Fine-Tuning Language Models on Multiple Datasets for Citation Intention Identification

Anonymous ACL submission

Abstract

Citation intention identification (CII) tools classify citations by their intention (e.g., background, motivation) and assist readers in comprehending scientific literature. Prior research has shown that pretrained language models (PLMs) such as SciBERT can achieve state-of-the-art performance on CII benchmarks. PLMs are trained via self-supervision tasks on a large corpus of general text and can quickly adapt to CII tasks via moderate fine-tuning on the corresponding dataset. Despite their advantages, PLMs can easily overfit small datasets during fine-tuning. In this paper, we propose a multi-task learning (MTL) framework that jointly fine-tunes PLMs on a dataset of primary interest together with multiple auxiliary CII datasets to take advantage of additional supervision signals. We develop a data-driven task relation learning method that controls the contribution of auxiliary datasets to avoid negative transfer and expensive hyper-parameter tuning. We conduct experiments on three CII datasets and show that fine-tuning with additional datasets can improve the PLMs' generalization performance on the primary dataset. PLMs fine-tuned with our proposed framework outperform the current state-of-the-art models by 7% to 11% on small datasets while aligning with the best-performing model on a large dataset.

1 Introduction

In a scientific paper, citations are critical components that help the reader establish context for the paper and its contributions (Manchanda and Karypis, 2021). A paper usually contains tens to hundreds of citations for different purposes. Among the citations, some of them are typically highly influential to the citing paper (e.g., the proposed method in the citing paper is an extension of the cited paper) while most of the others serve as background knowledge. Citation intention identification (CII) tools that classify citations by their intention can help readers quickly pinpoint relevant citations in scientific papers. Moreover, the intention of citations is essential information for building bibliometrics for evaluating the relevance and contribution of scientific publications (Hernández-Alvarez and Gomez, 2016).

Prior research commonly formulates CII as a text classification problem and solves it using machine learning methods (Jurgens et al., 2018; Cohan et al., 2019; Berrebbi et al., 2022). They extract the citation context from the citing papers (i.e., a span of text around the citation), and use it as an input to classifiers. Among these methods, pretrained language models (PLMs) (Devlin et al., 2019; Liu et al., 2019b) achieve the current state-of-the-art performance on CII benchmarks (Beltagy et al., 2019). Researchers apply PLMs to the CII problem by fine-tuning them on labeled citations. However, obtaining labeled citations for fine-tuning PLMs is challenging, as labeling citations in a scientific domain requires experts with in-depth domain knowledge.

Over the years, different CII datasets have been curated that assign the citations to different intention categories (Hernández-Alvarez and Gomez, 2016). They share an input space which is a set of citation contexts extracted from scientific papers. The set of citation labels of these datasets may contain semantically identical or similar intention categories (e.g., *"Background"* and *"Not Used"*). Accordingly, fine-tuning a PLM on one such dataset may benefit its generalization performance to another. In this paper, we aim to improve PLMs' generalization performance on CII datasets by leveraging supervision signals from additional CII datasets.

We propose a multi-task learning (MTL) framework that jointly fine-tunes PLMs on auxiliary CII datasets to improve the PLMs' generalizability on a primary CII dataset of interest. To prevent negative transfer (Wang et al., 2019) (wherein sharing information with unrelated tasks harms the per-

formance of the primary task) and to reduce the burden of hyper-parameter search, we propose a task relation learning (TRL) method to control the contribution of the auxiliary datasets to the training. The TRL method measures the relevance of an auxiliary dataset to the primary dataset by evaluating the *information gain* (Shannon, 1948) of a model trained on the auxiliary dataset on the primary dataset. We also find that the position of a citation within its context provides useful information for CII tasks and a position-aware readout function, i.e., a function that aggregates the PLM output token embeddings to a fixed length citation embedding, can improve PLMs' performance.

Our contributions are summarized as follows:

- We introduce a MTL framework that fine-tunes PLMs jointly on multiple CII datasets to improve their generalizability on the primary dataset of interest.
- We propose a data-driven TRL method that controls the contribution of auxiliary datasets in the MTL framework. It effectively and efficiently avoids negative transfer.
- We find that the position of the citation within the context is an informative signal for predicting citation intentions. We propose a position-aware readout function that outperforms the commonly used CLS and MEAN readout functions.
- We curate a new benchmark dataset called KIM that is specialized for the development of CII applications in materials science.

We carefully design and conduct experiments on three benchmark datasets which show that jointly fine-tuning PLMs on multiple datasets with the proposed MTL framework improves the PLMs' performance on the primary dataset. PLMs fine-tuned with our framework outperform the current state-of-the-art models by 7% to 11% on small datasets while align with the best-performing model on a large dataset.

2 Related Work

2.1 Citation Intention Identification

CII is a classification task that assigns citations into discrete intention categories such as *background* and *motivation*. A citation consists of several components, including a citation context (i.e., a span of text that contains the citation), topology information about the citing paper and the cited paper (e.g., their neighbors in citation graphs), meta information such as the title of the section that contains the citation, and etc.

Citation context is arguably the primary signal for CII methods. The majority of prior research formulates CII as a text classification problem and focuses on featurization of citation contexts. Early works (Abu-Jbara et al., 2013; Jurgens et al., 2018) represent citation contexts by hand-engineered features and pre-trained word embeddings (e.g., GloVe (Pennington et al., 2014), ELMo (Peters et al., 2018)). These methods apply traditional classification models such as support vector machines (SVM) (Schölkopf and Smola, 2002) to predict citation intentions. Deep learning-based methods (Cohan et al., 2019) use word embeddings together with a bi-directional long short-term memory network (BiLSTM) (Hochreiter and Schmidhuber, 1997) to learn context representations end-to-end for CII tasks. In recent years, Transformer (Vaswani et al., 2017)-based PLMs (Devlin et al., 2019; Liu et al., 2019b) revolutionized a wide range of NLP tasks, including CII. A PLM pre-trained on scientific literature, SciBERT (Beltagy et al., 2019), achieves the state-of-the-art performance on various CII datasets.

Some methods explore leveraging the other sources of information to improve CII performance. Cohan et al. (2019) demonstrate that training CII models with two auxiliary tasks, citation worthiness prediction and the section title prediction, effectively improves the generalization performance of CII models. Berrebbi et al. (2022) leverage the topology information from citation graphs as additional signals for predicting citation intentions and achieve better performance than methods that only predict by citation context.

2.2 Multi-task Learning

Multi-task learning (MTL) (Ruder, 2017) is a paradigm that jointly trains a model with shared parameters on multiple related tasks such that the knowledge acquired from them can benefit the learning of each other (Yu et al., 2020; Liu et al., 2019a; Tao and Busso, 2020). MTL formulations usually fall into two categories. The first category, which includes the present work, has one or more primary tasks (a.k.a. target tasks) and a set of auxiliary tasks that serve as regularizers to improve the

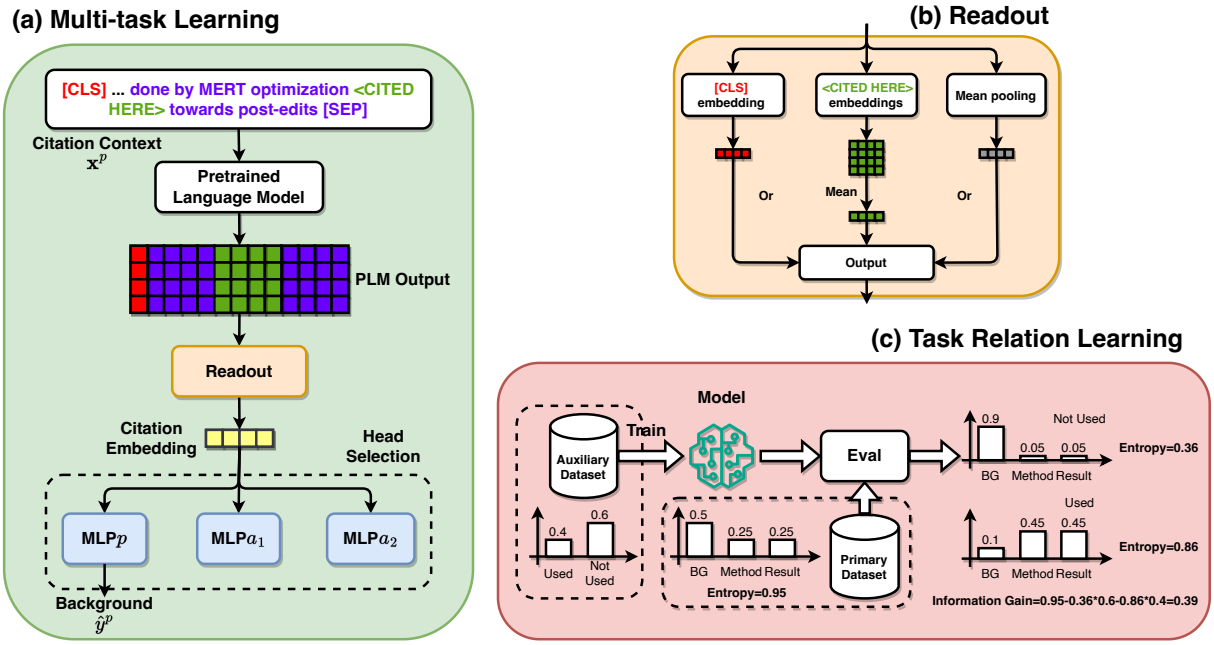

Figure 1: Overview of the architecture of our multi-task learning framework. (a) An overview of the MTL training process. The same language model parameters are shared across all datasets.; (b) The three readout operations (CLS, MEAN, CITED HERE) over the language model embeddings to generate representations for citation contexts; (c) The task relation learning (TRL) method. We train a classification model on an auxiliary dataset then evaluate its information gain on the primary dataset.

model's generalizability on the primary tasks (Ning et al., 2009; Liebel and Körner, 2018; Cohan et al., 2019). In the second category, tasks are equally important and the goal is to reach the Pareto frontier (Lin et al., 2019) over these tasks (Cao et al., 2022; Kendall et al., 2018).

Task weight assignment is an essential topic in MTL research. Earlier methods either assign task weights apriori by experts to reflect domain preference (Kokkinos, 2017; Eigen and Fergus, 2015) or tune the weights as hyper-parameters using validation sets (Cohan et al., 2019). Kendall et al. (2018) propose to weigh different tasks by a measurement of their uncertainties and show improved performance on all tasks compared to training separate models for each task. Our task relation learning method is a data-driven method that finds the relation between each auxiliary dataset (task) and the target dataset (task) to improve the model's performance on the target dataset.

2.3 Pretrained Language Model

Pretrained language models (PLM), a.k.a. large language models (LLM), are models pretrained on unlabeled text corpus with self-supervised language modeling tasks such as causal language modeling (Radford et al., 2018) and masked language modeling (Taylor, 1953). PLMs achieve state-of-the-art performance on a wide range of downstream NLP tasks via gradient-based fine-tuning on the corresponding datasets (Devlin et al., 2019; Liu et al., 2019b; Beltagy et al., 2019; Raffel et al., 2020; Radford et al., 2019).

Recently, Brown et al. demonstrate generative PLMs can be specialized for different tasks with in-context learning (ICL), a fine-tuning-free learning paradigm that unifies NLP tasks to the language generation task. Although the increasing scale of PLMs has been improving the performance of ICL on NLP tasks (Wei et al., 2023), we find that ICL with GPT4 (Bubeck et al., 2023) still underperforms our proposed method.

3 Methodology

3.1 Problem Setting

We formulate the CII task as a text classification problem. A CII dataset $\mathcal{D}^t = \{\mathbf{x}_i^t, y_i^t\}_{i=1}^{|\mathcal{D}^t|}$ is a set of instances where $x_i^t \in \mathcal{X}^t$ is a citation context and $y_i^t \in \mathcal{Y}^t$ is a discrete variable that indicates the intention of the citation. A citation context is a span of text around the citation. We posit the existence of multiple CII datasets and use the superscript t to distinguish different datasets. In the

MTL setting, there is a primary dataset of interest $\mathcal{D}^p$ and a set of auxiliary datasets $\{\mathcal{D}^a\}_{a \in A}$. We assume the datasets share an input space $\mathcal{X}^t = \mathcal{X}, \forall t \in \{p\} \cup A$ (the text of the citation context from scientific literature). We do not posit any assumptions on the label space $\mathcal{Y}^t$, different CII datasets can have different label spaces. Our goal is to train a CII model $\mathcal{M}$ such that

$$\mathcal{M}\left(\mathbf{x}_i^p\right) = y_i^p, \forall \left(\mathbf{x}_i^p, y_i^p\right) \in \mathcal{D}^p.$$

3.2 Multi-task Learning

Although the intention (label) spaces for different CII datasets are different, most of the CII datasets share an input space, citation context from scientific literature, for predicting the intention of the citations (Hernández-Alvarez and Gomez, 2016). Moreover, the label spaces of the datasets may contain semantically similar or even shared intention categories (e.g., "Background" and "Not Used"). It is intuitive to assume leveraging the supervision signal of one dataset to fine-tune a PLM can improve its performance on another dataset. To that end, we propose a MTL framework that jointly fine-tunes a PLM on additional CII datasets as auxiliary tasks to improve its generalizability on a primary dataset.

The MTL framework shares a PLM across datasets while using a separate prediction head (MLP) for each of them. During training, each dataset-specific MLP optimizes its parameters using gradients computed from its own prediction while losses of all prediction heads back-propagate to the PLM to update its parameters. Inference over the primary dataset is performed using its corresponding MLP. The MTL framework is shown in Figure 1.

Letting Θ_{LM} be the parameters of the PLM and Θ_{MLP_a} be the parameters of the MLP for dataset a, the objective function for the MTL framework is

$$\begin{aligned}\mathcal{L} = & \frac{1}{|\mathcal{D}^p|} \sum_{(x_i^p, y_i^p) \in \mathcal{D}^p} l^p\left(f^p\left(x_i^p; \Theta_{\text{LM}}, \Theta_{\text{MLP}_p}\right), y_i^p\right) \\ & + \sum_{a \in A} \frac{\lambda_a}{|\mathcal{D}^a|} \sum_{(x_i^a, y_i^a) \in \mathcal{D}^a} l^a\left(f^a\left(x_i^a; \Theta_{\text{LM}}, \Theta_{\text{MLP}_a}\right), y_i^a\right)\end{aligned} \tag{1}$$

where p is the primary dataset and A is the set of auxiliary datasets. $f^a(\cdot)$ and $l^a(\cdot)$ are the prediction function and the cross-entropy loss associated with dataset a, respectively. The coefficients $\lambda_a \in [0, 1]$, whose importance we discuss in the next section, control the contribution of dataset a towards the primary dataset p.

3.3 Task Relation Learning

In MTL, sharing information with relevant tasks may benefit the primary task but learning unrelated tasks may harm the performance of the primary task (negative transfer) (Wang et al., 2019). Identifying related tasks is thus critical in designing MTL frameworks. As shown in Equation 1, we use a coefficient λ_a to control how much an auxiliary dataset a contributes to the primary dataset. A common way to determine the value of λ_a is to treat it as a hyper-parameter and tune its value for best primary accuracy on a validation set via grid search. However, the search space grows exponentially as the number of auxiliary datasets increases, making a grid search too inefficient.

Since all datasets are for CII tasks and share input space, a model trained on one dataset may generate meaningful predictions in the label space of another dataset. For example, a citation that is labeled as *"Background"* in one dataset may be classified as *"Unused"* by a model trained on another dataset. Based on this insight, we propose a task relation learning (TRL) method to determine the value of λ_a. Our method trains a model on the auxiliary dataset a and evaluates it on the primary dataset p. If the model performs well on the primary dataset, we assume that jointly training on the auxiliary dataset could benefit the model's performance on the primary dataset. Considering the different label spaces of the datasets, it is intractable to employ traditional classification metrics such as accuracy and F1 scores for such evaluations. We propose to use information gain as the metric for the evaluation.

Information gain is defined on the basis of entropy (Shannon, 1948), a measure of uncertainty in information theory. Let $\mathcal{Y}^p$ and $\mathcal{Y}^a$ be the label spaces of the primary dataset and an auxilary dataset, respectively. The entropy of $\mathcal{Y}^p$ is

$$\text{Entr}(\mathcal{Y}^p) = -\sum_{i \in \mathcal{Y}^p} P(i) \log_{|\mathcal{Y}^p|}(P(i)), \tag{2}$$

which measures the uncertainty in set $\mathcal{Y}^p$ in the range of $[0, 1]$. $P(i)$ is the probability that intention i emerges in the dataset. When there is only one label $i \in \mathcal{Y}^p$ in the dataset, it is certain that any sample will have label i and the entropy is zero. When the labels are evenly distributed, a sample

from the dataset is completely uncertain and the value of the entropy is one. We compute entropy on the label distribution of the primary dataset.

After training a model on the auxiliary dataset, we apply the model to the primary dataset and obtain a predicted label from the auxiliary label space for each instance. We group the instances by their predicted labels and obtain a conditional primary label distribution for each auxiliary label $j \in \mathcal{Y}^a$. We compute entropy on each conditional label distribution by

$$\text{Entr}(\mathcal{Y}^p|j) = -\sum_{i \in \mathcal{Y}^a} P(i|j) \log_{|\mathcal{Y}^p|}(P(i|j)), \quad (3)$$

where $P(i|j)$ denotes the probability that a primary instance with label i is predicted into the auxiliary intention j. Note that, a small value of $\text{Entr}(\mathcal{Y}^p|j)$ indicates the predicted auxiliary label j is likely to be a sub-class of one label in the primary label space $\mathcal{Y}^p$ as most of the instances with predicted label j belong to one label in the primary label space $\mathcal{Y}^p$. We calculate information gain by

$$\text{IG}(\mathcal{Y}^p|\mathcal{Y}^a) = \text{Entr}(\mathcal{Y}^p) - \sum_{j \in \mathcal{Y}^a} P(j)(\text{Entr}(\mathcal{Y}^p|j)), \quad (4)$$

where $P(j)$ is the probability that a primary instance is predicted as label j. Information gain quantifies the uncertainty reduction when we group the primary instances by their predicted label in the auxiliary label space. A large value of information gain indicates that each label in the auxiliary space is a sub-class of a primary label which means the auxiliary label space is highly correlated with the primary label space.

We compute the value of λ_a as the relative reduction of entropy

$$\lambda_a = \frac{\text{IG}(\mathcal{Y}^p|\mathcal{Y}^a)}{\text{Entr}(\mathcal{Y}^p)}, \quad (5)$$

which falls into the range of $[0, 1]$. The closer λ_a is to one, the more similar dataset a is to the primary dataset. We show an example of the TRL method in Figure 1.

3.4 Readout Function

A common practice of fine-tuning PLMs for text classification is to use them as text encoders that convert citation contexts to vectors in a latent space and feed the latent vectors through prediction heads (MLPs) to obtain classification probabilities (Sun et al., 2019). PLMs output a contextualized embedding for tokens at each position of the citation contexts. For each context, we use a readout function to aggregate its contextualized token embeddings to a sentence embedding to feed into the downstream MLPs. We explore two standard readout functions, CLS that uses the output embedding of the <CLS> token as the context embedding, and MEAN that averages the contextualized embeddings of all tokens to be the context embedding (Beltagy et al., 2019; Devlin et al., 2019; Reimers and Gurevych, 2019). We also try a third approach called CITED HERE that is motivated by the fact that the position of the citation in the citation context can be informative for predicting the intention of the citation (Jurgens et al., 2018). We insert a special mark, <CITED HERE>, into the position of the citation in the context and apply mean pooling on the embeddings of the corresponding tokens as the representation of the citation to feed into MLPs (see an example in Figure 1). Our experiments in Section 4.7 show that the CITED HERE readout function performs better than the standard position-agnostic readout functions.

Dataset	# instances	# papers [1]	# labels
ACL	1904	186	6
KIM	804	614	3
SciCite	11020	6627	3

Table 1: Statistics of the datasets

4 Experiments

4.1 Datasets

We conduct experiments on three datasets: **ACL** (Jurgens et al., 2018), **SciCite** (Cohan et al., 2019), and a newly curated dataset called **KIM**. The ACL dataset was collected from the ACL Anthology Reference Corpus and consists of natural language processing papers. The SciCite dataset is the largest one we consider and contains citations from general computer science and medical domains. The papers are collected from the Semantic Scholar Open Research Corpus (Lo et al., 2020). Detailed statistics and descriptions of the datasets are shown in Table 1 and Table 4.

KIM Dataset. Despite the fact that ACL and SciCite are widely used as CII benchmarks, their focus

[1] Number of unique citing papers.

PLM	Method	ACL	KIM	SciCite
	Scaffolds	67.90	-	84.00
GPT4	ICL 0-shot	38.55	33.86	72.86
	ICL 5-shot	50.18	60.55	74.55
BERT	Default	57.44	57.30	83.46
	Ours (Search)	65.98	64.18	84.08
	Ours (TRL)	66.32	62.00	83.48
SciBERT	Default	67.25	60.27	85.22
	Ours (Search)	73.74	63.11	85.25
	Ours (TRL)	**75.57**	**64.56**	**85.35**

Table 2: Performance (Macro-F1) of the MTL fine-tuning approach compared to the baseline methods. Search indicates grid search while TRL indicates our proposed task relation learning method. Results are averaged over five runs, the best performing method for each primary dataset is in **bold**, and the second best results are underlined.

on the fields of computer science and medicine renders them insufficient for building CII models applicable to other scientific domains. In this paper, we curate a new CII dataset in the field of materials science called KIM. The KIM dataset was constructed by collecting the primary citations for interatomic models archived in the OpenKIM repository[2] (Tadmor et al., 2011), retrieving as many papers as possible from the literature[3] that cite any one of them, and extracting the associated citation context(s). Domain experts in the field of materials simulation, examined the titles and abstracts of the citing and cited papers, as well as the citation context(s) appearing in the citing paper, and assigned one of three distinct labels to each citation. The "Used" label indicates that the citing paper directly used the model described in the cited paper. The "Extended" label indicates that the citing paper used a modified form of the model, e.g. it used the same functional form of the original model but modified its parameters or added an additional term. Finally, the "Not Used" label indicates that the citing paper does not directly utilize the model described in the cited paper, but rather cites it to provide general scientific background. We will open-source the KIM dataset upon acceptance.

[2]openkim.org

[3]Many papers were acquired from the Semantic Scholar Open Research Corpus, while some were obtained directly by the authors through their institutional licenses.

4.2 Baselines

We compare our methods with **Scaffolds** (Cohan et al., 2019), **BERT** (Devlin et al., 2019), **SciBERT** (Beltagy et al., 2019), and **GPT4** (Bubeck et al., 2023). Scaffolds is the state-of-the-art RNN-based CII method that does not rely on PLMs. We report Scaffolds' results from the original paper for comparison as we use the same train-test split of the ACL and the SciCite dataset. BERT and SciBERT are transformer (Vaswani et al., 2017) models pre-trained on general domain text corpus and scientific literature, respectively. For BERT and SciBERT as baselines, we follow the setting from their original papers and use the output embedding of the CLS token as the context representation.

We evaluate zero-shot and few-shot ICL performance on one of the most capable generative PLMs, GPT4. We prompt GPT4 with a detailed text description of the CII task, definition of the intentions, and/or a few examples. For few-shot experiments, we follow the common practice in ICL (Brown et al., 2020) and randomly select five examples for each intention class as examples. Details about the prompts are shown in Section C.

4.3 Experimental Settings

We evaluate our proposed multi-task learning (MTL) framework, the task relation learning (TRL) method, and the position-aware CITED HERE readout function on two backbone PLMs, BERT and SciBERT. We employ two methods to compute the value of the aforementioned λ coefficients: grid search and the TRL method. For grid search, we explore values in the range 0.1 to 1.0 in increments of 0.1 using the validation set of the primary dataset. For the TRL method, we fine-tune one PLM on an auxiliary dataset and evaluate its information gain on the training set of the primary dataset to compute a λ that is associated with the PLM-primary-auxiliary triplet. This λ is used for jointly fine-tuning the PLM that computes it on the primary-auxiliary datasets. For fine-tuning a PLM on a primary dataset with more than one auxiliary dataset, we use the λs associated with the PLM-primary-auxiliary triplets, respectively. In all MTL experiments involving BERT and SciBERT, CITED HERE is the default readout function unless stated otherwise. A detailed experimental setting is shown in Section D

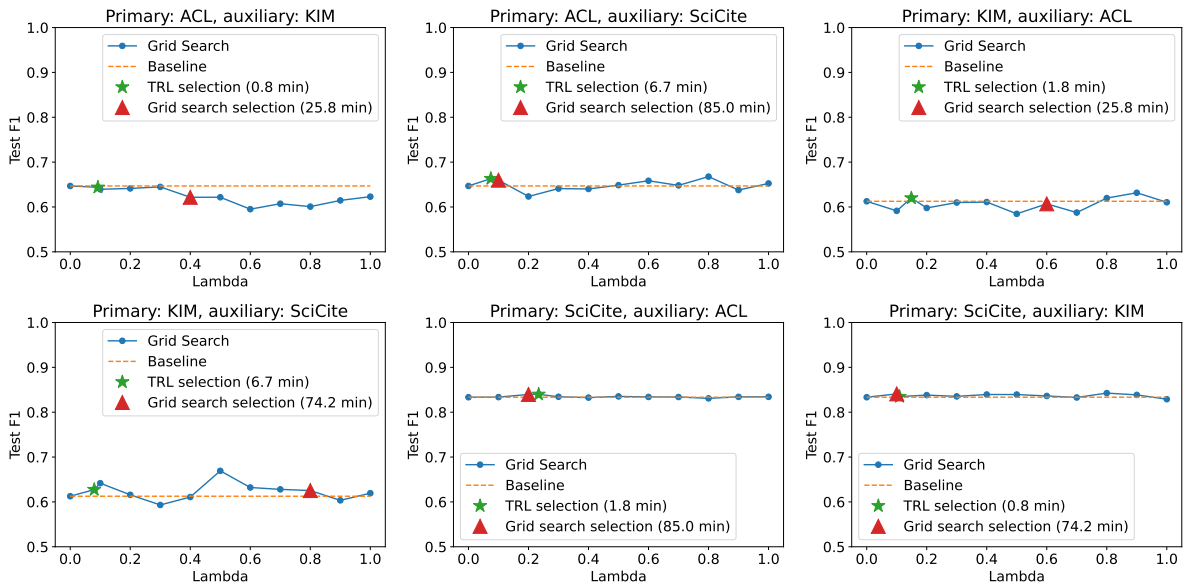

Figure 2: BERT performance of all binary combinations of primary and auxiliary datasets with different value of λs. The yellow line denotes the baseline performance of fine-tuning on only the primary dataset. The blue line denotes the performance of fine-tuning the primary and the auxiliary dataset with different λs. The star and the triangle indicate the λ found by our TRL method and the grid search method, respectively. Time in the brackets indicates the GPU time needed for the method.

4.4 Results

The main results of our experiments can be found in Table 2. On the two small benchmark datasets, ACL and KIM, our proposed fine-tuning framework significantly improves the backbone PLMs' performance compared to the default fine-tuning process. In particular, our method outperforms the current state-of-the-art by 7% on the KIM dataset and by 11% on the ACL dataset. On SciCite, our framework perform on par with the best-performing baselines. Our method outperforms the zero-shot and few-shot ICL methods on GPT4 by a significant margin. This demonstrates that, despite the increasing reasoning capability of LLMs, in the CII application with a few thousands training instances, our methodology is still necessary.

4.5 Multi-Dataset Fine-tuning

We explore all combinations of primary and auxiliary datasets and show the results in Table 3. In most of the cases, jointly fine-tuning PLMs on primary datasets with auxiliary datasets using our proposed MTL framework improves the PLMs' performance on the primary datasets. In cases when no improvement is achieved, jointly fine-tuning performs on par with fine-tuning the primary dataset by itself. However, adding more auxiliary dataset does not always lead to more improvement. In the table, we also observe a few cases (e.g., Primary: KIM, Auxiliary: SciCite, PLM: SciBERT) where adding auxiliary datasets degrades the PLMs' performance on the primary dataset. While auxiliary datasets can bring additional knowledge to PLMs, they may also cause distribution shift that degrades the performance of the PLMs.

4.6 Task Relation Learning

We investigate the effectiveness of our proposed TRL method for identifying the value of the λ coefficients. In Figure 2, we show BERT's performance of all binary combinations of primary and auxiliary datasets with different value of λs. For each pair of primary-auxiliary datasets, we show the λ values identified by our TRL method and the grid search method. We observe that the choice of λ is critical to the joint fine-tuning performance. A poorly chosen λ could amortize the benefit of auxiliary datasets and even degrade the performance.

When the auxiliary dataset can lead to positive transfer, our TRL method can effectively identify λs that improve the performance of the primary datasets and it performs on par or better than those selected by the grid search method (e.g., Primary: KIM, Auxiliary: SciCite). On the other hand, when the primary dataset causes negative transfer, the TRL method chooses a small value of λ that avoids

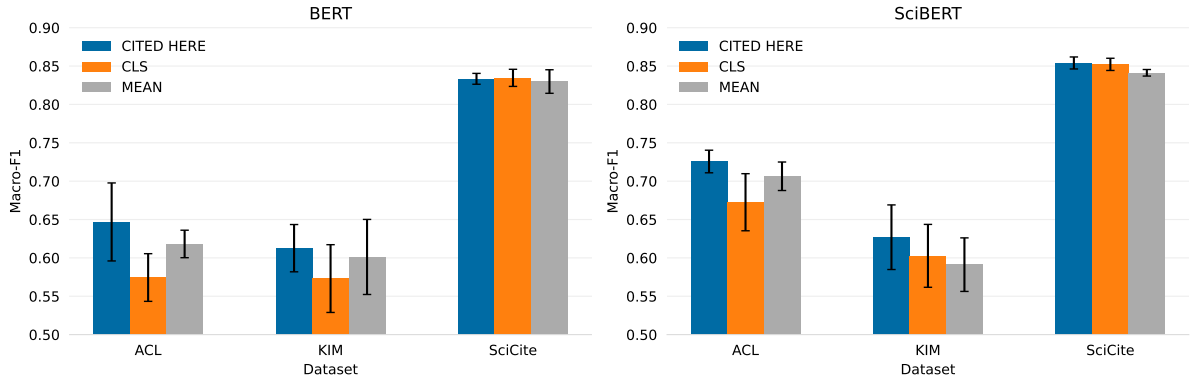

Figure 3: Performance (Macro-F1 with standard deviation) of the three readout functions: CLS, MEAN, and CITED HERE on BERT and SciBERT.

Primary	Auxiliary	BERT		SciBERT	
		Search	TRL	Search	TRL
ACL	-	64.68	64.68	72.56	72.56
	KIM	62.16	64.39	71.70	75.57
	SciCite	65.98	66.32	73.74	73.75
	KIM + SciCite	63.91	62.53	71.78	72.87
KIM	-	61.27	61.27	62.69	62.69
	ACL	60.70	62.00	63.11	63.44
	SciCite	63.21	62.74	58.88	63.26
	ACL + SciCite	64.18	62.98	61.86	64.56
SciCite	-	83.34	83.34	85.40	85.40
	ACL	83.96	83.98	85.25	85.35
	KIM	84.08	83.48	84.94	84.86
	ACL + KIM	84.34	84.00	85.43	84.55

Table 3: Performance (Macro-F1) of the MTL fine-tuning approach with different combinations of primary and auxiliary datasets and different methods of obtaining the value of λ. Search indicates grid search while TRL indicates our proposed task relation learning method. We group the results by the primary dataset. Baseline results of each primary dataset are shown in the first row of each group.

performance degradation on the primary dataset (e.g., Primary: ACL, Auxiliary: KIM). We reach similar conclusions on SciBERT and show the analysis in Section B.

Note that, the TRL method is also significantly faster compared to the grid search method, exhibiting a factor of 10 to 100 in run time improvement when only one auxiliary dataset is used. This advantage is more significant when dealing with more auxiliary datasets.

4.7 Readout Function

We compare the performance of the three readout functions, CLS, MEAN and CITED HERE and show the results in Figure 3. On the two small datasets ACL and KIM, the CITED HERE readout function significantly outperforms the other two readout functions. On the SciCite dataset, CITED HERE matches the performance of CLS and MEAN on BERT and slightly outperforms them on SciBERT.

This demonstrates that the position of the citation in the citation context is informative for the prediction of the its intention. Our proposed position-aware readout function, CITED HERE, has an edge over position-agnostic readout functions especially on small datasets. On the other hand, this inductive bias can be learned from the data when the dataset is large enough.

5 Conclusion

In this paper, we propose a multi-task learning (MTL) framework to fine-tune pretrained language models (PLMs) for citation intention identification (CII) tasks. Our framework treats additional CII datasets as auxiliary tasks to be jointly trained with a CII dataset that is of primary interest. We develop an efficient, data-driven task relation learning (TRL) method that controls the contribution of auxiliary datasets to avoid negative transfer. The proposed TRL method effectively identifies a coefficient that is critical to the performance of the MTL framework with magnitudes lower computational cost compared to grid search. We introduce a position-aware readout function and demonstrate that a citation's position within the context is informative for predicting its intention. Experimental results suggest that jointly fine-tuning PLMs on primary and auxiliary datasets with our proposed MTL framework effectively improves the their performance on the primary datasets.

6 Limitations

In this paper, we experiment with one way of finetuning PLMs, i.e., finetuning the pretrained PLM encoder with randomly initialized task specific multi-layer perceptrons while the multi-task learning (MTL) framework and the task relation learning (TRL) method proposed in this work are supposed to be applicable to any classification models. In the future, we will extend our study to different PLM finetuning techniques such as soft-prompt tuning (Lester et al., 2021) and adaptor-based finetuning (Houlsby et al., 2019; Hu et al., 2021) and different CII models such as GraphCite (Berrebbi et al., 2022).

7 Ethical Statement

Our work aims to improve the accuracy of citation intention identification (CII) tools that assist readers in comprehending scientific literature and evaluate the relevance and contribution of scientific publications. The method could be extended to other classification tasks. The datasets we used in this work are generated from scientific literature that are accessible through scientific publishers or pre-print servers. We believe our work should not raise any ethical concerns.

References

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 596–606, Atlanta, Georgia. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. 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 3615–3620.

Dan Berrebbi, Nicolas Huynh, and Oana Balalau. 2022. Graphcite: Citation intent classification in scientific publications via graph embeddings. In *Companion Proceedings of the Web Conference 2022*, pages 779–783.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Kaidi Cao, Jiaxuan You, and Jure Leskovec. 2022. Relational multi-task learning: Modeling relations between data and tasks. In *International Conference on Learning Representations*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

David Eigen and Rob Fergus. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658.

Myriam Hernández-Alvarez and José M Gomez. 2016. Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering*, 22(3):327–349.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

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

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames.

Transactions of the Association for Computational Linguistics, 6:391–406.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Iasonas Kokkinos. 2017. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6129–6138.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Lukas Liebel and Marco Körner. 2018. Auxiliary tasks in multi-task learning. *arXiv preprint arXiv:1805.06334*.

Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. 2019. Pareto multi-task learning. *Advances in neural information processing systems*, 32.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Saurav Manchanda and George Karypis. 2021. Evaluating scholarly impact: Towards content-aware bibliometrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6041–6053, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xia Ning, Huzefa Rangwala, and George Karypis. 2009. Multi-assay-based structure- activity relationship models: improving structure- activity relationship models by incorporating activity information from related targets. *Journal of chemical information and modeling*, 49(11):2444–2456.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Bernhard Schölkopf and Alexander J Smola. 2002. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China,*

October 18–20, 2019, Proceedings 18, pages 194–206. Springer.

Ellad B Tadmor, Ryan S Elliott, James P Sethna, Ronald E Miller, and Chandler A Becker. 2011. The potential of atomistic simulations and the knowledgebase of interatomic models. *Jom*, 63(7):17.

Fei Tao and Carlos Busso. 2020. End-to-end audiovisual speech recognition system with multitask learning. *IEEE Transactions on Multimedia*, 23:1–11.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. 2019. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11293–11302.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645.

A Intention Definitions

We show the definition of the intentions of the three CII datasets we used in our experiments in Table 4. Although the label spaces of the three datasets are different, they contain semantically similar and shared intention categories. For example, "Background" exists in both the ACL-ARC and the SciCite dataset. "Method" in the SciCite dataset and "Used" in the KIM dataset are semantically similar to each other.

B Additional Experimental Results

In Figure 4, we show SciBERT's performance of all binary combinations of primary and auxiliary datasets with different value of λs. We have similar observations to the results for BERT. The fine-tuning performance is sensitive to the choice of λs. Our TRL method can effectively identify λs that improve the performance of the primary datasets and performs on par or better than those selected by the grid search method.

C In-Context Learning Prompts

We present detailed examples of the zero-shot and few-shot prompts that we use for in-context learning on GPT4. We access to GPT4, specifically gpt-4-0125-preview, through the OpenAI API.

C.1 Zero-shot Prompt

```
I want you to act as a research
assistant with expertise in atom-
istic modeling.  I will provide
you with a piece of text from a
scientific paper that cites an-
other paper.  You will classify
the text into one of the follow-
ing labels that indicate the in-
tention of the citation:  [Back-
ground, Method, Result].  The
labels are defined as
  "Background":  "The citation
states, mentions, or points to
the background information giv-
ing more context about a problem,
concept, approach, topic, or im-
portance of the problem that is
discussed in the present paper."
  "Method":  "The present paper
uses a method, tool, approach or
dataset that is proposed in the
paper associated with the cita-
tion."
  "Result":  "The present paper
compares its results/findings
with the results/findings of the
paper associated with the cita-
tion."
  You will only respond with the
predicted label.  Below is the
input text:
  We used an active contour algo-
rithm <CITED HERE> to segment the
organs from 340 coronal slices
over the two patients.
```

Dataset	Intention	Definition
ACL-ARC	Background	The citation provides relevant information for the domain that the present paper discusses.
	Motivation	The citation illustrates the need for data, goals, methods, etc that is proposed in the present paper.
	Uses	The present paper uses data, methods, etc., from the paper associated with the citation.
	Extends	The present paper extends the data, methods, etc. from the paper associated with the citation.
	Compare or Contrast	The present paper expresses similarity / differences to the citation.
	Future	The citation is a potential avenue for future work of the present paper.
KIM	Used	The present paper uses at least one method that is proposed in the paper associated with the citation.
	Not Used	The present paper does not use or extend any methods that is proposed in the paper associated with the citation.
	Extended	The present paper uses an extended / modified version of the method proposed in the paper associated with the citation.
SciCite	Background	The citation states, mentions, or points to the background information giving more context about a problem, concept, approach, topic, or importance of the problem that is discussed in the present paper.
	Method	The present paper uses a method, tool, approach or dataset that is proposed in the paper associated with the citation.
	Result	The present paper compares its results/findings with the results/findings of the paper associated with the citation.

Table 4: Definition of intentions.

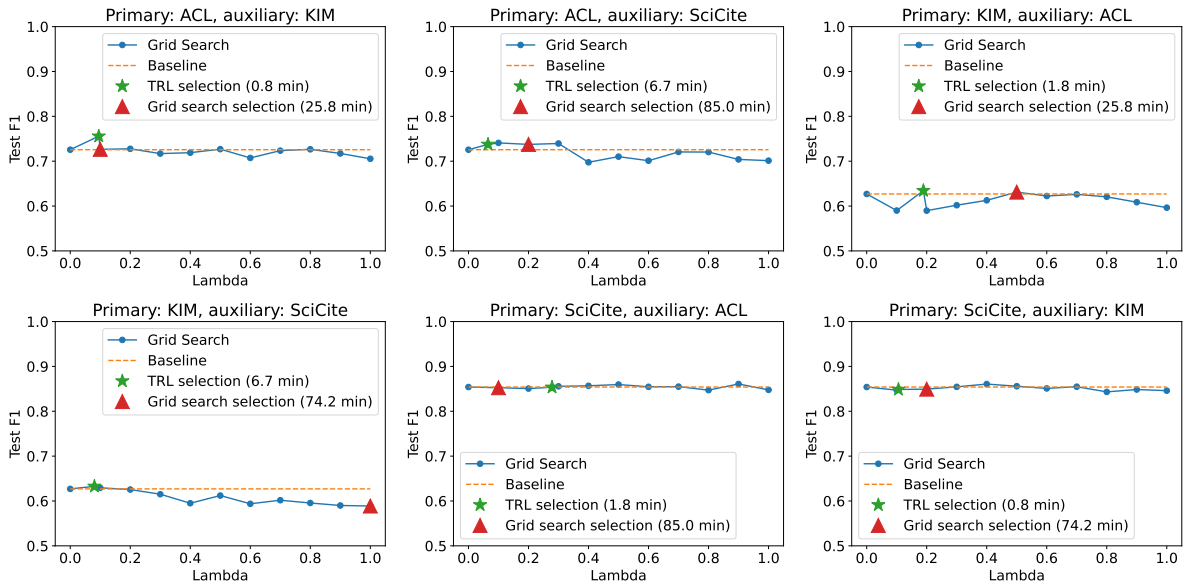

Figure 4: SciBERT performance of all binary combinations of primary and auxiliary datasets with different value of λs. The yellow line denotes the baseline performance of fine-tuning on only the primary dataset. The blue line denotes the performance of fine-tuning the primary and the auxiliary dataset with different λs. The star and the triangle indicate the *lambda* found by our TRL method and the grid search method, respectively. Time in the brackets indicates the GPU time needed for the method.

C.2 Few-shot Prompt

I want you to act as a research assistant with expertise in atomistic modeling. I will provide you with a piece of text from a scientific paper that cites another paper. You will classify the text into one of the following labels that indicate the intention of the citation: [Background, Method, Result]. The labels are defined as

"Background": "The citation states, mentions, or points to the background information giving more context about a problem, concept, approach, topic, or importance of the problem that is discussed in the present paper."

"Method": "The present paper uses a method, tool, approach or dataset that is proposed in the paper associated with the citation."

"Result": "The present paper compares its results/findings with the results/findings of the paper associated with the citation."

Here are some examples:

```
  Example:  We used an active con-
tour algorithm <CITED HERE> to
segment the organs from 340 coro-
nal slices over the two patients.
  Output:  Method
  Example:  The remnant of the
total plasma membranes after ex-
traction of caveolae is called
bulk plasma membranes <CITED
HERE> (Fig.
  Output:  Background
  Example:  More examples of con-
tradictory results have been ob-
served in bovines; some reports
<CITED HERE> indicated a signifi-
cant decrease in blastocyst
  Output:  Result
  You will only respond with the
predicted label.  Below is the
input text:
  Following <CITED HERE> and Koo
and Collins (2010), before train-
ing we transform the training set
trees to be the best achievable
within the model class (i.e.,
the closest projective tree or
1-Endpoint-Crossing tree).
```

D Experimental Settings

We use the Adam optimizer (Kingma and Ba, 2014) to minimize the cross-entropy loss in all our pre-training and fine-tuning tasks. The batch size is set to 32. For fine-tuning tasks, we set the learning rate to be 5e-5 and use a slated triangular scheduler (Howard and Ruder, 2018) to first warm up and then decrease the learning rate linearly. The model is fine-tuned for 10 epochs and evaluated on the validation set after every epoch. For the ACL and the SciCite datasets, we use the original train-test split and use 15% of the training set as validation set. For the KIM dataset, we randomly split the dataset into train, validation, and test sets with a 70%/15%/15% ratio. Test performance is reported on the checkpoint that performs the best on the validation set. We report the macro-F1 score as the evaluation metric. The macro-F1 scores that we report in this paper are averaged numbers of five independent runs. All experiments are conducted on a machine with an Intel(R) Core(TM) i9-10900F CPU and an Nvidia RTX 3090 GPU. Our methods and experiments are implemented using PyTorch (Paszke et al., 2019). For experiments including BERT and SciBERT, we use the implementation and pre-trained weights from the transformers library.[4]

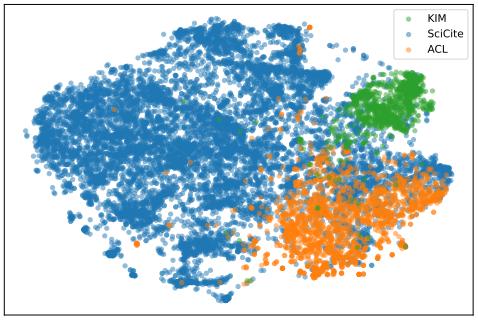

Figure 5: T-SNE visualization of the citation contexts in different datasets. Citation contexts are encoded by SciBERT using the CLS readout function.

E Visualization of Citation Contexts

In Figure 5, we show the t-SNE (Van der Maaten and Hinton, 2008) visualization of the citation contexts in the three datasets. We use SciBERT and the CLS readout function to convert the contexts to latent embeddings. We observe that the citation contexts of different datasets form into clusters because the fields of the papers are different, but there are significant overlaps between the clusters. It is reasonable to assume that the datasets share an input space.

[4] https://github.com/huggingface/transformers