	81.34	43.57 ↑
√				√		√				√		EE	GENIA2011	28.74	82.50	42.63 ↑
√				√		√					√	EE	GENIA2011	32.13	79.31	45.73 ↑
√					√	√			√			EE	GENIA2011	27.51	79.03	40.82 ↑
√					√	√					√	EE	GENIA2011	27.20	83.27	41.01 ↑
√					√	√				√		EE	GENIA2011	28.94	81.48	42.71 ↑
	√				√	√					√	EE	GENIA2011	51.29	51.29	51.29 ↑
	√				√	√				√		EE	GENIA2011	42.45	64.64	51.25 ↑
	√		√			√					√	EE	GENIA2011	32.99	73.61	45.56 ↑
	√			√		√					√	EE	GENIA2011	28.84	73.32	41.40 ↑
	√		√			√				√		EE	GENIA2011	34.84	57.39	43.36 ↑
	√			√		√				√		EE	GENIA2011	33.04	73.99	45.68 ↑
	√		√			√			√			EE	GENIA2011	28.94	80.85	42.62 ↑
	√			√		√			√			EE	GENIA2011	28.78	69.08	40.63 ↑
	√				√	√			√			EE	GENIA2011	32.81	77.67	46.13 ↑
		√	√			√					√	EE	GENIA2011	34.03	66.19	44.95 ↑
		√		√		√				√		EE	GENIA2011	35.05	76.31	48.03 ↑
		√		√		√			√			EE	GENIA2011	30.67	77.54	43.95 ↑
		√		√		√					√	EE	GENIA2011	31.50	73.83	44.16 ↑
		√			√	√				√		EE	GENIA2011	38.17	69.41	49.26 ↑
		√			√	√			√			EE	GENIA2011	31.54	81.54	45.48 ↑
		√			√	√					√	EE	GENIA2011	53.91	60.65	57.08 ★
		√	√			√				√		EE	GENIA2011	33.08	78.30	46.51 ↑
		√	√			√			√			EE	GENIA2011	32.51	77.04	45.73 ↑
							√					Baseline	GENIA2013	14.42	37.54	20.83
						√	√					EE	GENIA2013	13.61	35.49	19.68 ↓
							√	√				EE	GENIA2013	19.62	48.81	27.98 ↑
					√		√					EE	GENIA2013	9.31	27.99	13.97 ↓
		√					√					EE	GENIA2013	16.89	43.34	24.31 ↑
							√				√	EE	GENIA2013	16.49	42.66	23.79 ↑
						√	√	√				EE	GENIA2013	15.37	45.73	23.00 ↑
		√			√		√					EE	GENIA2013	13.08	39.25	19.62 ↓↑
		√					√				√	EE	GENIA2013	28.75	61.43	39.17 ★
					√		√				√	EE	GENIA2013	7.20	23.55	11.03 ↓
		√			√		√				√	EE	GENIA2013	24.76	60.41	35.12 ↑
√			√				√		√			EE	GENIA2013	16.75	45.39	24.47 ↑
√			√				√			√		EE	GENIA2013	15.59	47.10	23.43 ↑
√				√			√			√		EE	GENIA2013	20.14	69.28	31.21 ↑
√				√			√		√			EE	GENIA2013	9.22	33.11	14.42 ↓
√				√			√				√	EE	GENIA2013	22.08	53.58	31.27 ↑
√			√				√				√	EE	GENIA2013	20.69	59.39	30.69 ↑
√					√		√		√			EE	GENIA2013	9.46	31.40	14.53 ↓
√					√		√			√		EE	GENIA2013	19.35	65.19	29.84 ↑
√					√		√				√	EE	GENIA2013	17.22	47.44	25.27 ↑
		√	√				√			√		EE	GENIA2013	24.10	59.73	34.35 ↑
		√	√				√		√			EE	GENIA2013	12.70	35.84	18.75 ↓
		√	√				√				√	EE	GENIA2013	15.05	39.46	21.78 ↑
		√		√			√		√			EE	GENIA2013	8.60	30.61	13.43 ↓
	√		√				√				√	EE	GENIA2013	17.49	47.10	25.51 ↑
	√		√				√		√			EE	GENIA2013	16.48	49.83	24.77 ↑
		√		√			√			√		EE	GENIA2013	11.14	34.81	16.87 ↓
		√		√			√				√	EE	GENIA2013	26.68	70.31	38.69 ↑
	√		√				√			√		EE	GENIA2013	21.57	64.85	32.37 ↑
	√			√			√				√	EE	GENIA2013	22.03	59.39	32.13 ↑
	√				√		√		√			EE	GENIA2013	18.03	57.00	27.40 ↑
	√			√			√		√			EE	GENIA2013	12.66	40.27	19.27 ↑
	√			√			√			√		EE	GENIA2013	19.19	53.24	28.21 ↑
		√			√		√		√			EE	GENIA2013	17.22	44.37	24.81 ↑
		√			√		√			√		EE	GENIA2013	21.78	52.56	30.80 ↑
	√				√		√			√		EE	GENIA2013	17.19	64.63	27.16 ↑
	√				√		√				√	EE	GENIA2013	15.52	42.32	22.71 ↑
								√				Baseline	PHEE	48.51	93.75	63.94 ★
							√	√				EE	PHEE	55.56	93.15	69.60 ↑
						√		√				EE	PHEE	47.93	93.55	63.39 ↓
								√		√		EE	PHEE	49.73	92.34	64.64 ↓
				√				√				EE	PHEE	45.35	84.48	59.01 ↓
	√							√				EE	PHEE	76.19	89.01	82.10 ↑
						√	√	√				EE	PHEE	47.55	92.84	62.89 ↓
	√			√				√				EE	PHEE	46.41	90.02	61.25 ↓
	√							√		√		EE	PHEE	47.36	92.24	62.59 ↓
				√				√		√		EE	PHEE	47.39	92.34	62.63 ↓
	√			√				√		√		EE	PHEE	69.35	88.51	77.77 ↑
√			√					√	√			EE	PHEE	47.67	92.84	63.00 ↓
√			√					√		√		EE	PHEE	47.62	92.94	62.98 ↓
√				√				√	√			EE	PHEE	47.37	92.44	62.64 ↓
√				√				√		√		EE	PHEE	47.49	92.64	62.79 ↓
√				√				√			√	EE	PHEE	49.86	92.44	64.78 ↓
√					√			√	√			EE	PHEE	47.10	91.53	62.19 ↓
√			√					√			√	EE	PHEE	47.42	92.54	62.70 ↓
√					√			√		√		EE	PHEE	49.35	91.43	64.10 ↓
√					√			√			√	EE	PHEE	47.78	93.15	63.16 ↓
	√		√					√	√			EE	PHEE	46.53	90.62	61.49 ↓
		√			√			√		√		EE	PHEE	51.09	89.82	65.13 ↓
		√			√			√			√	EE	PHEE	47.83	79.94	59.85 ↓
	√		√					√		√		EE	PHEE	46.07	89.31	60.79 ↓
	√		√					√			√	EE	PHEE	18.18	21.77	19.82 ↓
	√				√			√	√			EE	PHEE	53.36	67.24	59.50 ↓
	√				√			√		√		EE	PHEE	47.67	92.94	63.02 ↓
	√				√			√			√	EE	PHEE	49.23	90.83	63.86 ↓
	√			√				√			√	EE	PHEE	48.48	93.04	63.74 ↓
		√		√				√	√			EE	PHEE	46.95	90.02	61.71 ↓
		√	√					√		√		EE	PHEE	48.53	91.63	63.46 ↓
		√		√				√		√		EE	PHEE	94.23	92.14	93.17 ↑
		√	√					√	√			EE	PHEE	48.32	92.84	63.56 ↓
		√		√				√			√	EE	PHEE	47.64	92.64	62.92 ↓
		√	√					√			√	EE	PHEE	53.27	95.97	68.51 ↓
		√			√			√	√			EE	PHEE	47.72	93.04	63.09 ↓

Table 4: Data preparation for the TC task, including ADE, HealthAdvice, and PubMed20krct datasets. ★ indicates the best F1 score and best combination in the data preparation phase. Each line represents one experiment and √ means the dataset is selected for this run.

Training sets																
RE task			NER task			EE task			TC task					Metrics		
BioRED	DDI	GIT	BC2GM	BC4CHEMD	BC5CDR	GENIA2011	GENIA2013	PHEE	ADE	HealthAdvice	PubMed20krct	Task	Test Set	Precision	Recall	F1 Score
									√			TC	ADE	93.20	93.20	93.20 ★
									√	√		TC	ADE	80.30	80.30	80.30 ↓
									√		√	TC	ADE	59.40	59.40	59.40 ↓
			√						√			TC	ADE	90.10	90.10	90.10 ↓
						√			√			TC	ADE	60.28	89.40	72.01 ↓
√									√			TC	ADE	81.31	88.30	84.66 ↓
									√	√	√	TC	ADE	86.90	86.90	86.90 ↓
√						√			√			TC	ADE	72.40	89.70	80.13 ↓
			√			√			√			TC	ADE	86.60	89.20	87.88 ↓
√			√						√			TC	ADE	46.89	87.30	61.01 ↓
√			√			√			√			TC	ADE	36.46	84.00	50.85 ↓
		√			√		√		√			TC	ADE	85.00	85.00	85.00 ↓
		√		√			√		√			TC	ADE	89.20	89.20	89.20 ↓
		√		√		√			√			TC	ADE	56.42	80.80	66.45 ↓
		√		√				√	√			TC	ADE	68.17	70.90	69.51 ↓
		√	√				√		√			TC	ADE	87.30	87.30	87.30 ↓
√					√		√		√			TC	ADE	83.00	83.00	83.00 ↓
		√	√					√	√			TC	ADE	83.47	83.80	83.63 ↓
	√		√				√		√			TC	ADE	82.30	82.30	82.30 ↓
√					√	√			√			TC	ADE	29.17	33.20	31.06 ↓
		√	√			√			√			TC	ADE	44.69	90.80	59.89 ↓
	√			√			√		√			TC	ADE	79.80	79.80	79.80 ↓
√					√			√	√			TC	ADE	30.24	33.90	31.97 ↓
	√			√		√			√			TC	ADE	23.99	32.50	27.60 ↓
	√				√			√	√			TC	ADE	87.60	87.60	87.60 ↓
	√				√		√		√			TC	ADE	84.50	84.50	84.50 ↓
		√			√	√			√			TC	ADE	55.58	89.60	68.61 ↓
		√			√			√	√			TC	ADE	89.81	89.90	89.86 ↓
√				√		√			√			TC	ADE	48.14	53.00	50.45 ↓
√			√				√		√			TC	ADE	80.71	82.00	81.35 ↓
√			√					√	√			TC	ADE	67.46	68.00	67.73 ↓
√				√				√	√			TC	ADE	40.39	49.20	44.36 ↓
√				√			√		√			TC	ADE	80.40	80.40	80.40 ↓
	√				√	√			√			TC	ADE	86.80	86.80	86.80 ↓
	√		√			√			√			TC	ADE	59.61	78.80	67.87 ↓
	√		√					√	√			TC	ADE	70.58	72.70	71.63 ↓
										√		TC	HealthAdvice	91.00	91.00	91.00 ★
									√	√		TC	HealthAdvice	89.70	89.70	89.70 ↓
										√	√	TC	HealthAdvice	90.10	90.10	90.10 ↓
	√									√		TC	HealthAdvice	87.30	87.30	87.30 ↓
								√		√		TC	HealthAdvice	83.93	86.70	85.29 ↓
				√						√		TC	HealthAdvice	85.70	85.70	85.70 ↓
									√	√	√	TC	HealthAdvice	87.20	87.20	87.20 ↓
	√							√		√		TC	HealthAdvice	79.74	80.30	80.02 ↓
	√			√						√		TC	HealthAdvice	84.80	84.80	84.80 ↓
				√				√		√		TC	HealthAdvice	62.05	83.40	71.16 ↓
	√			√				√		√		TC	HealthAdvice	77.60	77.60	77.60 ↓
		√			√		√			√		TC	HealthAdvice	85.20	85.20	85.20 ↓
		√		√			√			√		TC	HealthAdvice	85.20	85.20	85.20 ↓
		√	√			√				√		TC	HealthAdvice	78.51	81.10	79.78 ↓
		√		√				√		√		TC	HealthAdvice	83.50	83.50	83.50 ↓
		√		√		√				√		TC	HealthAdvice	53.56	81.30	64.58 ↓
		√	√				√			√		TC	HealthAdvice	86.50	86.50	86.50 ↓
	√			√			√			√		TC	HealthAdvice	55.81	80.20	65.82 ↓
		√	√					√		√		TC	HealthAdvice	80.06	84.70	82.31 ↓
√					√	√				√		TC	HealthAdvice	41.20	92.50	57.01 ↓
		√			√			√		√		TC	HealthAdvice	84.80	84.80	84.80 ↓
		√			√	√				√		TC	HealthAdvice	82.00	87.90	84.85 ↓
	√		√				√			√		TC	HealthAdvice	84.33	85.60	84.96 ↓
√					√			√		√		TC	HealthAdvice	86.90	86.90	86.90 ↓
	√		√			√				√		TC	HealthAdvice	59.73	78.60	67.88 ↓
	√				√			√		√		TC	HealthAdvice	77.30	77.30	77.30 ↓
	√		√					√		√		TC	HealthAdvice	84.40	84.40	84.40 ↓
	√				√		√			√		TC	HealthAdvice	81.50	81.50	81.50 ↓
√					√		√			√		TC	HealthAdvice	57.08	83.40	67.78 ↓
√				√		√				√		TC	HealthAdvice	69.02	80.20	74.19 ↓
	√				√	√				√		TC	HealthAdvice	82.92	83.00	82.96 ↓
√			√			√				√		TC	HealthAdvice	83.01	85.00	83.99 ↓
√			√				√			√		TC	HealthAdvice	84.32	84.40	84.36 ↓
√				√				√		√		TC	HealthAdvice	68.04	72.80	70.34 ↓
√			√					√		√		TC	HealthAdvice	78.00	78.00	78.00 ↓
√				√			√			√		TC	HealthAdvice	58.51	85.60	69.51 ↓
	√			√		√				√		TC	HealthAdvice	83.17	84.50	83.83 ↓
											√	TC	PubMed20krct	87.30	87.30	87.30 ★
									√		√	TC	PubMed20krct	84.60	84.60	84.60 ↓
										√	√	TC	PubMed20krct	83.10	83.10	83.10 ↓
		√									√	TC	PubMed20krct	84.20	84.20	84.20 ↓
					√						√	TC	PubMed20krct	85.10	85.10	85.10 ↓
							√				√	TC	PubMed20krct	84.36	84.70	84.53 ↓
									√	√	√	TC	PubMed20krct	85.10	85.10	85.10 ↓
		√					√				√	TC	PubMed20krct	82.28	83.10	82.69 ↓
		√			√						√	TC	PubMed20krct	86.30	86.30	86.30 ↓
					√		√				√	TC	PubMed20krct	82.64	83.30	82.97 ↓
		√			√		√				√	TC	PubMed20krct	81.40	81.40	81.40 ↓
		√	√				√				√	TC	PubMed20krct	82.90	82.90	82.90 ↓
		√		√			√				√	TC	PubMed20krct	84.10	84.10	84.10 ↓
		√			√	√					√	TC	PubMed20krct	84.30	84.30	84.30 ↓
		√	√			√					√	TC	PubMed20krct	82.02	83.00	82.50 ↓
	√			√			√				√	TC	PubMed20krct	56.46	59.90	58.13 ↓
		√		√		√					√	TC	PubMed20krct	75.84	81.60	78.61 ↓
		√	√					√			√	TC	PubMed20krct	84.40	84.40	84.40 ↓
		√		√				√			√	TC	PubMed20krct	69.37	71.10	70.22 ↓
		√			√			√			√	TC	PubMed20krct	81.30	81.30	81.30 ↓
√					√	√					√	TC	PubMed20krct	48.16	83.60	61.11 ↓
√					√			√			√	TC	PubMed20krct	48.26	72.00	57.78 ↓
	√				√			√			√	TC	PubMed20krct	58.81	61.40	60.08 ↓
	√				√		√				√	TC	PubMed20krct	69.57	72.00	70.76 ↓
	√		√					√			√	TC	PubMed20krct	82.90	82.90	82.90 ↓
	√			√				√			√	TC	PubMed20krct	59.93	68.80	64.06 ↓
	√		√				√				√	TC	PubMed20krct	66.98	78.90	72.45 ↓
√					√		√				√	TC	PubMed20krct	82.30	82.30	82.30 ↓
	√				√	√					√	TC	PubMed20krct	75.75	77.80	76.76 ↓
√			√			√					√	TC	PubMed20krct	49.03	78.30	60.30 ↓
√				√		√					√	TC	PubMed20krct	34.66	80.40	48.43 ↓
√			√				√				√	TC	PubMed20krct	71.18	72.60	71.88 ↓
√				√				√			√	TC	PubMed20krct	27.59	92.10	42.46 ↓
√				√			√				√	TC	PubMed20krct	42.18	72.00	53.20 ↓
√			√					√			√	TC	PubMed20krct	55.82	63.80	59.54 ↓
	√			√		√					√	TC	PubMed20krct	81.71	82.20	81.95 ↓
	√		√			√					√	TC	PubMed20krct	80.84	81.00	80.92 ↓