Adapting Pretrained Language Models for Citation Classification via Self-Supervised Contrastive Learning

Tong Li
tlice@connect.ust.hk
Hong Kong University of Science and Technology
Kowloon, Hong Kong SAR, China

Jiachuan Wang
jwangey@connect.ust.hk
Hong Kong University of Science and Technology
Kowloon, Hong Kong SAR, China

Yongqi Zhang*
yongqizhang@hkust-gz.edu.cn
Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, Guangdong, China

Shuangyin Li
shuangyinli@scnu.edu.cn
South China Normal University
Guangzhou, Guangdong, China

Lei Chen
leichen@ust.hk
Hong Kong University of Science and Technology
Kowloon, Hong Kong SAR, China

Abstract

Citation classification, which identifies the intention behind academic citations, is pivotal for scholarly analysis. Previous works suggest fine-tuning pretrained language models (PLMs) on citation classification datasets, reaping the reward of the linguistic knowledge they gained during pretraining. However, directly fine-tuning for citation classification is challenging due to labeled data scarcity, contextual noise, and spurious keyphrase correlations. In this paper, we present a novel framework, Citss, that adapts the PLMs to overcome these challenges. Citss introduces self-supervised contrastive learning to alleviate data scarcity, and is equipped with two specialized strategies to obtain the contrastive pairs: sentence-level cropping, which enhances focus on target citations within long contexts, and keyphrase perturbation, which mitigates reliance on specific keyphrases. Compared with previous works that are only designed for encoder-based PLMs, Citss is carefully developed to be compatible with both encoder-based PLMs and decoder-based LLMs, to embrace the benefits of enlarged pretraining. Experiments with three benchmark datasets with both encoder-based PLMs and decoder-based LLMs demonstrate our superiority compared to the previous state of the art. Our code is available at: github.com/LITONG99/Citss

CCS Concepts

• **Computing methodologies** → **Information extraction**.

Keywords

Citation Classification, Pretrained Language Models, Contrastive learning

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference acronym 'XX, Woodstock, NY
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Tong Li, Jiachuan Wang, Yongqi Zhang, Shuangyin Li, and Lei Chen. 2018. Adapting Pretrained Language Models for Citation Classification via Self-Supervised Contrastive Learning. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In scholarly writings, citations act as intellectual bridges, linking researchers and their ideas across time and disciplines to provide a connected view of the scientific literature. The analytic study of citations is the cornerstone for understanding the structure, evolution, and impact of scientific contributions, attracting growing attention in recent years [30, 43]. One of the critical focuses is the citation classification, which identifies and categorizes the authors' intention of using citations in their writing, empowering a range of applications, including research evaluation [28, 59, 64], research trends identification [21, 59], paper recommendation [17, 53, 56], and scientific texts summarization [8, 25, 57].

For a specific citation, its surrounding textual context is essential for revealing the underlying intention behind its explicit mentioning. Traditional works in citation classification extract informative features from the contexts and then train supervised classifiers to assign labels [30]. Some research [28] relies on hand-engineered features involving the in-text cue words, metadiscourse, part-of-speech tags, dependency relationships, etc. Others [11, 54] also bring in sophisticated deep-learned features such as the word representations from popular embedding models [14, 47, 50]. Nevertheless, these methods struggle to capture the semantic nuances between different citation intentions due to their limited model capacity.

Recent strides in transformer-based pretrained language models (PLMs) offer significant opportunities for developing more effective citation classification systems. Plenty of studies [33, 55] have demonstrated that PLMs acquired extensive linguistic knowledge during pretraining, which can be fine-tuned on citation classification datasets to discern subtle semantic nuances in citation intentions. They leverage encoder-based PLMs, such as BERT [15], Longformer [4], SciBERT [3], to encode the context into representations, which is optimized end-to-end to automatically mine rich semantic information for the task. Despite their efforts in fine-tuning

PLMs on citation data, these methods are limited in addressing the following unique challenges and only consequent with sub-par performance.

Challenge 1: Scarcity of labeled citation data. It usually requires domain-specific knowledge and expertise of annotators to accurately interpret the scientific texts and assign citation labels. Hence, existing approaches either call on author self-annotation [34, 52] or employ annotators with relevant academic backgrounds [18, 28]. To date, publicly available citation classification datasets remain limited to a few thousand samples, which fall short in manifesting the task-specific textual patterns and restrain the performance of deep learning systems [30, 48]. With a great number of parameters, fine-tuning PLMs is more vulnerable confronted data scarcity.

Challenge 2: Defocusing on the target citation. Most existing methods [33, 41, 55] only work with a highly related but extremely local context, which is the sentence directly containing the citation, referred to as the *citance* [13, 57]. However, the necessary semantic clues that enable us to determine the type of citation can be far from the citation and even fragmented in the writing [32]. Although the larger input window of PLMs makes it possible to include such long-range dependencies, a broader context can also introduce excessive irrelevant descriptions, such as mentions of other citations and general discourses, which are likely to distract the model from the target citation [5] and even cause the problem of *lost-in-the-middle* [37].

Challenge 3: Spurious correlation based on keyphrases. The scientific keyphrases that are repeatedly mentioned in the texts, including research subjects, tasks, techniques, etc., are usually semantically significant for the PLMs, implying that the PLMs can easily establish spurious correlations from them to the label. For example, in the case where each citation context in the training data discusses a unique technique, the model can easily fit on the dataset by learning a toxic mapping from the keyphrases to the observed label, resulting in a corrupted model. This problem can be further intensified by the insufficient training data discussed in Challenge 1.

In this paper, we address the above challenges and propose a framework, Citss, that adapts pretrained language models for **Cit**ation classification via **s**elf-**s**upervised contrastive learning. For Challenge 1, our framework is equipped with two transform strategies, sentence-level cropping (SC) and keyphrase perturbation (KP), which generate contrastive pairs for citation classification in a self-supervised manner and derive contrastive loss to provide extra supervision signals for model fine-tuning, alleviating the demand for annotated citation contexts Given the original sample, each of the SC and KP strategies not only produces diverse and realistic artificial samples serving as its contrastive pairs, but is also deliberately aimed to enhance the ability of PLMs in facing Challenge 2 and Challenge 3 correspondingly. Specifically, SC facilitates the model to focus on the target citation and improve model robustness against irrelevant noises, and KP helps to mitigate the spurious correlation established between specific keyphrases and the observed label.

Besides developing our fine-tuning method specialized in elevating those previously highlighted encoder-based PLMs for citation classification, we are also ambitious to embrace the currently booming large language models (LLMs). With the number of parameters scaling up to billions, the advantages of pretraining are amplified for these LLMs [6] with decoder-based architecture, making it appealing to harness their power for citation classification. However, according to the latest attempts [33, 45], adopting the cutting-edge LLMs, such as GPT-3.5-turbo, GPT-4, and SciGPT2 [40], for citation classification in a language generation style still falls behind the fine-tuned "small" encoder-based PLMs, leaving the question of how to benefit from LLMs on citation classification open. In this regard, we establish our framework carefully so it can not only be applied to the decoder-based architectures of LLMs but also seamlessly incorporated with the prevalent parameter-efficient fine-tuning (PEFT) paradigms, such as Lora [63], to further reduce the trainable parameters in LLMs. With our framework, we successfully fine-tuned a Llama3-8B backbone on the existing limited citation classification data and achieved noticeable improvements compared with baselines.

To summarize, our main contributions are as follows

- We propose a novel self-supervised contrastive learning framework that is applicable to fine-tuning both encoder-based PLMs and decoder-based LLMs[1] for citation classification under the scarcity of labeled citation data. To the best of our knowledge, this is the first work to effectively fine-tune LLMs for citation classification.
- We propose a sentence-level cropping strategy that enhances the ability of PLMs to extract beneficial information for the target citation from long contexts and defend against irrelevant noises.
- We propose a keyphrase perturbation strategy that assists the PLMs in predicting the citation intentions based on the context logic rather than the occurrence of specific keyphrases.
- Experiments on three datasets with both an encoder-based PLM and a decoder-based LLM demonstrate the consistent superiority of our framework.

2 Related Work

The analytical study of citations boasts a long scholarly history. Within this domain, the citation classification task we investigate belongs to natural language processing grounded in citation context analysis [23]. In the literature on related data mining research, this task has been specifically designated as either "citation function classification" [18, 28, 30] or "citation intent classification" [5, 11, 54, 55, 58]. While some scholarly works posit distinctions between these two concepts, in a sense that the former adopts an objective perspective centered on how citations serve scholarly writing, whereas the latter emphasizes authors' subjective psychological processes [44], we observe that their classification schemas are frequently identical or substantially overlapping. Consequently, this study refrains from further differentiation between these terminologies and collectively refers to them as "citation classification."

Other tasks that categorize the citations based on the contexts include: polarity classification [62], which identifies authors' sentiment stance toward cited content; influence classification [31, 64], which seeks to define the significance of a cited work and recognize the influential citations; role classification [66], which discerns the types of resources (code, data, website, media, etc.) provided by the

[1]We use "PLMs" to collectively refer to both types.

citation link. These endeavors primarily serve specific application objectives, and their classification schemas exhibit substantial divergence from the research focus of this study, with marked differences in problem characteristics.

Additional citation analytical tasks may not necessarily originate from citation context analysis. For instance, citation prediction [12, 20, 27, 65] aims to forecast potential citations from a candidate paper collection for the target document. Likewise, citation recommendation [22] attempts to recommend scholarly references to the authors during the paper drafting. These analytical tasks differ more profoundly from the present research, primarily in that they focus on the stages where formal academic texts and actual citations may not yet exist.

3 Preliminary

3.1 Citation Classification

For the target citation i, its *citation anchor* [1, 2] in the text indicates the occurrence of the citation, which can be in various formats (numerical, author-with-year, etc.) depending on the writing conventions. The citation context T_i is the textual content surrounding the citation anchor [23]. As shown in Figure 1 (b), we replace the citation anchor with a special tag (#CITATION_TAG) to distinguish it from other citations. Given the types of citation intentions C, our task takes T_i as the input and outputs the predicted label $y_i \in C$.

3.2 Pretrained Language Models

From a high level, the interaction between our framework and the backbone PLM can be regarded as a function,

$$x = \mathcal{M}(T, \mathbf{P}), \tag{1}$$

where x is the output vector, T is the citation context, and $\mathbf{P}$ is a textual prompt. $\mathbf{P}$ contains possible meta-information of the citation classification tasks, and defines the format in which T is presented to the PLM, in order to help the model adapt to the specific task. The formatted input will be converted into a series of tokens and forwarded through the stacked transformer layers of the backbone $\mathcal{M}$. Correspondingly, there will be a series of hidden state vectors at the last layer, and we retrieve x from them in different ways for encoder-based PLMs and decoder-based LLMs, depending on their distinct characteristics of language modeling. (1) The **encoder-based PLM** consists of an encoder that is pre-trained with the blank infill task (masked language modeling). They are the leading models for numerous supervised text classification tasks [7], among which the SciBERT [3] is optimized for scientific text and pertinent to citation classification. Because these models are skilled at reconstructing information at the masked input token, we insert a mask token in the prompt and read out x at the masked position. (2) The **decoder-based LLM**, such as GPT-4, Llama3-8B, and Llama3-70B, achieves groundbreaking performance recently in various natural language understanding task [6]. Unlike traditional classifiers, these models usually leverage language generation for text classification through question-answering or instruction-based paradigms [45]. Architecturally, they employ decoder-only structures pretrained via next-token prediction objectives [9]. Their output hidden state in the last position contains information for generating the next token, and is authentically intended to be decoded into a textual response. In our framework, instead of decoding, we incorporate the novel "Explicit-One-word-Limitation" trick (EOL) [26], which compels the model to condense all contextual understanding into the immediate next token's representation by explicitly instructing it to output exactly one more word in the prompt. We will then read x out at the last position.

As the interaction with PLMs formulated as Equation 1, our methodology only cares about how to prompt for the backbone model and to obtain the x properly but is indifferent to the inner architectures. Therefore, our framework can work together with those PEFT methods [19], which select, reparameterize [63], or insert [24] trainable modules in the PLMs to reduce trainable parameters while keeping the input and output interface unchanged, making it possible to efficiently finetune the LLMs with the limited citation data.

3.3 Self-Supervised Contrastive Learning

Contrastive learning [10, 46] generates additional supervision signals by utilizing the similarity between data samples. It first defines the *contrastive pairs*, i.e., positive pairs and negative pairs, then the contrastive loss forces the positive pairs to be similar in the latent space, while the negative pairs to be dissimilar, encouraging the model to capture invariant information between positive pairs as well as distinguishable information between individual negative pair. There are many ways to construct contrastive pairs. In a self-supervised manner [61], carefully designed strategies can be employed to make modifications on the original sample T_i and transform it into its positive pair $\tilde{T}_i$. Other transformed samples in the batch, $\{\tilde{T}_{j\neq i} | j \in \mathcal{B}\}$, will serve as the negative pairs. With z_i and $\tilde{z}_i$ denoting the representations of T_i and $\tilde{T}_i$ in the latent space, the classic InfoNCE loss [46] for the batch can be written as

$$L^{InfoNCE} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp sim(z_i, \tilde{z}_i)}{\sum_{j \in \mathcal{B}} \exp sim(z_i, \tilde{z}_j)}, \tag{2}$$

where $sim(q, k) = q^\top k / \tau$ is the similarity metric and τ is the temperature hyperparameter that adjusts the strength of contrast.

In order to benefit from the contrastive learning for citation classification, it is substantial to develop proper transformation strategies tailored to the task. On the one hand, indicative information about the citation intention contained in the original sample is supposed to be preserved after the transformation. On the other hand, the transformation needs to introduce sufficient input diversity to effectively guide the model to gain discriminative ability in desired aspects.

4 Methodology

4.1 Framework Architecture

An overview of Citss is depicted in Figure 1 (a). In this section, we will first describe the framework architecture to show how to obtain context representations and create extra supervision signals by contrastive learning. Then, we will discuss the sentence-level cropping strategy and explain why it helps the model to focus on the target citation against contextual noises. Next, we will elaborate on the definition and algorithm of the keyphrase perturbation strategy and how it helps to mitigate the spurious correlations. Finally, we derived the complexity of our framework.

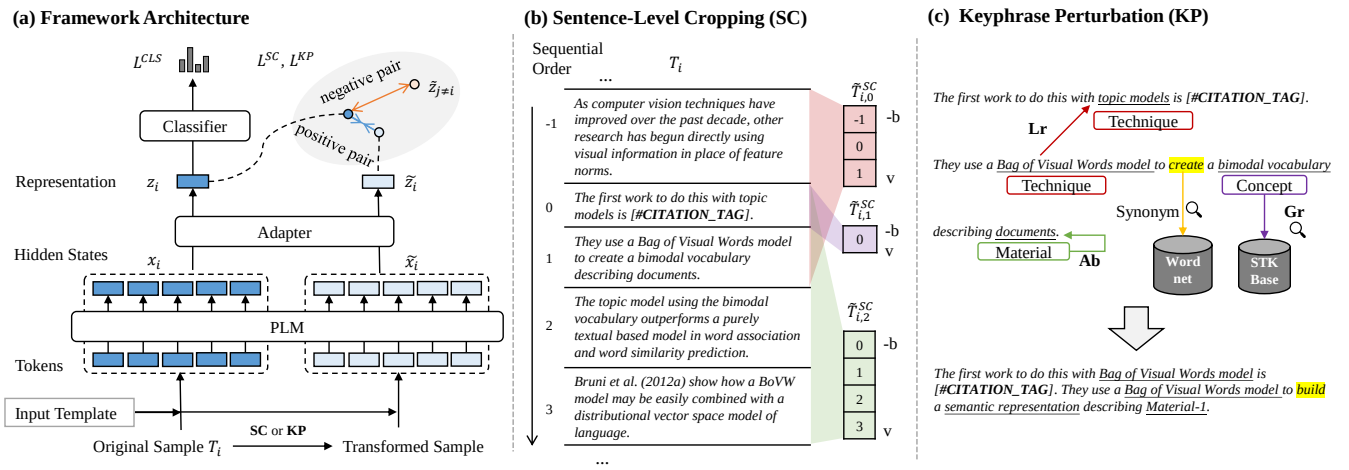

Figure 1: Overview of Citssframework: (a) exhibits the architecture and workflow. (b) shows an example of sentence-level cropping. (c) shows an example of the keyphrase perturbation. The underlined text is the keyphrases and the shaded text is the word for synonym replacement.

Given the context T_i of sample i, each of the two strategies, sentence-level cropping (SC) and keyphrase perturbation (KP), will modify it into the corresponding transformed sample $\tilde{T}_i$. T_i and $\tilde{T}_i$ are input into the PLM separately to obtain the hidden state vector, $x_i = \mathcal{M}(T_i, \mathbf{P})$, and $\tilde{x}_i = \mathcal{M}(\tilde{T}_i, \mathbf{P})$.

The output vectors directly from the PLM comply with the intrinsic distribution of the hidden states, thereby limiting in characterizing task-specific information for citation classification. Therefore, we introduce a lightweight MLP adapter module to map them into task-specific context representation. The adapter is also able to perform dimension reduction, so that the subsequent modules can compute the similarity between contrastive pairs in an appropriate latent space. Formally,

$$f(x) = W_2\mathsf{LN}(\mathsf{GeLU}(W_1 x + b_1)) + b_2, \tag{3}$$

where $\mathrm{LN}(\cdot)$ is the layer normalization, $\mathrm{GeLU}(\cdot)$ is the activation function, and W_1, W_2, b_1, b_2 are learnable parameters. The representation of the original context $z_i = f(x_i)$ is sent to a linear classifier for prediction, and calculate the multi-class classification loss.

$$y_i = g(z_i) = \mathrm{softmax}(W_3 z_i + b_3), \tag{4}$$

$$L^{CLS} = \sum_{i \in \mathcal{B}} CrossEntropy(y_i, \hat{y}_i), \tag{5}$$

where $\mathcal{B}$ is the batch index set and W_3, b_3 are learnable parameters.

As for the representations of transformed contexts $\tilde{z}_i = f(\tilde{x}_i)$, they are only used for contrastive learning. Under each strategy, we compute the the InfoNCE loss L^{SC} or L^{KP}, so the overall optimization target is

$$L = L^{CLS} + \lambda_1 L^{SC} + \lambda_2 L^{KP} + \omega L^{pnt}, \tag{6}$$

where $\lambda_1, \lambda_2, \omega$ are hyperparameters controlling the magnitude of the loss terms, and L^{pnt} is the weight decay penalty loss [39] for overfitting prevention. It is worth mentioning that the transformations and contrastive learning are only conducted for the fine-tuning stage to aid parameter learning. During inference, we collect the citation context and forward it sequentially through modules for prediction.

4.2 Sentence-Level Cropping

The sentence-level cropping helps the model focus on the target citation with the presence of contextual noises, and its intuition resembles image cropping [10, 16] in computer vision, which guides the model to attend on the target object against background noises.In the textual context, we crop on the sentence level because a sentence is a unit to convey a complete thought in natural language.

As illustrated in Figure 1 (b), SC splits the long context into a sequence of sentences. In every epoch, it randomly crops a subsequence with the same citance but different context ranges. Minimizing the contrastive loss imposes the representation to be correlated with the target citation regardless of the randomly allocated input range, thus encouraging the model to focus on the target citation within long input. Rather than only containing the citance, each positive pair may also randomly overlap on a portion of surrounding sentences with the original sample, so the representation immediately after this optimization step tends to encode such random contextual information. Supposing the information contributes to the correct prediction, it is likely to lead to better classification loss in the following steps, and such changes will be retained during optimization across epochs. Otherwise, if the information disturbs the correct prediction, the instant influence of this step will be counteracted by other steps after epochs. Therefore, the SC contrastive loss across epochs assists the model in dynamically extracting valuable contextual information that enhances the prediction while get rid of the noises.

Denoting the citance for sample i as s_i^0, the input of Citss is an enlarged context $T_i = \langle s_i^{-l}, ..., s_i^{-1}, s_i^0, s_i^1, ..., s_i^l \rangle$, representing a sequence of $2l + 1$ sentences where l is a hyperparameter defining the maximum one-side range. SC produces contexts with perturbed

ranges,

$$\mathcal{SC}(T_i) = \{\langle s_i^{-b}, ..., s_i^0, ..., s_i^v\rangle | \forall b, v, -l \le -b \le 0 \le v \le l\}/\{T_i\}.$$

Each transformed sample is comprised of the citance, b preceding sentences, and v succeeding sentences. SC can produce at most $(l+1)^2-1$ transformed samples and the minimum resultant context is $\langle s_i^0 \rangle$. Among them, we repeatedly iterate to obtain the positive pair for the current epoch e, $\tilde{T}_{i,e}^{SC} \sim \mathcal{SC}(T_i)$. With temperature hyperparameter τ_1, the contrastive loss is

$$L^{SC} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp(z_i^\top \tilde{z}_{i,e}^{SC}/\tau_1)}{\sum_{j \in \mathcal{B}} \exp(z_i^\top \tilde{z}_{j,e}^{SC}/\tau_1)}. \tag{7}$$

4.3 Keyphrase Perturbation

The intuition of the keyphrase perturbation strategy is that modifying the scientific keyphrases in the context usually does not affect the citation intention. The scientific keyphrases are associated with detailed topics, while the residue of the sentence organizes the knowledge and outlines the writing logic, playing distinct roles in scientific writing. Hence, the residue is usually more dominant in deciding the citation intention, an example is as follows.

Example 4.1. Here are two texts, **S1**: "The first work to do this with topic models is [1].", and **S2**: "We use topic models [1] to find hidden semantic structures in documents." They are similar if we consider general semantics because they are both written around the technique "topic models". However, for citation classification, the label of **S1** is "BACKGROUND" while the label of **S2** is "USE", since they express totally distinct logic regarding the intention of the target citation. Further, consider **S3**: "The first work to do this with data augmentation is [1].", which is an artificial text created by altering the technique in **S1**. **S3** is quite plausible to appear in a paper on the topic "data augmentation" and is very likely to belong to the same citation label "BACKGROUND" as **S1**.

As illustrated in Figure 1 (c), KP recognizes and changes the scientific keyphrases within the original context to construct its positive pairs with the same residue. Through subsequent contrastive learning, this encourages the representation to model the logical semantics of sentence residue instead of specific keyphrase semantics.

We first standardize the definition of keyphrases in our research as the scientific typed keyphrases (STKs) [35], which are the entities that have indispensable semantic meaning in the scientific domain, including specific proper nouns such as "BERT" and significant words or phrases in the text, such as "citation classification". Analogized to the named entities [49] in the general domain which can be typed as person, organization, and location, STKs also come with types, such as task, technique, materials, and concept. In this work, we will use the terms keyphrase and STKs interchangeably. There are several tools and methods for STK extraction under the supervised setting or few-shot setting [35]. Here we consider the simple one-shot setting so as to reduce manual efforts to label for this intermediate task, and leverage the instructional generation to extract the STKs with LLMs. This training-free approach performs well even under the one-shot setting, thanks to the great generalization ability of LLMs.

Algorithm 1 Keyphrase perturbation for epoch e

Input: K_i, T_i, Op_e, β, γ, synonym base $\mathcal{SN}$, global STK base $\mathcal{K}$

$\tilde{T}_{i,e}^{KP} \leftarrow T_i$
for $k \in K_i$ **do**
  Sample indicator $q \sim Bernoulli(\beta)$
  if $q = 1$ **then**
    Perturb k in $\tilde{T}_{i,e}^{KP}$ with Op_e
  end if
end for
Split T_i into word list $\mathcal{W}$ and eliminate the stop words.
for $w \in \mathcal{W}$ and w not in any $k \in K_i$ **do**
  if w has synonyms in $\mathcal{SN}$ **then**
    Sample indicator $q \sim Bernoulli(\gamma)$
    if $q = 1$ **then**
      Sample w' from $\mathcal{SN}[w]$
      Replace w in $\tilde{T}_{i,e}^{KP}$ with w'
    end if
  end if
end for

Output: $\tilde{T}_{i,e}^{KP}$

Formally, for the training set $\mathcal{D}$, let K_i be the STKs appeared in T_i and $\mathcal{K} = \bigcup_{i \in \mathcal{D}} K_i$ be the set of observed STKs. KP perturbs each $k \in K_i$ at a predefined probability β. We consider the following three different operations to perturb k, which will be invoked periodically at different epochs.

- **G**lobal **r**eplacement. The mention of k is replaced by another $k' \in \mathcal{K}/\{k\}$. We additionally require k and k' to have the same type because different types of STKs tend to function differently in the context.
- **L**ocal **r**eplacement. The mention of k is replaced by $k' \in K_i/\{k\}$ of the same type. It is a localized version of **Gr** that only allows replacement between keyphrases occurring in the same context, ensuring k' and k' are semantically relevant to each other.
- **Ab**straction. The mention of k is replaced by its type name, such as "Task-1" and "Technique-2". We add the extra numerical IDs to distinguish between perturbed keyphrases of the same type. This operation masks out k by making it anonymous without introducing new keyphrases.

After the perturbing STKs, KP also performs synonym replacement [60] on the residue of the context to introduce semantic diversity in the general domain. Particularly, each word except for the stop words is perturbed at predefined probability γ and we use the WordNet [42] synonym base $\mathcal{SN}$ to query for the synonyms. This step introduces general semantic differences between the positive pairs, in case there are few or no STKs in the context. Let the scheduled perturbation operation for epoch e is $\text{Op}_e \in \{\textbf{Gr}, \textbf{Lr}, \textbf{Ab}\}$, the overall KP algorithm to generate $\tilde{T}_{i,e}^{KP}$ is in Algorithm 1. With temperature hyperparameter τ_2, the contrastive loss for epoch e is

$$L^{KP} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp(z_i^\top \tilde{z}_{i,e}^{KP}/\tau_2)}{\sum_{j \in \mathcal{B}} \exp(z_i^\top \tilde{z}_{j,e}^{KP}/\tau_2)}. \tag{8}$$

4.4 Complexity Analysis

Denote the number of trainable parameters in $\mathcal{M}$ as N_0, the hidden state size of $\mathcal{M}$ as $d_x = |x|$, the model size as $d_z = |z|$, the intermediate embedding size of the adapter as d, $C = |C|$ is the size of label set, the number of trainable parameters of Citssis $N_0 + d(d_x + d_z) + d + d_z + C(d_z + 1)$. With $C \ll d_z$ and $d_z < d_x$, the number of parameters is in $O(N_0 + 2d \cdot d_x)$. Additionally, our method can work seamlessly with prevalent parameter-efficient finetuning methods for LLM, such as Lora, which wraps $\mathcal{M}$ and reduces N_0 without changing its interface with our framework. As for time complexity, since SC only shortens the T_i and KP performs replacement between synonyms or STKs, the length of $\tilde{T}_i$ after tokenization resembles that of T_i, implying similar forwarding complexity. Assume the PLM forwarding complexity per sample is $O(|\mathcal{M}|)$, the amortized complexity to perform contrastive learning amongst the batch is $O(|\mathcal{B}| \cdot d_z^2)$, hence the training complexity of our framework per sample is $O(3|\mathcal{M}|+6d \cdot d_x+2|\mathcal{B}| \cdot d_z^2) \approx O(3|\mathcal{M}|)$ and the inference complexity is $O(|\mathcal{M}| + 2d \cdot d_x) \approx O(|\mathcal{M}|)$. Here the time for for SC and KP can be neglected, because with l' be the length of T_i, it takes $O(l')$ time to scan the context and generate a transformed sample, while $|\mathcal{M}|$ is generally in $O(l'^2)$.

5 Experiments

We conduct experiments to investigate the proposed framework and answer the following research questions.

- **RQ1**: How is the overall performance of our framework?
- **RQ2**: How effective is the proposed SC strategy in extracting contextual information and defending the irrelevant noise?
- **RQ3**: How effective and robust is the proposed KP strategy?

5.1 Setup

5.1.1 Datasets. We use two domain-specific datasets and a multidisciplinary dataset, and each dataset consists of 6 categories as labels. For ACL-ARC and ACT2, we use the original test split, and reserve 15% of the training data as the validation split since the release does not include a validation split. For FOCAL, we use the original split. The statistics are summarized in Table 1.

- **ACL-ARC** [32] is in the domain of computational linguistics, which is initially annotated and released by Jurgens et al. [28] and processed for citation classification by Cohan et al. [11]. Although the original version is used by several later works, it is pointed out that there are duplicates, data leakage, and incomplete sentences, which may be caused by limited OCR techniques in the early years. We use the cleaned version by Kunnath et al. [32].
- **FOCAL** [18] is from the astrophysical domain. The original labels seem to further divide the 'Compare/Contrast' into 3 fine-grained classes by sentiment (similarities, differences, neutrality). We regard them as one class to align with other datasets.
- **ACT2** [34, 51] is a highly heterogeneous multidisciplinary dataset that is challenging for existing citation classification methods, comprised of samples from over 20 domains including medicine, psychology, computer science, business, economics, etc.

5.1.2 Backbones. For encoder-based PLMs, we experiment with the SciBERT [3], a bert-based model pretrained on papers from the corpus of Semantic Scholar [2] and owns vocabulary that is built to best match the training corpus. SciBERT results in state-of-the-art performance on a wide range of scientific domain NLP tasks and is highlighted by the previous works in citation classification [33, 36, 41]. Following the experimental results of previous work [33], we use a null template [38] that does not include any task-specific patterns.

P1: {T}. [MASK].

For decoder-based LLMs, we experiment with the instruction-tuned Llama3-8B since it is one of the leading open-source LLMs among models of similar size. Specifically, we use the bfloat16 version of instruction-tuned Llama3-8B [3] and further apply the Low-Rank Adaptation [63] on $\mathcal{M}$ to reduce trainable parameters. As for the prompt, we write a task description suggested by previous work [36] to elevate the quality of task-specific output from the LLM. The overall prompt is as follows.

P2: *You are provided a context from a paper P citing a paper Q, with the specific citation marked as the "#CITATION_TAG" tag. Please analyze the citation function of the context which represents the author's motive or purpose for citing Q. Here is the context:"*{T}*". Only output one word as the answer:*

Our motivation for experiments with the LLMs is to offer a potential way to take advantage of the LLMs for citation classification, as well as to shed light on the possible performance. We did not opt for the larger models because the task is inherently in data shortage, which may not be affordable to blindly upgrade the model scales. For instance, we found finetuning 21M to 42M parameters out of the 8B is already a sweet point with maximal performance.

5.1.3 Baselines. We first introduce the best-performed feature-based baseline that does not include finetuning of any PLMs. (1) **Scaffold** [11]: It concatenates the Glove and ElMo embeddings of the words in the context as features and employs a BiLSTM-Attention model to aggregate among them. It further designs two auxiliary tasks, predicting whether a sentence contains a citation and predicting the section name, to handle the data scarcity.

We then introduce 3 baselines that are dedicated to finetune encoder-based PLM. (2) **TRL** [55]: It is a multi-task learning framework that uses the labeled data from auxiliary datasets to aid the finetuning on the primary dataset. A task relation learning procedure automatically computes the task weights. In our experiments, we use ACL-ARC and FOCAL as the auxiliary datasets for each other; for ACT2, neither of the other datasets is helpful so we only use itself. (3) **IREL** [41]: It is the winning system in the 2021 SDP Citation Context Classification Shared task [31], which finetune SciBERT and a linear classifier end-to-end with a class-balanced classification loss [29]. (4) **PET** [33]: It explored several closed-form prompts to finetune SciBERT in the Pattern Explicit Tuning style [2]. It reports **P1** as one of the best prompts across different datasets, resulting in previously state-of-the-art performance.

[2] www.semanticscholar.org
[3] huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Table 1: Dataset statistics.

	Splits				Citation Types (%)					
Dataset	#All	#Train	#Validation	#Test	Background	Compare/Contrast	Uses	Motivation	Extend	Future
ACL-ARC	1,929	1,399	246	284	51.3	18.1	3.7	3.6	4.6	18.7
FOCAL	4,166	2,617	660	889	41.2	24.7	0.3	0.9	5.8	27.1
ACT2	3,000	2,550	450	1,000	54.8	12.2	5.8	1.9	9.6	15.8

For the decoder-based LLMs, we implement 2 straightforward baselines as there are no previous works dedicated to tuning them for citation classification. (5) **IFP** [33]: It is a training-free method that adopts the model for text classification via instruction-following prompting and searches for the textual label from the decoded response. The prompt is listed in the Appendix. (6) **LoRA** [63]: It is a PEFT technique introducing a small number of trainable rank decomposition matrices to adapt the pretrained LLMs. We use **P2** in this baseline.

5.1.4 Implementations. We implement our framework and all LLM baselines with the python transformer library. For other baselines, we use their authorized implementation. We use a Llama3-70B model for the STKs extraction and the detials are in the Appendix. In all experiments, the context range l=3, the weight decay coefficient ω=0.01, the learning rate is $2e^{-5}$, the synoynym replacement ratio γ=0.1. In the LoRA component, r=16 for ACL-ARC and FOCAL, $r = 8$ for ACT2, and α=16 for all settings. For ACL-ARC, the final hyperparamers are d=1024, d_z=256, τ_1=1, τ_2=1, $|\mathcal{B}|$=4 for both backbones; λ_1=0.2, 0.1, λ_2=0.1, 0.2, β=0.6, 0.4, for SciBERT and Llama3-8B. For FOCAL, the final hyperparamers are λ_1=0.2, λ_2=0.1, τ_1=5, τ_2=1 for both backbones; d=256, 1024, d_z=128, 256, $|\mathcal{B}|$=16, 4, β=0.6, 0.7, for SciBERT and Llama3-8B. For ACT2, the final hyperparamers are λ_1=0.1, τ_1=0.1; λ_2=0.2, τ_2=10 for both backbones; d=256, 128, d_z=128, 64, $|\mathcal{B}|$=16, 4, β=0.3, 0.4 for SciBERT and Llama3-8B. The reported performance under each setting is averaged over 3 runs. At each run, the model is trained for at most 10 epochs with early stopping based on the summation of Macro-f1 and Accuracy computed based on the validation set. All the experiments are conducted on a server equipped with Intel(R) Xeon(R) Gold 6240 CPU and two NVIDIA A800 (80GB Memory). More implementation details are summarized in the Appendix.

5.2 RQ1: Overall Comparison

5.2.1 Classification Performance. The overall comparison is reported in Table 2. Macro-f1 represents the average performance on each class since the citation types are unevenly distributed as in Table 1. Based on the results, we have the following four main observations. (1) Citss achieves the state-of-the-art performance, outperforming existing methods on 5 out of 6 metrics when using SciBERT and on all metrics when using the Llama3-8B backbone. This highlights the versatility and effectiveness of our approach with both encoder-based PLMs and decoder-based LLMs. On ACT2, it is hard to achieve high scores on both metrics due to the difficulties in predicting its minority classes. IREL prioritizes Macro-F1 by weighting minority classes more heavily, compared to which our framework strikes a better balance and achieves significantly higher accuracy than IREL without substantial loss in Macro-F1. (2) Turning to the experiments with SciBERT, PET achieves the second-best overall performance. The success of both our method and PET, which leverage prompting strategies, highlights the important role of task-specific patterns in adapting PLMs for citation classification. TRL attains the second-best accuracy on FOCAL and ACT2, albeit for different reasons. On FOCAL, its performance validates the effectiveness of supplementing labeled data from the auxiliary ACL-ARC dataset. However, it becomes a dummy classifier on ACT2 that overwhelmingly predicts the majority class, resulting in the worst Macro-F1 score. (3) In experiments utilizing decoder-based LLMs, the finetuning methods essentially outperform the IFP. This superiority extends even when comparing finetuning to IFP using the significantly larger Instruction-tuned 70B model. This reflects that there is still a large gap between the generation approach, which relies solely on task descriptions and examples, and the data-driven finetuning approaches of cutting-edge LLMs for the complex text classification task of citation classification. (4) With Citss , Llama3-8B outperforms SciBERT on ACL-ARC, achieving significantly better results. While the performance gap between Llama3-8B and SciBERT narrows on Focal and ACT2, this suggests that Focal and ACT2 may benefit more from scientific text-specialized pretraining of SciBERT. ACL-ARC appears to leverage Llama3-8B's larger scale pretraining on general domain knowledge.

5.2.2 Efficiency. We now detail the absolute time costs under our experiment environments. (1) Citss : With SciBERT, it takes 17 to 23 minutes for training and 3 to 12 seconds for inference on the entire test set. For Llama3-8B with LoRA, it takes around 89, 160, and 131 minutes for training and 12, 71, and 53 seconds for inference on ACL-ARC, FOCAL, and ACT datasets, respectively. The time for producing transformed samples with KP and SC is negligible in comparison. (2) Among the baselines, the feature-based Scaffold is the most efficient, requiring only a few minutes for training and seconds for inference. However, this efficiency comes at the cost of significant manual effort in data preprocessing, particularly in collecting and cleaning the section name feature. (3) All SciBERT-based baselines share similar inference times. As for the training, IREL and PET require similar training time, which is around 6 to 12 minutes. TRL involves augmenting the training data and determining auxiliary task weights, resulting in a comparable overall training time to Citss with SciBERT. (4) For LLM baselines, while IFP is training-free, its inference speed is significantly slower due to the sequential decoding, requiring roughly 5 and 15 seconds per sample for Llama3-8B and Llama3-70B, respectively. This limits its practical applicability. LoRA finetuning, on the other hand, takes 40 to 92 minutes for training and 11 to 52 seconds for inference.

Table 2: Overall performance comparison. The best score is bolded and the second-best score is underlined for each backbone. And * denotes significant improvements (measured by t-test, p<0.05) compared with other baselines with the same backbone.

Method	Backbone	ACL-ARC		FOCAL		ACT2	
		Macro-f1	Accuracy	Macro-f1	Accuracy	Macro-f1	Accuracy
Scaffold	NA	0.496 ± 0.021	0.649 ± 0.012	0.145 ± 0.041	0.447 ± 0.133	0.146±0.006	0.363±0.017
IREL	SciBERT	0.614 ± 0.037	0.721 ± 0.025	0.580 ± 0.086	0.742 ± 0.007	**0.262**±0.012	0.468±0.028
TRL	SciBERT	0.476 ± 0.024	0.610 ± 0.007	0.604 ± 0.009	0.756 ± 0.009	0.118±0.001	0.544±0.000
PET	SciBERT	0.616 ± 0.022	0.723 ± 0.019	0.641 ± 0.044	0.750 ± 0.008	0.258± 0.018	0.537±0.020
Citss	SciBERT	**0.665*** ± 0.018	**0.743** ± 0.006	**0.679** ± 0.023	**0.777*** ± 0.005	0.254±0.012	**0.563***±0.009
IFP	Llama3-8B	0.422	0.575	0.243	0.398	0.213	0.446
LoRA	Llama3-8B-bfloat16	0.670 ± 0.050	0.745 ± 0.028	0.670 ± 0.035	0.757±0.001	0.242 ± 0.034	0.529 ± 0.014
Citss + LoRA	Llama3-8B-bfloat16	**0.744** ± 0.010	**0.819*** ± 0.007	**0.682** ± 0.024	**0.768*** ± 0.001	**0.266** ± 0.006	**0.549** ± 0.017
IFP	Llama3-70B	0.569	0.701	0.430	0.623	0.242	0.545

Table 3: Ablation Study by disabling the contrastive learning loss term in Citss. Imp. (%) shows the average improvements of both metrics. For Llama3-8B, we use the same LoRA setting and bfloat16 version.

SciBERT	ACL-ARC			FOCAL			ACT2		
	Macro-f1	Accuracy	Imp.(%)	Macro-f1	Accuracy	Imp.(%)	Macro-f1	Accuracy	Imp.(%)
$\lambda_1, \lambda_2 = 0$	0.616 ± 0.012	0.714 ± 0.022	-	0.641 ± 0.029	0.742 ± 0.004	-	0.262 ± 0.016	0.539 ± 0.021	-
$\lambda_2 = 0$	0.660 ± 0.025	0.745 ± 0.032	5.7	0.666 ± 0.029	0.762 ± 0.006	3.3	0.246 ± 0.036	0.558 ± 0.008	0.3
$\lambda_1 = 0$	0.646 ± 0.032	0.730 ± 0.017	3.5	0.654 ± 0.004	0.764 ± 0.009	2.5	0.260 ± 0.003	0.549 ± 0.017	0.9
Llama3-8B	Macro-f1	Accuracy	Imp.(%)	Macro-f1	Accuracy	Imp.(%)	Macro-f1	Accuracy	Imp.(%)
$\lambda_1, \lambda_2 = 0$	0.729 ± 0.007	0.799 ± 0.003	-	0.673 ± 0.008	0.762 ± 0.003	-	0.227 ± 0.006	0.544 ± 0.017	-
$\lambda_2 = 0$	0.736 ± 0.021	0.805 ± 0.013	0.9	0.681 ± 0.009	0.769 ± 0.004	1.0	0.248 ± 0.016	0.542 ± 0.016	2.5
$\lambda_1 = 0$	0.739 ± 0.008	0.804 ± 0.011	1.0	0.675 ± 0.005	0.766 ± 0.004	0.4	0.236 ± 0.009	0.556 ± 0.019	2.8

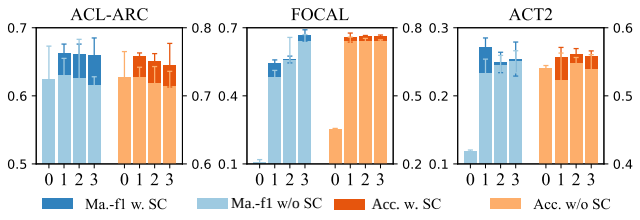

Figure 2: Ablation study of SC with different T_i. The x-axis is l, and $l = 0$ corresponds to the citance.

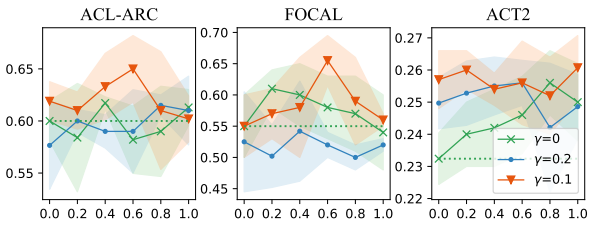

Figure 3: Performance with varying β (x-axis) and γ. The dashed line is the performance without KP.

5.3 RQ2: Investigation of Sentence-Level Cropping

We first conduct the ablation study by disabling the contrastive learning loss term in Citss, and list the result in Table 3. With all datasets and backbone models, using a single strategy is better than no strategy. Combining with the overall performance in Table 2, we can conclude that simultaneously using SC and KP further outperforms a single strategy.

Further, we adjust l in $T_i = \langle s_i^{-l}, ..., s_i^{l} \rangle$ to input context with different ranges to the SciBERT, evaluating the performance both with and without SC in Figure 2. On ACL-ARC, $l = 0$ is a decent input range with the highest accuracy, just as the setting of previous studies [33, 41, 55]. Increasing l without SC leads to performance degradation, suggesting that the surrounding sentences introduce excessive noise that overrides beneficial information. For FOCAL and ACT2, expanding the context window generally improves performance. This may indicate more citation-relevant long-range dependencies and less noise compared to ACL-ARC, making a broader context more advantageous. Across all datasets, SC consistently improves both metrics for $l = 1, 2, 3$, and outperforms $l = 0$ significantly. This validates the efficacy of SC in mitigating in-context noise and leveraging broader contexts effectively.

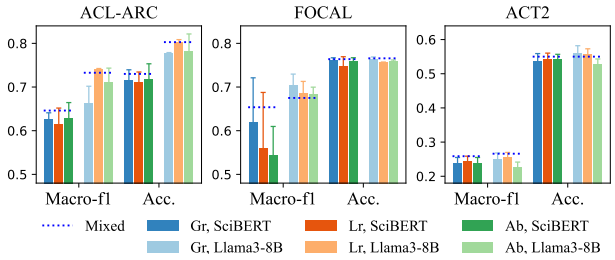

Figure 4: Performance with perturbation operation Op.

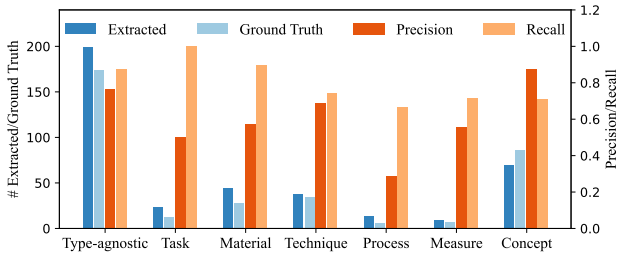

Figure 5: Extracted STKs estimated by humans.

5.4 RQ3: Investigation of Keyphrase Perturbation

We analyze the impact of the keyphrase perturbation ratio β and synonym replacement ratio γ on performance, reporting Macro-F1 scores with SciBERT in Figure 3 (with a similar pattern for Accuracy). Increasing γ initially improves performance within a certain range, but exceeding an optimal value leads to degradation. This suggests that introducing general domain diversity to scientific writing is beneficial to a point. On ACL-ARC and FOCAL, performance peaks as β increases under an optimal γ, indicating that gradually increasing the differentiation of keyphrases between positive samples enhances performance. But entirely different keywords between positive pairs may pose challenges for model optimization. On ACT2, varying β has a less significant effect with the optimal γ, possibly due to its high heterogeneity, where keyphrases are markedly dissimilar.

We also compare different proposed perturbation operations in Figure 4 by evaluating performance under single and mixed modes. In most cases, using the mixed operations achieves optimal or near-optimal results, facilitating the implementation of the framework.

Finally, we analyzed the extracted STKs, beginning with a quantitative assessment. Across the ACL-ARC, FOCAL, and ACT2 datasets, we obtained 6801, 21244, and 13731 STKs, respectively. Subsequently, we conducted a qualitative analysis by randomly sampling 20 contexts from ACL-ARC. Human annotators identified 174 STKs in these samples, while the LLM extracted 199. The quality of these extractions is further detailed in Figure 5. The main differences between the LLM-mined and manually annotated STKs reside in two aspects, neither of which poses a dangerous risk to our framework. First, the LLM exhibits a more lenient criterion for STK identification, resulting in a recall ranging from 0.667 to 1, which usually exceeds its precision. For example, the LLM identified "generating an initial description" as a Process, while human annotators did not acknowledge it as an STK. Introducing perturbations based on such high-order semantic elements, which are not strictly STKs, is unlikely to severely compromise the validity of the KP algorithm. Indeed, exploring alternative transformation strategies based on perturbing these higher-order elements, or even paraphrasing the context entirely, presents a promising avenue for future research. Second, the LLM demonstrated a tendency to categorize a larger proportion of STKs under the "Concept" category, which encompasses STKs not readily classified into other defined types. In contrast, human annotators were more adept at identifying specific STK types. Fortunately, the KP algorithm can perturb the STK in the original context regardless of its assigned type. The type-agnostic precision and recall, at 76 and 87%, are high, lending strong support to our KP strategy.

6 Conclusion

We introduce a framework, Citss , to finetune PLMs for citation classification via self-supervised contrastive learning. Our framework employs sentence-level cropping and keyphrase perturbation strategies to construct contrastive pairs without the need for labels. Taking the citation context as input, our framework acquires task-specific representation from the output of PLMs, and performs contrastive learning together with the supervised loss over the labeled data. Experiments with three benchmark datasets demonstrated that our framework improves the classification performance by an average 3.9% with SciBERT and 6.3% with Llama3-8B, compared to the previous state of the art.

References

[1] Riaz Ahmad and Muhammad Tanvir Afzal. 2018. CAD: An algorithm for citation-anchors detection in research papers. *Scientometrics* 117 (2018), 1405–1423.
[2] Riaz Ahmad, Muhammad Tanvir Afzal, and Muhammad Abdul Qadir. 2017. Pattern analysis of citation-anchors in citing documents for accurate identification of in-text citations. *IEEE Access* 5 (2017), 5819–5828.
[3] 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)*. 3615–3620.
[4] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).
[5] 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*. 779–783.
[6] 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 (2020), 1877–1901.
[7] Martin Juan José Bucher and Marco Martini. 2024. Fine-Tuned'Small'LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification. *arXiv preprint arXiv:2406.08660* (2024).
[8] Jingqiang Chen and Hai Zhuge. 2019. Automatic generation of related work through summarizing citations. *Concurrency and Computation: Practice and Experience* 31, 3 (2019), e4261.
[9] Liang Chen, Zekun Wang, Shuhuai Ren, Lei Li, Haozhe Zhao, Yunshui Li, Zefan Cai, Hongcheng Guo, Lei Zhang, Yizhe Xiong, et al. 2024. Next Token Prediction Towards Multimodal Intelligence: A Comprehensive Survey. *arXiv preprint arXiv:2412.18619* (2024).
[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
[11] 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). 3586–3596.

[12] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2270–2282.

[13] Arman Cohan and Nazli Goharian. 2018. Scientific document summarization via citation contextualization and scientific discourse. *International Journal on Digital Libraries* 19 (2018), 287–303.

[14] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2018. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. arXiv:1705.02364 [cs.CL] https://arxiv.org/abs/1705.02364

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[16] Terrance DeVries. 2017. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv preprint arXiv:1708.04552* (2017).

[17] Jiayuan Ding, Tong Xiang, Zijing Ou, Wangyang Zuo, Ruihui Zhao, Chenhua Lin, Yefeng Zheng, and Bang Liu. 2022. Tell me how to survey: literature review made simple with automatic reading path generation. In *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 3426–3438.

[18] Felix Grezes, Thomas Allen, Tirthankar Ghosal, and Sergi Blanco-Cuaresma. 2023. Function of citation in astrophysics literature (focal): Findings of the shared task. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*. 143–147.

[19] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. arXiv:2403.14608 [cs.LG] https://arxiv.org/abs/2403.14608

[20] Qianyue Hao, Jingyang Fan, Fengli Xu, Jian Yuan, and Yong Li. 2024. HLM-Cite: Hybrid Language Model Workflow for Text-based Scientific Citation Prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[21] Saeed-Ul Hassan, Iqra Safder, Anam Akram, and Faisal Kamiran. 2018. A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis. *Scientometrics* 116 (2018), 973–996.

[22] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*. 421–430.

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

[24] 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. arXiv:1902.00751 [cs.LG] https://arxiv.org/abs/1902.00751

[25] Rahul Jha, Amjad-Abu Jbara, Vahed Qazvinian, and Dragomir R Radev. 2017. NLP-driven citation analysis for scientometrics. *Natural Language Engineering* 23, 1 (2017), 93–130.

[26] Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2023. Scaling sentence embeddings with large language models. *arXiv preprint arXiv:2307.16645* (2023).

[27] Bowen Jin, Wentao Zhang, Yu Zhang, Yu Meng, Xinyang Zhang, Qi Zhu, and Jiawei Han. 2023. PATTON: Language Model Pretraining on Text-Rich Networks. In *61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*. Association for Computational Linguistics (ACL), 7005–7020.

[28] 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 (2018), 391–406.

[29] Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political analysis* 9, 2 (2001), 137–163.

[30] Suchetha N Kunnath, Drahomira Herrmannova, David Pride, and Petr Knoth. 2021. A meta-analysis of semantic classification of citations. *Quantitative science studies* 2, 4 (2021), 1170–1215.

[31] Suchetha N Kunnath, David Pride, Drahomira Herrmannova, and Petr Knoth. 2021. Overview of the 2021 SDP 3C citation context classification shared task. Association for Computational Linguistics.

[32] Suchetha Nambanoor Kunnath, David Pride, and Petr Knoth. 2022. Dynamic Context Extraction for Citation Classification. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 539–549.

[33] Suchetha N Kunnath, David Pride, and Petr Knoth. 2023. Prompting strategies for citation classification. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 1127–1137.

[34] Suchetha Nambanoor Kunnath, Valentin Stauber, Ronin Wu, David Pride, Viktor Botev, and Petr Knoth. 2022. ACT2: A multi-disciplinary semi-structured dataset for importance and purpose classification of citations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 3398–3406.

[35] Avishek Lahiri, Pratyay Sarkar, Medha Sen, Debarshi Sanyal, and Imon Mukherjee. 2024. Few-TK: A Dataset for Few-shot Scientific Typed Keyphrase Recognition. In *Findings of the Association for Computational Linguistics: NAACL 2024*. 4011–4025.

[36] Yibin Lei, Di Wu, Tianyi Zhou, Tao Shen, Yu Cao, Chongyang Tao, and Andrew Yates. 2024. Meta-task prompting elicits embedding from large language models. *arXiv preprint arXiv:2402.18458* (2024).

[37] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.

[38] Robert Logan IV, Ivana Balažević, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models. In *Findings of the Association for Computational Linguistics: ACL 2022*. 2824–2835.

[39] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[40] Kelvin Luu, Xinyi Wu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A. Smith. 2021. Explaining Relationships Between Scientific Documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 2130–2144. doi:10.18653/v1/2021.acl-long.166

[41] Himanshu Maheshwari, Bhavyajeet Singh, and Vasudeva Varma. 2021. SciBERT sentence representation for citation context classification. In *Proceedings of the Second Workshop on Scholarly Document Processing*. 130–133.

[42] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[43] Josh M. Nicholson, Milo Mordaunt, Patrice Lopez, Ashish Uppala, Domenic Rosati, Neves P. Rodrigues, Peter Grabitz, and Sean C. Rife. 2021. scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies* 2, 3 (11 2021), 882–898. doi:10.1162/qss_a_00146 arXiv:https://direct.mit.edu/qss/article-pdf/2/3/882/1970740/qss_a_00146.pdf

[44] Jeppe Nicolaisen. 2007. Citation analysis. *Annual review of information science and technology* 41, 1 (2007), 609–641.

[45] Kai Nishikawa and Hitoshi Koshiba. 2024. Exploring the applicability of large language models to citation context analysis. *Scientometrics* 129, 11 (2024), 6751–6777.

[46] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[47] 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)*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). Association for Computational Linguistics, Doha, Qatar, 1532–1543. doi:10.3115/v1/D14-1162

[48] Julien Perier-Camby, Marc Bertin, Iana Atanassova, and Frédéric Armetta. 2019. A preliminary study to compare deep learning with rule-based approaches for citation classification. In *8th international workshop on bibliometric-enhanced information retrieval (bir) co-located with the 41st european conference on information retrieval (ecir 2019)*, Vol. 2345. 125–131.

[49] Georgios Petasis, Alessandro Cucchiarelli, Paola Velardi, Georgios Paliouras, Vangelis Karkaletsis, and Constantine D Spyropoulos. 2000. Automatic adaptation of proper noun dictionaries through cooperation of machine learning and probabilistic methods. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. 128–135.

[50] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. arXiv:1802.05365 [cs.CL] https://arxiv.org/abs/1802.05365

[51] David Pride and Petr Knoth. 2020. An authoritative approach to citation classification. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. 337–340.

[52] David Pride, Petr Knoth, and Jozef Harag. 2019. Act: An annotation platform for citation typing at scale. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 329–330.

[53] Xiang Ren, Jialu Liu, Xiao Yu, Urvashi Khandelwal, Quanquan Gu, Lidan Wang, and Jiawei Han. 2014. Cluscite: Effective citation recommendation by information network-based clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 821–830.

[54] Muhammad Roman, Abdul Shahid, Shafiullah Khan, Anis Koubaa, and Lisu Yu. 2021. Citation intent classification using word embedding. *Ieee Access* 9 (2021), 9982–9995.

[55] Zeren Shui, Petros Karypis, Daniel S Karls, Mingjian Wen, Saurav Manchanda, Ellad B Tadmor, and George Karypis. 2024. Fine-Tuning Language Models on Multiple Datasets for Citation Intention Classification. *arXiv preprint arXiv:2410.13332* (2024).

[56] Yizhou Sun, Hongzhi Yin, and Xiang Ren. 2017. Recommendation in context-rich environment: An information network analysis approach. In *Proceedings of the*

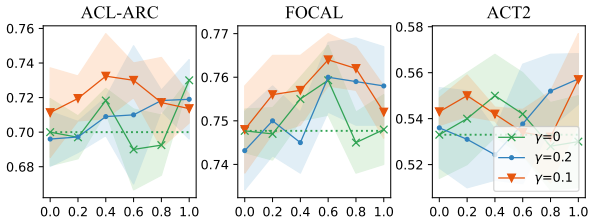

Figure 6: Accuracy with varying β (x-axis) and γ. The dashed line is the performance without KP.

26th International Conference on World Wide Web Companion. 941–945.
[57] Shahbaz Syed, Ahmad Hakimi, Khalid Al Khatib, and Martin Potthast. 2023. Citance-Contextualized Summarization of Scientific Papers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 8551–8568.
[58] Hong-Jin Tsai, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Citation intent classification and its supporting evidence extraction for citation graph construction. In *Proceedings of the 32nd ACM international conference on information and knowledge management*. 2472–2481.
[59] Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*.
[60] Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. 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)*. 6382–6388.
[61] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466* (2020).
[62] Wenhao Yu, Mengxia Yu, Tong Zhao, and Meng Jiang. 2020. Identifying referential intention with heterogeneous contexts. In *Proceedings of The Web Conference 2020*. 962–972.
[63] Yu Yu, Chao-Han Huck Yang, Jari Kolehmainen, Prashanth G Shivakumar, Yile Gu, Sungho Ryu Roger Ren, Qi Luo, Aditya Gourav, I-Fan Chen, Yi-Chieh Liu, et al. 2023. Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 1–8.
[64] Fanjin Zhang, Kun Cao, Yukuo Cen, Jifan Yu, Da Yin, and Jie Tang. 2024. PST-Bench: Tracing and Benchmarking the Source of Publications. *arXiv preprint arXiv:2402.16009* (2024).
[65] Yu Zhang, Hao Cheng, Zhihong Shen, Xiaodong Liu, Ye-Yi Wang, and Jianfeng Gao. 2023. Pre-training Multi-task Contrastive Learning Models for Scientific Literature Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 12259–12275.
[66] He Zhao, Zhunchen Luo, Chong Feng, and Yuming Ye. 2019. A context-based framework for resource citation classification in scientific literatures. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1041–1044.

A Hyperparameter Ranges

We tune $\lambda_1, \lambda_2 \in [0.01, 0.3]$, $\tau_1, \tau_2 \in [0.1, 20]$, the model dimension $d_z \in \{256, 128, 64\}$, $d \in \{1024, 512, 256, 128\}$, the batch size $|\mathcal{B}| \in \{4, 8, 16\}$, learning rate $lr \in \{ie^{-5}\}_{i\in\{1,2,4,10\}}$, dropout rate $dr = \{ie^{-2}\}_{i\in\{0,1,5\}}$. For the Llama3 experiments, we tune the lora rank $r \in \{8, 16, 32, 64\}$, the lora alpha $\alpha \in \{8, 16, 32\}$.

B Details of STKs

The statistics of the extracted STKs are summarized in Table 4.

C Supplementary Experiments

Figure 6 is the respective Accuracy of Figure 3.

D Instructions for LLMs

Prompt for STKs extraction

You are provided a context from a paper P, and please ignore the #CITATION_TAG. Your task is to identify scientific keyphrases from the context. Each scientific keyphrase belongs to one of the following classes:

- [Task]: The scientific problem or research focus addressed in the paper. It outlines the specific objectives or questions that the study aims to answer.
- [Material]: All materials utilized in the study, such as experimental tools, datasets, and the objects or subjects of investigation. It details the resources of the research.
- [Technique]: The specific methods, models, frameworks, or systems. It identifies the approaches taken to analyze data or solve problems.
- [Process]: It describes a sequence of steps or operations involved in a particular procedure, algorithm, or workflow. It emphasizes the procedural aspects.
- [Measure]: This class pertains to the metrics, indicators, or criteria used to assess or quantify the outcomes of the study.
- [Concept]: This category encompasses scientific keyphrases that do not fit into the aforementioned classes. It may include phenomena, theoretical terms, or entities relevant to the field of study.

Output your answer only in JSON format and be consistent with the text in the original context. Specifically, if there is any keyphrase of a certain class, use the class label as the key and the list of keyphrases as the value.

Here is an example: "The framework represents a generalization of several predecessor NLG systems based on Meaning-Text Theory: FoG (Kittredge and Polgu re, 1991), LFS (Iordanskaja et al., 1992), and JOYCE (Rambow and Korelsky, 1992). The framework was originally developed for the realization of deep-syntactic structures in NLG (#CITATION_TAG)"
Output:{'Technique': ['NLG systems', 'FoG', 'LFS', 'JOYCE', 'Meaning-Text Theory'], 'Concept':['deep-syntactic structures']}
Here is the context:{T}

Table 4: STKs statistics. "Type-agnostic" is the total quantity of STKs. Under each class, we report the average number of keyphrases of that class per sample.

Dataset	Type-agnostic	Task	Material	Technique	Process	Measure	Concept
ACL-ARC	6,801	1.14	1.87	1.72	0.76	0.48	3.07
FOCAL	21,244	0.91	3.73	1.44	1.42	2.48	4.42
ACT2	13,731	1.11	2.20	0.83	0.73	1.03	4.15

Prompt for the IFP baseline

You are provided a context from a paper P citing a paper Q, with the specific citation marked as the '#CITATION_TAG' tag. Please analyze the citation function of the context, which represents the author's motive or purpose for citing Q. The six classes of citation functions are:

- [BACKGROUND]: The cited paper Q provides relevant information or is part of the body of literature in this domain.
- [COMPARES_CONTRASTS]: The citing paper P expresses similarities or differences to, or disagrees with, the cited paper Q.
- [EXTENSION]: The citing paper P extends the data, methods, etc. of the cited paper Q.
- [FUTURE]: The cited paper Q is a potential avenue for future work.
- [MOTIVATION]: The citing paper P is directly motivated by the cited paper Q.
- [USES]: The citing paper P uses the methodology or tools created by the cited paper Q.

Here is the context: "{T}"

Only output the most appropriate class to categorize #CITATION_TAG and enclose the label within square brackets [].