IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 April 2025, accepted 25 June 2025, date of publication 8 July 2025, date of current version 15 July 2025.

*Digital Object Identifier 10.1109/ACCESS.2025.3586729*

RESEARCH ARTICLE

# Overcoming Data Scarcity: Guiding Citation Function Classification With Prompt-Based Few-Shot Learning

**KRITTIN CHATRINAN, THANAPON NORASET, AND SUPPAWONG TUAROB, (Member, IEEE)**
Faculty of Information and Communication Technology, Mahidol University, Salaya 73170, Thailand

Corresponding author: Suppawong Tuarob (suppawong.tua@mahidol.edu)

This work was supported in part by the Office of Higher Education Commission (OHEC), Thailand, and in part by Thailand Research Fund (TRF) under Grant MRG6080252.

**ABSTRACT** Citation function analysis is crucial for understanding how cited literature affects the narrative in scientific publications, as citations serve multiple purposes that need accurate distinction and classification. The field faces challenges due to insufficient labeled data and the complexity of defining and categorizing citation functions, requiring specialized knowledge and a deep understanding of scholarly literature. This limitation leads to the imprecise identification and classification of citation functions. To mitigate this challenge, we propose a meta-learning strategy that utilizes prompt learning with pre-trained language models, also known as prompt-based tuning, for the task of few-shot learning in citation function classification. Our findings demonstrate that prompt-based tuning with SciBERT surpasses state-of-the-art pre-trained language models with the conventional fine-tuning approach when labeled data is scarce. Furthermore, we present an analysis that sheds light on the impact of template selection on the prompt-based tuning methodology.

**INDEX TERMS** Citation function analysis, document classification, natural language processing, prompt-learning.

## I. INTRODUCTION

In scientific publishing, citations serve as a critical mechanism for elucidating an author's rationale for referencing specific works [1], thereby bolstering the arguments and contributions presented in a paper. The core elements of a citation include the surrounding text, known as the citation context, which draws from relevant scientific literature [2]. By analyzing the semantic content of cited publications, researchers gain valuable insights into their contributions to a scientific discourse [3]. Consequently, citation analysis has emerged as a vibrant research domain, attracting scholars seeking to understand the roles and significance of citations within academic papers.

Recent studies have explored various dimensions of citation analysis, with a particular emphasis on citation function analysis. This approach seeks to determine the purpose and significance of cited works within citing papers, examining the diverse roles citations play in scientific literature [4]. The complexity of citation functions arises from their varied applications across different contexts. For instance, one hypothesis posits that the structural organization of scientific papers—divided into sections such as Introduction, Methodology, Results, and Discussion—influences citation usage within these sections [5]. Additionally, some research investigates the practical applications of algorithms in scholarly work, exploring their utilization or adaptation in novel ways, such as identifying influential algorithms that drive subsequent innovations [6]. These multifaceted perspectives underscore the diverse ways citations convey meaning and contribute to the scientific narrative within academic papers.

Several datasets have been developed to study citation functions, each focusing on distinct aspects of citation roles. For example, the ACL-ARC dataset [5] examines citation

The associate editor coordinating the review of this manuscript and approving it for publication was Fu Lee Wang.

objectives across various sections of academic papers, while SciCite [7] broadens its scope to investigate citation purposes across diverse domains. Similarly, SciRes [8] analyzes citation roles based on their contextual usage within texts. These datasets provide valuable insights into citation functions but are primarily designed for general domains where labeled data is relatively abundant, facilitating efficient model training. However, this abundance is not replicated in specialized fields, such as medical research [9], legal studies [10], or algorithm development [6], where annotating citation functions demands significant expertise. The complex terminologies and nuanced author intentions in these domains require experienced researchers with deep disciplinary knowledge, making annotation time-intensive and resource-heavy. This reliance on expert annotations poses a significant barrier to scaling citation function analysis across disciplines, as the limited availability of labeled data hinders comprehensive studies that fully leverage citation contexts. Addressing this challenge requires the creation of additional labeled datasets tailored to capture the nuanced semantics of specific citation roles and objectives, while clearly delineating boundaries between them.

This study focuses on the analysis of algorithm citation functions, a task that presents unique challenges due to the intricate nature of algorithm-related contexts, their associated terminologies, and the diverse intentions behind authors' citations. These challenges are compounded by variations in authors' writing styles and the scarcity of labeled data, which necessitates expert knowledge to accurately interpret both the context of algorithm usage and the underlying purposes of citations. To address the issue of limited labeled data in algorithm citation function analysis, this study employs few-shot learning techniques, enabling models to achieve high predictive accuracy with minimal labeled examples. By curating representative samples of broader citation contexts within academic literature, this approach enhances the model's ability to generalize across diverse scenarios. Unlike traditional fine-tuning methods, which adapt pre-trained models to specific tasks, this study leverages prompt-based learning, reframing the citation function classification task as a text generation problem. This method assumes that carefully crafted prompts can effectively guide the model to produce accurate predictions for specialized tasks. By adopting this approach, we aim to mitigate the constraints imposed by insufficient annotated data and enhance the practical applicability of citation function prediction in algorithm-focused research.

This study makes several significant contributions to the field of citation function analysis.

1) We address the major challenge of limited labeled data by implementing a few-shot learning method, which enables models to achieve high accuracy even with minimal training data.
2) Through empirical evaluation, we demonstrate our proposed prompt learning approach outperforms traditional fine-tuning methods in scenarios where there is a scarcity of labeled data, likely due to its improved generalization capabilities and more effective use of limited data.
3) We analyze the impact of different template styles on model understanding and performance, providing valuable insights into how to effectively utilize this technique for citation function analysis.

These contributions should help advance the field of citation function analysis by improving the practical applicability of findings and broadening the implications that can be drawn from them.

In the following sections, we will delve into the background of our citation function analysis, detail the methodology employed, describe the experiments conducted, analyze the obtained results, and finally draw conclusions from our findings.

## II. BACKGROUND AND RELATED WORK

Analyzing citation functions has been a long-standing research domain in computer science and digital libraries [11], [12], [13]. While a citation is typically used to infer the reference relationship between a citing and the corresponding cited papers, researchers have attempted to understand the semantics behind citations in terms of functions, roles, and the author's attitudes towards the cited work [14]. This section first aims to lay out the foundation of previous work that involves citation function analysis for readers who are not familiar with this domain. Then, the subsequent subsections will provide relevant background on pre-trained language models and the prompt learning strategy.

### A. CITATION FUNCTION ANALYSIS

Citation function analysis represents a pivotal topic within the vast domain of citation analysis, which primarily involves the identification and examination of the purpose or role that citations play within academic articles. By scrutinizing citation functions, researchers can gain insights into how authors integrate existing research to support their own arguments and understand the interconnectedness of diverse disciplines through shared references.

Citation function analysis covers numerous aspects of citation roles as the target class and offers a variety of supervised learning methods to categorize the purposes of these roles based on the contexts found in scientific literature. Numerous studies differ in their definitions of citation roles based on the assumptions of authors. For example, Teufel et al. [3] developed a system for classifying citations into three categories, namely ''contrast,'' ''weaknesses of other work,'' and ''similarities between work.'' Furthermore, Cohan et al. [7] introduced a novel dataset that offers a more comprehensive representation of citation intents in scientific discourse across various domains. Recently, Tuarob et al. [6] concentrated on the diverse usages of algorithms in scientific literature through citation function analysis. These studies collectively contribute to the advancement of knowledge within this field

by explaining the complexities and differences of citation functions in academic writing.

While the citation function holds immense potential for academic research, its annotation process poses significant challenges due to several factors. Primarily, the task necessitates the involvement of domain experts who must interpret the author's intentions to establish the ground truth. Such a process is not only resource-intensive but also time-consuming, which, if not carefully supervised, could easily lead to false annotations and subsequently compromise the quality of the training data. To mitigate these limitations, an alternative approach is proposed: few-shot learning, a technique that enables models to learn from a small amount of high-quality data rather than relying on extensive resources and prolonged timeframes. By adopting this methodology, it is possible to streamline the citation annotation process, ultimately improving the overall accuracy and efficiency of the trained predictive models.

### B. PRE-TRAINED LANGUAGE MODELS IN SCIENTIFIC DOMAINS

Currently, the scientific domain holds immense importance due to its vast scope and critical role in advancing human knowledge across various disciplines such as physics, chemistry, biology, engineering, and social sciences. Given the complexity of scientific discourse, researchers have developed pre-trained language models specifically designed for the scientific domain, which serve as initial representation models that encompass all relevant scientific knowledge. These models are subsequently applied to address problems in diverse areas of research. In this study, two such pre-trained language models are selected for the citation function classification tasks, including:

- **SciBERT** [15], a pre-trained language model based on BERT [16] that is trained on multiple scientific corpora to perform different downstream scientific natural language processing tasks, including sequence tagging, sentence classification, and dependency parsing.
- **SPECTER** [17], a pre-trained language model that generates document-level embedding of documents that is pre-trained with the conjunction of scholarly documents and the citation graphs. SPECTER was reported to outperform competitive baselines on various document-level tasks in scientific domains.

### C. FEW-SHOT LEARNING

Few-shot learning, a subfield of meta-learning, trains models on various related tasks during a meta-training phase. This enables them to effectively generalize to new, unseen tasks using only a limited number of examples during meta-testing. This approach is particularly valuable in Natural Language Processing (NLP) for tasks such as text classification, sentiment analysis, and question answering, where obtaining large labeled datasets can be challenging and expensive [18].

Although few-shot learning has seen significant adoption in various NLP tasks, its application in citation analysis for downstream tasks remains relatively underexplored. Addressing this gap, a recent paper introduces ALPET, a novel method that combines active learning strategies with few-shot learning techniques, leveraging pre-trained language models to enhance Citation Worthiness Detection, especially in low-resource languages. Applied to Catalan, Basque, and Albanian Wikipedia datasets, ALPET demonstrates superior performance compared to the existing baseline, showcasing its potential in data-scarce scenarios [19]. This success suggests a promising avenue for further research into the application of few-shot learning across various aspects and tasks within citation analysis.

### D. PROMPT LEARNING

Prompt learning constitutes a significant topic within the domain of natural language processing. This method employs pre-trained language models, which have been trained on vast amounts of textual data, to address various downstream tasks such as text classification, machine translation, named entity recognition, and text summarization. Prompt learning operates under relaxed constraints by requiring no task-specific data for initial training. Unlike traditional supervised learning approaches that involve teaching a model to map input to output by learning a specific function, prompt learning is based on language models that directly model the probability of text generation [20]. This approach has the potential to simplify model training and improve performance, particularly in scenarios where task-specific data is limited or unavailable.

A substantial body of research has elucidated the synergistic advantages of combining prompt-based learning with few-shot learning methodologies, particularly in the context of natural language processing (NLP). For instance, Luo et al. [21] conducted a comprehensive study exploring the efficacy and computational efficiency of prompt-learning frameworks when applied to small-scale language models for domain-specific text classification tasks. Their results underscored the pronounced benefits of prompt learning, especially in scenarios with limited training data, such as few-shot and zero-shot learning environments. Specifically, the study demonstrated that fine-tuning models using few-shot prompt-based approaches consistently yielded superior performance compared to conventional fine-tuning techniques in data-scarce settings, highlighting the potential of prompt learning to enhance model adaptability and generalization.

Similarly, Lahiri et al. [22] introduced CitePrompt, an innovative prompt-based learning tool designed to classify citation intents within scientific literature. By leveraging carefully crafted prompts, CitePrompt achieved state-of-the-art or highly competitive performance on benchmark datasets, even in challenging few-shot and zero-shot scenarios. This work further emphasized the versatility of prompt-based learning in handling specialized tasks, such as intent classification, by capitalizing on the contextual understanding

of pre-trained language models. Collectively, these studies illustrate the transformative potential of integrating prompt learning with few-shot techniques, offering robust solutions for tasks requiring high performance with minimal labeled data.

These results emphasize the value of prompt learning as a powerful strategy for addressing text classification challenges, particularly in contexts where task-specific labeled data is scarce or absent. In contrast to conventional citation analysis tasks, our study focuses on enhancing the performance of citation function classification, a more granular and nuanced task. This task presents unique interpretive challenges, even for domain experts, due to its localized nature and the inherent complexity of discerning citation functions, especially when labeled data is limited.

## III. METHODOLOGY

In our research, we conducted an investigation into few-shot learning utilizing prompt learning, a technique known as prompt-based tuning. We compared this approach to traditional fine-tuning methods when dealing with limited labeled data for citation function classification tasks.

Mathematically, we can represent the text classification problem as $T = \{X, Y\}$, where $X$ denotes the collection of input instances and $Y$ represents the set of classes. Each instance $x \in X$ is comprised of multiple tokens, such as $w_1, w_2, \ldots, w_{|x|}$, and is assigned a label $y_x \in Y$.

### A. FINE-TUNING FOR PRE-TRAINED LANGUAGE MODELS

During the fine-tuning process of a pre-trained language model (PLM), the input instance $x = \{w_1, w_2, \ldots, w_{|x|}\}$ is first transformed into an input sequence of tokens $\{[CLS], w_1, w_2, \ldots, w_{|x|}, [SEP]\}$, where $[CLS]$ and $[SEP]$ are special tokens used to mark the start and end of the input sequence, respectively. Subsequently, the PLM transforms each token in the input sequence into a numerical representation. For a downstream classification task, a dedicated classification module is utilized to generate a probability distribution over the predefined class set $Y$ by applying the softmax function:

$$p(\cdot|x) = Softmax(W \cdot h_{[CLS]} + b) \tag{1}$$

where the hidden state of $[CLS]$ token is represented by $h_{[CLS]}$, $b$ is a trainable bias vector, and $W$ is a learnable matrix, initialized randomly prior to fine-tuning. This approach involves training the model on a task-specific dataset to adapt it to the target classification task while preserving the general knowledge learned by the PLM during pre-training.

### B. PROMPT-BASED TUNING FOR PRE-TRAINED LANGUAGE MODELS

Prompt-based tuning presents an alternative strategy for adapting Pre-trained Language Models (PLMs) to specific tasks. This method works by aligning the task format with the PLM's original training, often mimicking cloze-style objectives. Instead of retraining large parts of the model, we guide its predictions using carefully structured prompts. A prompt consists of a template, denoted as $T(\alpha)$ and a predefined set of relevant label words, $V$. The template takes an input instance $x$ and transforms it into a prompt input $x_{prompt} = T(x)$. This transformation involves arranging the original tokens from $x$ and potentially inserting supplementary tokens, critically including at least one $[MASK]$ token into $x_{prompt}$. The PLM is then tasked with predicting a suitable label word $v$ from the set $V$ to replace the masked position. Consider a binary classification example where the template is $T(\alpha) =$ "$\alpha$ *The purpose is* $[MASK]$" and map $x$ to $x_{prompt} =$ "$x$ *The purpose is* $[MASK]$". The hidden representation $h_{[MASK]}$ is computed for $[MASK]$ token. Given the vocabulary item $v \in V$, the probability that token $v$ can occupy the masked position is calculated using the softmax function:

$$p([MASK] = v|x_{prompt}) = \frac{exp(v \cdot h_{[MASK]})}{\sum_{\tilde{v} \in V} exp(\tilde{v} \cdot h_{[MASK]})} \tag{2}$$

A mapping function $\phi : Y \rightarrow V$, also known as a verbalizer, connects the set of classes to the set of label words. The probability distribution over the class set $Y$ is then formalized using the probability distribution over $V$ at the masked position:

$$p(y|x) = p([MASK] = \phi(y)|x_{prompt}) \tag{3}$$

This entire process allows the PLM to acquire task-specific knowledge while retaining the rich linguistic understanding gained during pre-training. The illustrations for both fine-tuning and prompt-based tuning approaches are shown in Figure 1. Table 1 shows the settings of the label words for each class set for this study.

**TABLE 1. Label words for each class set.**

| Scheme | Label Words |
|---|---|
| UTILIZATION | **UTILIZE**: utilize,<br>**NOT-UTILIZE**: not-utilize |
| USAGE | **USE**: use,<br>**EXTEND**: extend,<br>**MENTION**: mention,<br>**NOT-ALGO**: not-algorithm |

## IV. EXPERIMENTS AND RESULTS

This section reports the dataset used in this research and the key experiment results.

### A. DATA COLLECTION AND PRE-PROCESSING

To validate the proposed method, this research employed the algorithm citation function classification task [6] as the case study. Such a task aims to categorize a citation and its context into an algorithmic utilization function. The dataset under consideration consists of two classification schemes: UTILIZATION and USAGE. The former is a binary scheme that determines whether a citation context indicates utilization (UTILIZE) or non-utilization (NOT-UTILIZE) of the

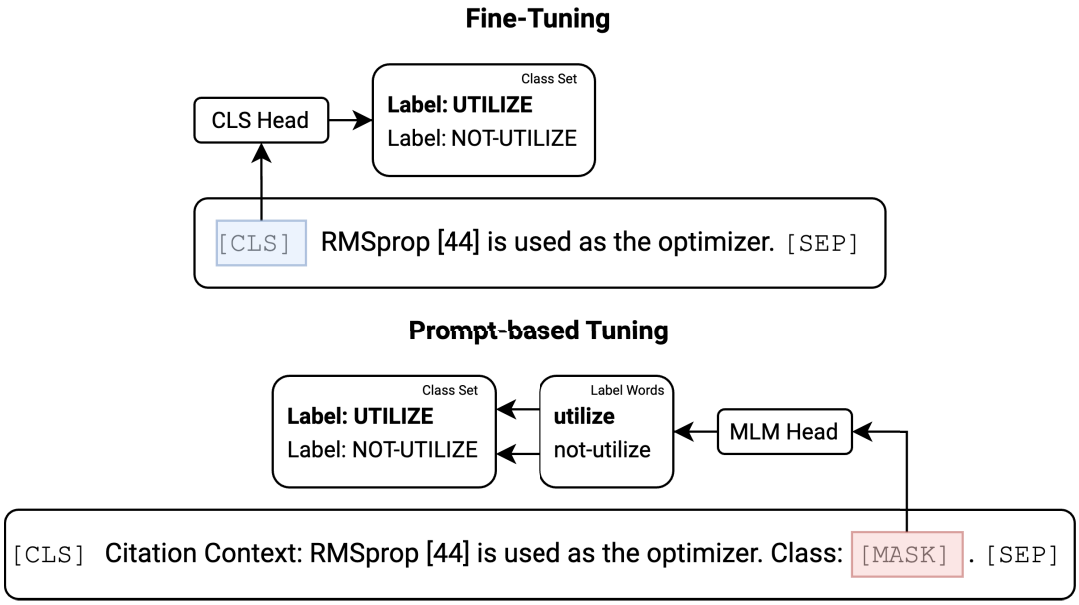


**FIGURE 1.** High-level methodology of the proposed fine-tuning and prompt-based tuning approaches.

**TABLE 2.** List of algorithm citation function classes and examples taken from [23], [24], [25], and [26], respectively.

| UTILIZATION Scheme | USAGE Scheme | Example |
|---|---|---|
| UTILIZE | USE | We used the enhanced version of the DeBERTa model named DeBERTaV3 **(He et al., 2021)**. The DeBERTaV3 model used the ELECTRA style pre-training by replacing mask language modeling (MLM) with the replaced token detection (RTD) strategy where the model is trained as a discriminator to determine whether an input token is either original or replaced by a generator. |
| | EXTEND | For computational efficiency, whenever the Wiki + Books corpora were used for pre-training, we initialized BioBERT with the pre-trained BERT model provided by **Devlin et al. (2019)**. We define BioBERT as a language representation model whose pre-training corpora includes biomedical corpora |
| NOT-UTILIZE | MENTION | Training ELECTRA-Large further results in an even stronger model that outperforms ALBERT **(Lan et al., 2019)** on GLUE and sets a new state-of-the-art for SQuAD 2.0. Taken together, our results indicate that the discriminative task of distinguishing real data from challenging negative samples is more compute-efficient and parameter-efficient than existing generative approaches for language representation learning. |
| | NOT-ALGO | Metastatic breast cancer occurs when breast tumor cells spread to other organs, such as the liver, brain, bones, or lungs, through the bloodstream or lymphatic system **[6]**. Breast tissue is mostly made up of glandular (milk-producing) and fat tissues, as well as lobes and ducts. |

cited paper's content. The latter is a more refined scheme, narrowing down the function of the cited algorithm into four categories: simple usage (USE), extension (EXTEND), mention (MENTION), and citations that are not related to algorithms (NOT-ALGO). The classification schemes and their examples of citation contexts are presented in Table 2. Interested readers are encouraged to consult the original dataset paper [6] for details on the classification scheme and criteria. The labeled dataset comprises 8,796 citation contexts, offering valuable insights into the utilization and usage of citations across a wide range of scholarly works. The dataset details are shown in Table 3. While this study

**TABLE 3. Citation function class distribution in algorithm citation function dataset.**

| UTILIZATION Scheme | USAGE Scheme | # Samples | Proportion |
|---|---|---|---|
| UTILIZED | USE | 752 | 8.50% |
| | EXTEND | 378 | 4.30% |
| NOT-UTILIZED | MENTION | 5750 | 65.40% |
| | NOT-ALGO | 1916 | 21.80% |

concentrates on algorithm citation functions, the presented approach has the potential for generalization and application to other aspects of citation function analysis that rely on classifying citations into their functional categories.

During the pre-processing stage, we standardized the citation signs, which appear in various formats across different scholarly works. By transforming these diverse citation formats into a uniform standard format, we facilitated better model understanding and prediction accuracy. This normalization also helps the model to focus explicitly on the specific citations under consideration, thereby minimizing potential confusion arising from the wide variety of citation styles commonly encountered in academic literature.

### B. EXPERIMENT SETTINGS

In order to emulate real-life situations where gathering a significant volume of labeled data can be difficult, we utilized small labeled datasets that were randomly selected from the overall labeled dataset. These datasets consisted of 10%, 20%, and 30% of the total labeled data, which is reflective of typical data availability in actual scenarios. Each model went through the training process utilizing these reduced datasets and the class. To guarantee reliable results, we carried out a 3-fold cross-validation and averaged the evaluation metrics which consist of Precision, Recall, and F1-score across all folds.

We compared the performance of three different models in our case study. The first model is FT-SciBERT, which serves as our baseline model. This model is a fine-tuned version of the SciBERT pre-trained language model, specifically developed for scientific texts. For FT-SciBERT, we employed four epochs of training with a batch size of 32 and a learning rate of 1e-5.

The second model is SPECTER, a recently introduced pre-trained language model tailored for scientific knowledge, which served as another state-of-the-art baseline. Similar to FT-SciBERT, we trained the SPECTER model using four epochs, a batch size of 32, and a learning rate of 1e-5 to ensure consistency in our experiments.

The third model is our proposed prompt-based tuning approach applied to SciBERT (PT-SciBERT). This method involves training the model for ten epochs, with a batch size of 32 and a learning rate of 5e-5. The use of prompts allows the model to better understand scientific contexts and improve its performance in our task. By comparing the results obtained from these three models, we aim to determine the most effective approach for addressing the citation function classification using pre-trained language models.

### C. RESULTS AND ANALYSIS

Tables 4 and 5 present the classification performance of various candidate models across multiple settings with limited labeled data, addressing our research question on the efficacy of prompt-based learning in enhancing model performance under data-scarce conditions. Notably, PT-SciBERT consistently outperforms other models, achieving the highest F1-scores across all evaluated settings. The most substantial improvement is observed when comparing PT-SciBERT to the baseline FT-SciBERT, with statistically significant gains in F1-score ($p < 0.05$ to $p < 0.001$) when utilizing 20% or 30% of labeled data. These results underscore the superiority of prompt-based tuning, particularly as the availability of labeled data increases beyond minimal levels, aligning with our research objective to investigate how prompt-based approaches can optimize citation function classification performance in low-resource scenarios.

To further explore the role of prompt design, a key component of our theoretical framework rooted in prompt-based learning theories, we evaluated three distinct template designs using 30% of the labeled data. These templates included a basic control template and two enhanced variations incorporating additional instructions and demonstrations (Table 6). Performance analysis (Table 7) revealed that the template integrating both instructions and demonstrations achieved the highest F1-scores across all classification schemes. This template exhibited a statistically significant improvement over the baseline in the USAGE task ($p=0.014$), highlighting its particular efficacy for nuanced classification tasks. These findings validate our theoretical assumption, drawn from prompt-based learning literature, that structured prompts with explicit guidance and contextual examples enhance a model's ability to interpret complex tasks.

These results directly address our research questions by demonstrating that well-crafted prompts, enriched with instructions and demonstrations, significantly improve classification accuracy in few-shot settings. They also contribute to the theoretical framework by illustrating how prompt-based learning leverages the contextual understanding of pre-trained models to reduce reliance on extensive fine-tuning, a core tenet of the paradigm. By providing clear instructions and representative examples, the enhanced templates align with cognitive-inspired theories of learning, where structured guidance facilitates task comprehension and generalization. This synergy between empirical performance and theoretical principles underscores the importance of tailored prompt designs in achieving optimal outcomes for specialized classification tasks, particularly in domains with limited labeled data. Furthermore, these findings suggest practical implications for developing robust models that can adapt to diverse scientific contexts, reinforcing the transformative potential of prompt-based learning in advancing citation function analysis.

**TABLE 4.** The classification results on the UTILIZATION scheme.

| Labeled Data (%) | Models | Precision | Recall | F1-score | P-value |
|---|---|---|---|---|---|
| 10 | FT-SciBERT | 0.77 | **0.744** | 0.75 | - |
| | SPECTER | 0.797 | 0.697 | 0.725 | 0.499 |
| | PT-SciBERT | **0.809** | 0.724 | **0.756** | 0.697 |
| 20 | FT-SciBERT | 0.796 | 0.743 | 0.765 | - |
| | SPECTER | **0.844** | 0.742 | 0.78 | 0.095 |
| | PT-SciBERT | 0.816 | **0.793** | **0.803** | 0.004 |
| 30 | FT-SciBERT | 0.821 | 0.847 | 0.802 | - |
| | SPECTER | 0.835 | 0.817 | 0.824 | 0.353 |
| | PT-SciBERT | **0.867** | **0.848** | **0.856** | 0.003 |

**TABLE 5.** The classification results on the USAGE scheme.

| Labeled Data (%) | Models | Precision | Recall | F1-score | P-value |
|---|---|---|---|---|---|
| 10 | FT-SciBERT | 0.549 | 0.51 | 0.491 | - |
| | SPECTER | 0.527 | 0.518 | 0.52 | 0.844 |
| | PT-SciBERT | **0.615** | **0.525** | **0.553** | 0.218 |
| 20 | FT-SciBERT | 0.625 | 0.613 | 0.601 | - |
| | SPECTER | 0.709 | 0.666 | 0.681 | 1.00E-04 |
| | PT-SciBERT | **0.723** | **0.677** | **0.694** | 3.00E-06 |
| 30 | FT-SciBERT | 0.676 | 0.685 | 0.675 | - |
| | SPECTER | 0.759 | **0.73** | 0.739 | 1.00E-07 |
| | PT-SciBERT | **0.767** | **0.73** | **0.744** | 1.00E-08 |

**TABLE 6.** Various template styles for algorithm citation function classification.

| ID | Template |
|---|---|
| 1 | Citation Context: X  Class: [MASK] |
| 2 | [INSTRUCTION]<br>Citation Context: X  Class: [MASK] |
| 3 | [INSTRUCTION]<br>Citation Context: [EXAMPLE]  Class: [LABEL]<br>Citation Context: X  Class: [MASK] |

**TABLE 7.** Classification results with varying settings of template styles in the UTILIZATION and USAGE schemes.

| Scheme | Template ID | Precision | Recall | F1-score | P-value |
|---|---|---|---|---|---|
| UTILIZATION | 1 | 0.86 | 0.822 | 0.825 | - |
| | 2 | 0.867 | **0.848** | 0.856 | 0.237 |
| | 3 | **0.874** | 0.845 | **0.859** | 0.188 |
| USAGE | 1 | 0.755 | 0.699 | 0.72 | - |
| | 2 | 0.762 | 0.725 | 0.736 | 0.232 |
| | 3 | **0.781** | **0.741** | **0.758** | 0.014 |

## V. DISCUSSION

This section examines the error analysis of misclassified samples and explores the limitations of the study.

### A. ERROR ANALYSIS

Further analysis was conducted to examine misclassified samples from the UTILIZE and USAGE schemes, aiming to uncover patterns that could guide future enhancements in model performance. For the UTILIZE scheme, a notable proportion of misclassifications occurred when the model incorrectly labeled a citation context as NOT-UTILIZE, despite the ground truth indicating UTILIZE. Detailed inspection of these misclassified instances revealed a key limitation tied to the density of citations within a given context. Specifically, contexts featuring multiple cited works often lacked clarity regarding the distinct contribution of each work. This issue was particularly pronounced in cases where authors cited several papers in close proximity to support a single argument or provide a broad overview, which obscured the specific role of individual citations. For example, consider the following modified citation context from [27]:

> "Identify similar edit operation sequences using the bi-gram matching. To find similar edit sequences, Repertoire uses the bi-gram matching algorithm [13]. We use a bi-gram matching instead of the longest common subsequence algorithm [14], because the bi-gram matching could allow slight variations in edit sequences."

In this instance, the presence of multiple citations within a concise context may confound the model, making it challenging to accurately determine how each cited work is utilized. Consequently, the model may erroneously classify the context as NOT-UTILIZE, misinterpreting the lack of explicit focus on individual citations as an absence of specific utilization.

Regarding the USAGE scheme, the predominant misclassification errors involved incorrect predictions between the MENTION and NOT-ALGO categories. These errors were frequently observed in contexts where authors employed abbreviated algorithm names, which appeared to impede the model's ability to correctly identify them as algorithms. For instance, consider the following example taken from [28]:

> ''There are also additional advantages of using the TTA such as, e.g., cycle minimization, implementation flexibility, performance scalability, etc. They are all explained in more details in [1]. Hence, these architectures have serious potential in important applications in the future.''

In this case, the model inaccurately predicted NOT-ALGO, likely failing to recognize the abbreviation ''TTA'' as an algorithm due to its brevity, despite contextual indicators suggesting its algorithmic significance and further explication in the referenced source. To address the identified limitations in model performance, particularly regarding citation density and abbreviated terminology, several targeted strategies can be implemented to enhance classification accuracy. One approach involves replacing the target citation with unique special tokens to increase the model's attention toward the specific citation being analyzed. This technique aims to emphasize the focal citation within dense citation contexts, thereby improving the model's ability to discern its distinct role. Additionally, a prompt-based learning strategy can be employed to incorporate key contextual information, such as the abstract section of the cited work, which often clarifies abbreviated terminology used in the study. By integrating such context into the prompt, the model gains a better understanding of core terminology, enabling a more accurate interpretation of its significance within the citation context. These combined approaches are designed to mitigate misclassification errors and enhance the model's robustness in handling complex academic texts.

### B. LIMITATIONS

Although the findings indicate excellent performance, this method is highly sensitive to both the provided instructions and the representative samples used. If the model misinterprets either the instructions or the examples, the performance can vary significantly, which underscores the need to optimize the instructions and carefully prepare representative samples. Additionally, the generalizability of these results is limited, as they apply only to the specific tasks and datasets utilized in this study. For other citation functions or more complex classification tasks, the effects of instructions and examples may differ. The inclusion of more complex representative samples and label words is also essential for performance improvement. Future research could investigate whether there is a saturation point beyond which adding more examples, excessively detailed instructions, or additional label words no longer improves results. Despite these limitations, this approach shows promise for optimizing classification outcomes within the context of this study.

## VI. CONCLUSION

In our research, we utilized pre-trained language models (PLMs) to address the challenge of citation function classification in scientific domains with limited labeled data. We explored two approaches: fine-tuning and prompt-based tuning, which showed promising results. While fine-tuning involves adapting the PLM's parameters for a specific task, prompt-based tuning relies on providing additional context or a template to guide the model's behavior. Our experiments demonstrated that prompt-based tuning outperformed fine-tuning in terms of F1-score, indicating its potential benefits for tasks with difficult label annotation, such as citation function analysis. Furthermore, we found that the choice of template style also impacted the classification performance. Specifically, a template with explicit instructions and demonstrations yielded better results than other template styles. In our future work, we intend to explore prompt-based tuning more extensively by integrating more complex representative samples and additional label words to better map to the class set. This approach aims to enhance generalizability across various domains of citation functions. By enabling greater flexibility during inference, we anticipate improvements in performance across multiple dimensions of citation function classification.

## REFERENCES


[1] R. Jan, ''Citation analysis of library trends,'' *Webology*, vol. 6, Mar. 2009. [Online]. Available: http://www.webology.org/2009/v6n1/a67.html

[2] L. Bornmann and H. Daniel, ''What do citation counts measure? A review of studies on citing behavior,'' *J. Document.*, vol. 64, no. 1, pp. 45–80, Jan. 2008.

[3] S. Teufel, A. Siddharthan, and D. Tidhar, ''An annotation scheme for citation function,'' in *Proc. 7th SIGdial Workshop Discourse Dialogue-SigDIAL*, Apr. 2006, p. 80.

[4] M. J. Moravcsik and P. Murugesan, ''Some results on the function and quality of citations,'' *Social Stud. Sci.*, vol. 5, no. 1, pp. 86–92, Feb. 1975.

[5] D. Jurgens, S. Kumar, R. Hoover, D. McFarland, and D. Jurafsky, ''Measuring the evolution of a scientific field through citation frames,'' *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 391–406, Dec. 2018.

[6] S. Tuarob, S. W. Kang, P. Wettayakorn, C. Pornprasit, T. Sachati, S.-U. Hassan, and P. Haddawy, ''Automatic classification of algorithm citation functions in scientific literature,'' *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 10, pp. 1881–1896, Oct. 2020.

[7] A. Cohan, W. Ammar, M. van Zuylen, and F. Cady, ''Structural scaffolds for citation intent classification in scientific publications,'' in *Proc. North Amer. Chapter Assoc. Comput. Linguistic*, Jul. 2019, pp. 3586–3596.

[8] H. Zhao, Z. Luo, C. Feng, A. Zheng, and X. Liu, ''A context-based framework for modeling the role and function of on-line resource citations in scientific literature,'' in *Proc. 9th Int. Joint Conf. Natural Lang. Process. Conf. Empirical Methods Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Mar. 2019, pp. 5205–5214. [Online]. Available: https://aclanthology.org/D19-1524/

[9] H. Kilicoglu, Z. Peng, S. Tafreshi, T. Tran, G. Rosemblat, and J. Schneider, ''Confirm or refute? A comparative study on citation sentiment classification in clinical research publications,'' *J. Biomed. Informat.*, vol. 91, Mar. 2019, Art. no. 103123. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046419300413

[10] D. Locke and G. Zuccon, ''Towards automatically classifying case law citation treatment using neural networks,'' in *Proc. 24th Australas. Document Comput. Symp.* New York, NY, USA: ACM, Dec. 2019, pp. 1–8, doi: 10.1145/3372124.3372128.

[11] M. Hernández-Alvarez and J. M. Gomez, "Survey about citation context analysis: Tasks, techniques, and resources," *Natural Lang. Eng.*, vol. 22, no. 3, pp. 327–349, May 2016.
[12] M. Hernández-Alvarez, J. M. G. Soriano, and P. Martínez-Barco, "Citation function, polarity and influence classification," *Natural Lang. Eng.*, vol. 23, no. 4, pp. 561–588, Jul. 2017.
[13] C.-S. Lin, "An analysis of citation functions in the humanities and social sciences research from the perspective of problematic citation analysis assumptions," *Scientometrics*, vol. 116, no. 2, pp. 797–813, Aug. 2018.
[14] I. Budi and Y. Yaniasih, "Understanding the meanings of citations using sentiment, role, and citation function classifications," *Scientometrics*, vol. 128, no. 1, pp. 735–759, Jan. 2023.
[15] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proc. 9th Int. Joint Conf. Natural Lang. Process. Conf. Empirical Methods Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., ACM, 2019, pp. 3615–3620. [Online]. Available: https://aclanthology.org/D19-1371
[16] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, J. Burstein, C. Doran, and T. Solorio, Eds., Jan. 2018, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423
[17] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. Weld, "SPECTER: Document-level representation learning using citation-informed transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Aug. 2020, pp. 2270–2282.
[18] M. Shah, D. Garg, A. Kothari, P. Hansora, A. Shah, and M. Parikh, "Advances and challenges in few-shot learning for natural language processing: A pilot study," in *Intelligent Strategies for ICT*, M. S. Kaiser, J. Xie, and V. S. Rathore, Eds., Singapore: Springer, 2024, pp. 31–40.
[19] A. Halitaj and A. Zubiaga, "ALPET: Active few-shot learning for citation worthiness detection in low-resource Wikipedia languages," *Expert Syst. Appl.*, vol. 281, Jul. 2025, Art. no. 127503. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095741742501125X
[20] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surveys*, vol. 55, no. 9, pp. 1–35, Jan. 2023, doi: 10.1145/3560815.
[21] H. Luo, P. Liu, and S. Esping, "Exploring small language models with prompt-learning paradigm for efficient domain-specific text classification," 2023, *arXiv:2309.14779*.
[22] A. Lahiri, D. K. Sanyal, and I. Mukherjee, "CitePrompt: Using prompts to identify citation intent in scientific papers," in *Proc. ACM/IEEE Joint Conf. Digit. Libraries (JCDL)*, Jun. 2023, pp. 51–55.
[23] A. Aziz, M. A. Hossain, and A. N. Chy, "Enhancing the DeBERTa transformers model for classifying sentences from biomedical abstracts," in *Proc. 20th Annu. Workshop Australas. Lang. Technol. Assoc.*, Dec. 2022, pp. 156–160. [Online]. Available: https://aclanthology.org/2022.alta-1.21
[24] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.
[25] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in *Proc. 8th Int. Conf. Learn. Represent.*, Addis Ababa, Ethiopia, Jan. 2020. [Online]. Available: https://www.openreview.net/pdf?id=r1xMH1BtvB
[26] A. Nomani, Y. Ansari, M. H. Nasirpour, A. Masoumian, E. S. Pour, and A. Valizadeh, "PSOWNNs-CNN: A computational radiology for breast cancer diagnosis improvement based on image processing using machine learning methods," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–17, May 2022.
[27] B. Ray and M. Kim, "A case study of cross-system porting in forked projects," in *Proc. ACM SIGSOFT 20th Int. Symp. Found. Softw. Eng.*, Nov. 2012, pp. 1–11.
[28] V. A. Zivkovic, R. J. W. T. Tangelder, and H. G. Kerkhoff, "Design and test space exploration of transport-triggered architectures," in *Proc. Design, Autom. Test Eur. Conf. Exhib.*, 2000, pp. 146–153.

**KRITTIN CHATRINAN** received the bachelor's degree in information and communication technology and the master's degree in computer science from the Faculty of Information and Communication Technology, Mahidol University, Thailand. His research focus involves natural language processing and the application of machine learning methods in the social media and healthcare fields.

**THANAPON NORASET** received the B.Sc. degree from the Faculty of Information and Communication Technology, Mahidol University, Thailand in 2007, and the Ph.D. degree in computer science from Northwestern University, USA, in 2017. He is a Faculty Member with the Faculty of Information and Communication Technology, Mahidol University. His research interests are in the field of natural language processing and machine learning.

**SUPPAWONG TUAROB** (Member, IEEE) received the B.S.E. and M.S.E. degrees in computer science and engineering from the University of Michigan-Ann Arbor, and the M.S. degree in industrial engineering and the Ph.D. degree in computer science and engineering from Pennsylvania State University. Currently, he is an Associate Professor of Computer Science and the Director of Machine Intelligence and Knowledge Engineering Research Clusters (MIKE) with the Faculty of Information and Communication Technology, Mahidol University in Thailand. His research primarily focuses on data mining in large-scale scholarly data, software engineering, social sciences, and healthcare. He is also interested in the applications of intelligent technologies for the betterment of society.