Prompting Strategies for Citation Classification

Suchetha N. Kunnath
snk56@open.ac.uk
Knowledge Media Institute
The Open University
Milton Keynes, UK

David Pride
david.pride@open.ac.uk
Knowledge Media Institute
The Open University
Milton Keynes, UK

Petr Knoth
petr.knoth@open.ac.uk
Knowledge Media Institute
The Open University
Milton Keynes, UK

ABSTRACT
Citation classification aims to identify the purpose of the cited article in the citing article. Previous citation classification methods rely largely on supervised approaches. The models are trained on datasets with citing sentences or citation contexts annotated for a citation’s purpose or function or intent. Recent advancements in Large Language Models (LLMs) have dramatically improved the ability of NLP systems to achieve state-of-the-art performances under zero or few-shot settings. This makes LLMs particularly suitable for tasks where sufficiently large labelled datasets are not yet available, which remains to be the case for citation classification. This paper systematically investigates the effectiveness of different prompting strategies for citation classification and compares them to promptless strategies as a baseline. Specifically, we evaluate the following four strategies, two of which we introduce for the first time, which involve updating Language Model (LM) parameters while training the model: (1) Promptless fine-tuning, (2) Fixed-prompt LM tuning, (3) Dynamic Context-prompt LM tuning (proposed), (4) Prompt + LM fine-tuning (proposed). Additionally, we test the zero-shot performance of LLMs, GPT3.5, a (5) Tuning-free prompting strategy that involves no parameter updating. Our results show that prompting methods based on LM parameter updating significantly improve citation classification performances on both domain-specific and multi-disciplinary citation classifications. Moreover, our Dynamic Context-prompting method achieves top scores both for the ACL-ARC and ACT2 citation classification datasets, surpassing the highest-performing system in the 3C shared task benchmark. Interestingly, we observe zero-shot GPT3.5 to perform well on ACT2 but poorly on the ACL-ARC dataset.

CCS CONCEPTS
• Computing methodologies → Natural language processing: Information extraction.

KEYWORDS
Citation Classification, Research Evaluation, Large Language Models, Prompt Training

ACM Reference Format:

1 INTRODUCTION
Citation classification, which aims to identify author’s intent for citing a specific article, is key for many scholarly document processing-based applications, including research evaluation [13, 35], information retrieval [39], document summary generation [12, 32] and so forth. Recently, the domain has benefited from the fine-tuning of large pre-trained scientific transformer-based models [2, 17, 30] and domain or task adaptive pre-training of generic models [8, 11, 27]. Although these semi-supervised models have significantly improved performance compared to the earlier fully-supervised feature-based [13, 38] and neural architecture-based models [6], the existing methods for citation classification still rely on supervised methods, utilising large annotated datasets for fine-tuning. The model performances largely depend on the size of the training corpus used, which for citation function classification is difficult to procure, given the challenges involved in determining the author’s intent from the citing sentence [15, 38].

Figure 1: Prompting strategies evaluated in this paper

The need to minimize supervision in NLP tasks and leverage the implicit knowledge gained by pre-trained Language Models (PLMs) from self-supervision using massive corpora has led to a shift from a pre-train fine-tune strategy to a pre-train, prompt, and predict one. This strategy involves reformulating the downstream task to match the LM training objective rather than adapting the model to the end-task [10, 23, 36]. LLMs, such as GPT-3-175B and PaLM-540B, have been very successful in solving a wide range of complex NLP tasks in few and zero-shot setups [4, 5], thus contributing significantly to this paradigm shift.
As the model size continues to grow exponentially, it has reached a point where a single LLM, trained in an unsupervised manner, is sufficient for solving any NLP problem. The latest introduction of conversational agents like ChatGPT has revolutionised research in NLP, demonstrating the capability for solving complex problems with little labelled data. However, the real challenge is finding the right prompt for such models to make reasonable predictions [23]. The existing prompt engineering methods range from hand-crafted [34, 36] to automatically generated prompts [10, 20]. Another design choice is whether the entire [3] or few [24] or none [10] of the LM parameters are updated, while used in conjunction with the prompts.

Despite advancements in prompt-based strategies for LMs, these methods have not been fully utilised for citation classification. Recently, Lahiri et al. [16] incorporated external knowledge related to different sections of the research paper as a mapping to the citation types and evaluated prompt-based learning for citation classification. However, the authors limit their analysis only to domain-specific datasets and a single prompting method [16]. A recent study by Nambanoor Kunnath et al. [30] reported that such domain-specific datasets are not sufficiently representative of the diverse citation behaviours of authors across research disciplines.

In this paper, we evaluate different prompting strategies based on whether LM parameters are updated or not for citation classification on two datasets with similar classification schema but with different domain distributions: (1) the multi-disciplinary ACT2 dataset [31, 35] and the (2) ACL-ARC dataset [13]. We evaluate the strategies illustrated in Figure 1, the typology of which is adapted from [23]. Our contributions are:

- We systematically investigate the effectiveness of prompt-based training strategies for citation classification on domain-specific and multi-disciplinary datasets.
- We propose a new method to choose prompts dynamically by using extended fixed and dynamic citation contexts.
- We also propose a new method for Prompt + LM fine-tuning by using citation-based prompts in the form of the relationship between citing and the cited articles generated by GPT3.5.
- Finally, we evaluate zero-shot GPT3.5 under different question-answer and instruction-based settings, including testing the model’s ability as an annotator for citation classification.

We release the source code and datasets used for all experiments here→https://github.com/oacore/prompt_citation_classification.

## 2 BACKGROUND

### 2.1 Parameter Updating

The pre-train fine-tune paradigm for citation classification requires large scientific task-agnostic LMs like SciBERT to be fine-tuned on a task-specific dataset [2]. The entire LM parameters are tuned in adapting to the task at hand. One limitation of model fine-tuning is the requirement for a large annotated corpus; obtaining that manually is difficult, especially for citation function classification, due to the challenges involved in understanding the author’s actual intent for citation from a single citing sentence [1, 12]. Moreover, the differences in training objectives of LM and task-specific fine-tuning prevent the knowledge representations learned by the pre-trained LMs from being fully utilised in the downstream tasks. Besides, pre-train followed by fine-tuning strategy is more efficient when there is no domain distribution discrepancy between the data used during both stages [19]. Hence, fine-tuning SciBERT, which is pre-trained on specific domains (Computer Science and BioMedicine), in a multi-disciplinary setting produces sub-optimal performances for citation function classification when compared to domain-specific datasets [30].

To adapt the model to the domain or the task corresponding to the dataset distribution, Gururangan et al. [11] proposes Domain Adaptive and Task Adaptive Pre-Training (DAPT and TAPT), which utilises a second stage pre-training. However, further pre-training is computationally intensive for domain adaptation since this requires training on a massive corpus. Although computational overhead for
TAPT is comparatively minimal as further pre-training is performed on the task dataset itself, the basic assumption here also depends on the availability of more humanly curated labelled training instances [8, 11].

Recent developments in LM tuning, as an alternative to fine-tuning, are based on reformulating the task by augmenting input with prompts. It converts the classification task to a cloze-style format by masking words, which the LM then predicts. This allows the model to take full advantage of the masked language modeling objective by performing LM tuning. This approach achieved state-of-the-art performance in both few-shot and zero-shot settings [4, 37]. One of the prominent prompt-engineering methods, Pattern Exploiting Training (PET), proposed by [36, 37] uses manually and automatically constructed prompts and verbalizers that map to the class labels for converting the classification task to cloze-style format. Yet another approach that combines the benefits of prompts and fine-tuning uses autoregressive models for generating prompts. Subsequently, the generated example-specific prompts, in combination with each instance, are then fine-tuned with the LM parameters [3].

### 2.2 No Parameter Updating

LLMs have achieved remarkable results for many NLP tasks without requiring further model parameter tuning. These include zero-shot prompt learning, the setting devoid of examples [4, 5]. Zero-shot learning has benefited from instructions to the model in the form of label description [9]. Another line of research inspired by decomposing complex problems into multiple more straightforward steps, known as Chain of Thought (CoT) prompting [40], uses simple instructions like “Let’s think step by step” for enhancing the reasoning ability of the model. The rationales from the model using CoT for a set of examples have been shown effective for in-context few-shot learning [14]. However, some previous studies have also noted that learning under few-shot settings is sensitive to the number or even the order of demonstrations used [22, 28]. Choosing the appropriate samples that are semantically similar to each test input query, rather than randomly extracted static demonstrations, has succeeded for text classification and reasoning tasks in the case of GPT3 [22, 42]. However, a recent study finds that adding more than one demonstration to test instances deteriorates results in the case of text classification with ChatGPT [41].

### 3 CITATION CLASSIFICATION: TASK AND DATASETS

Let P and Q be two research articles, where P is the citing paper and the latter is the cited article. Citation classification involves determining the motive behind referencing paper Q in P. It aims to capture the function or purpose of the citations to Q by utilizing the context from paper P, also known as citation context. Citation context explicitly contains citations to the cited article. It may be a single citing sentence or a set of multiple sentences, including the citing sentence.

We use the following two datasets for all experiments in this paper: (1) Domain-specific ACL-ARC dataset [13] and (2) Multi-disciplinary ACT2 [31, 35] dataset. ACL-ARC is a computational linguistics-based, domain expert annotated corpus derived from ACL Anthology3. However, ACT2 citation classification dataset is developed from multiple domains.

Both datasets have citing sentences or citation contexts annotated for 6 classes as illustrated in Table 1. The table also shows the descriptions of each class. The class distribution for ACL-ARC and ACT2 is highly skewed, with the majority of data points belonging to the background class [13, 31]. Due to the diversity in the domains present in ACT2, unlike the homogenous nature of ACL-ARC, identifying the author’s intent is challenging. The presence of symbols, mathematical notations and equations makes the classification task even more difficult [29]. Also, over half of the dataset has multiple citations in the citing sentence [31].

Figure 2 illustrate the domain distribution of the train and test set for ACT2. 148 research papers in training set belong to the top-level 22 domains, with Medicine, Physics, Psychology, Computer Science and Biology dominating the list. On the other hand, the test set contains 81 publications, with the majority of instances belonging to the domains: Medicine, Psychology, Environmental Sciences and Computer Science4.

---

3 In this paper, we use the terms citation classification and citation function classification interchangeably.

https://aclanthology.org/

4 We used the top-level domains from the Semantic Scholar API - https://api.semanticscholar.org/api-docs/graph. Using a semi-automated approach, we tag ACT2 dataset with their respective domains.
For ACL-ARC, we follow the train (1,647 instances) and test split (284 instances) used by [30]. ACT2, however, has 4,000 data points, of which, the first 1,000 instances form the test set, and the remaining 3,000 comprises the training set. We split the dataset in such a way as to maintain the original class distribution. 15% of the training set is further divided for both datasets to create a validation set.

4 EXPERIMENTS

4.1 Parameter Updating

4.1.1 Promptless Fine-Tuning. We experiment with head-based fine-tuning in this setting. A scientific LM is fine-tuned on both datasets using a classifier head. The pre-trained representation from the scientific LM using citation context is passed to a final linear head layer to obtain the output citation class. We test both the zero and few shot settings for model fine-tuning. For zero-shot classification, there is no updation of the LM parameters; hence no fine-tuning is involved. We further evaluate few-shot predictions using multiple training instance sets (10, 50, 100 etc.).

Besides LM fine-tuning, we additionally explore the effectiveness of TAPT + fine-tuning [11] on both datasets to evaluate how task-specific continued pre-training followed by fine-tuning improves scores. For this experiment, we use additional 3,000 citation contexts from the original version of the ACT2 dataset [35]. We use the same additional data for ACL-ARC pre-training as well.

4.1.2 Fixed-Prompt LM Tuning. We incorporate cloze-style manually constructed patterns into the citation context, as shown below;

“It is important not to confuse these findings with an impression shown by the title and summary of an article by #CITATION_TAG, that mind-wandering as such lowers mood. Function of #CITATION_TAG is <mask>.

The task-specific manually curated prompts augmented with citation context contain a single masked token. Along with a verbalizer, which maps the output label (the numeric value) to the predicted masked token, [Background: 0, Compare-Contrasts: 1, Extension: 2, Future: 3, Motivation: 4, Uses: 5], this prompting strategy thus involves the reformulation of citation classification task as a masked token prediction problem.

We test the following 3 manual prompts:

(1) prompt_manual_1: [citation_context][#CITATION_TAG is <mask>]
(2) prompt_manual_2: [citation_context][Function of #CITATION_TAG is <mask>]
(3) prompt_manual_3: [citation_context][Why does the citing paper cite #CITATION_TAG?<mask>]

The first two patterns are cloze-style phrases, while the third pattern-verbalizer forms a question and answer.

Additionally, we also evaluate null prompts [25], where a mask follows citation context and does not include any task-specific patterns:

(1) prompt_null_1: [citation_context]<mask>

The approach mentioned above aims to minimize the manual effort required for crafting prompts, which typically involves domain knowledge. The method seeks to eliminate any biases that may arise from creating task-specific templates, as claimed by the authors.

We follow the method based on PET, proposed by [36, 37] for the fixed-prompt LM tuning strategy. During the first step, PLM is fine-tuned using different citation function-specific prompts. Ensemble from these models is then used to annotate an additional unlabeled corpus. Finally, this auxiliary soft-labelled dataset is trained on a linear classifier for predicting the function labels. We also train LM across different few-shot settings to see how fixed-prompt LM tuning performs across various scales of data points, ranging from 10 to the entire training set, following [18].

4.1.3 Dynamic Context-Prompt LM Tuning. We introduce a novel method for prompt template engineering, employing dynamic patterns as an alternative to the fixed-prompt LM tuning approach. Specifically, we investigate whether incorporating additional contextual information around the citing sentences as prompts improve the prediction of the masked token, <mask>. To explore this, we employ dynamic patterns in the form of extended citation context as prompts. Following [30], we test the best-performing extended citation contexts using (a) fixed context, (b) Dynamic contiguous context and (c) Dynamic non-contiguous context for both datasets.

The fixed context involves extracting a predetermined number of sentences preceding or succeeding the citing sentence. The contiguous context pertains to a consecutive sequence of sentences within the citation context, while the non-contiguous context entails the selection of only relevant sentences to comprise the citation context. The corresponding prompt template in this setting would be of the form:

(1) prompt_dynamic_fc: [fixed_context]<mask>
(2) prompt_dynamic_cc: [dynamic_context]<mask>
(3) prompt_dynamic_ncc: [dynamic_non_contiguous_context]<mask>

In addition to citing sentences from both datasets, we used embedding similarity with citing and cited title, citing and cited abstract and the paragraph containing #CITATION_TAG as features for extracting the dynamic contiguous and non-contiguous contexts. For citing and cited feature representations, we use citation-informed SPECTER [7] and SciNCL [33] document embeddings. We use the CORE API1 for retrieving abstracts from citing papers for both ACL-ARC and ACT2. The abstracts from cited papers are obtained from multiple sources; CORE, Semantic Scholar8 and PubMed9.

4.1.4 Prompt + LM Fine-Tuning. Inspired by PADA [3], we propose a straightforward approach for evaluating the effectiveness of combining example-based prompts with LM fine-tuning. For the prompt engineering, our approach relies on [26], where the generated prompts are single sentences describing the relationship between the citing and the cited research paper for each sample in the dataset. The rationale is that the relationship between the citing

---

1https://github.com/oacore/dynamic_citation_context/tree/main/data/aclarc, after removing the duplicates in the dataset
2https://github.com/oacore/ACT2
3https://core.ac.uk/services/api
4https://www.semanticscholar.org/product/api
and the cited paper informs the author’s reason for citing. [26] developed SciGEN [10], a scientific GPT2-based model by pre-training GPT2 [11] using Computer Science papers to form SciGPT2. The resulting model is then fine-tuned on citation contexts (SciGEN), assuming that the association between citing and cited documents are best explained using citing sentences [26].

The generative prompt is combined with the input samples, followed by LM fine-tuning. Unlike PADA, which uses the same Text-to-Text Transfer Transformer (T5) model for prompt generation and text classification, we use separate models for citation classification. For prompt generation, we use GPT3.5, besides SciGPT2. Fine-tuning, however, is performed with a separate scientific LM.

We used the same features as dynamic prompts for generating text in this experiment too. We tested with various combinations of citing/cited title, citing/cited abstract, citation context and the entire paragraph containing #CITATION_TAG. Finally, we chose the following three prompts, which resulted in the best macro and micro f-scores for this experiment:

1. **prompt_generative_1**: [CLS][citation_context] + [citing_title] + [cited_title]
2. **prompt_generative_2**: [CLS][citation_context] + [citing_title] + [citing_abstract] + [cited_title]
3. **prompt_generative_3**: [CLS][citation_context] + [citing_title] + [cited_title]

where the <citation_context>, <citing_title>, <cited_title>, <citing_abstract> and <cited_abstract> represent the features used for generating prompts, and the [CLS] represents the fine-tuning classifier head token.

In this study, we did not evaluate parameter-efficient continuous/soft prompts because of their previously reported low performance on tasks with limited training data [21, 24]. Additionally, unlike hard prompts, these prompts lack interpretability.

4.2 No Parameter Updating

4.2.1 Zero-shot Tuning-free Prompting. We evaluate GPT3.5 with zero demonstrations to the model in this no LM parameter update setting. The following four prompt-based strategies are analysed for zero-shot classification: three Question-Answer Prompts (QAP) and one Instruction-based Prompt (IP). For QAP, the citation classification task is re-formulated to question and an answer for each class. To extract the answer from the given options, we append the trigger words - "Therefore, the answer is", following [14]. To generate the rationale for the model’s decision to choose specific classes and to decompose the task, we also examine the Chain-of-Thought prompt - "Let’s think step by step" [14, 40].

5 EXPERIMENTAL SETTINGS

5.1 Models and evaluation. We use the uncased sciavoc version of SciBERT [2] as the scientific LM for all the experiments that updates parameters [12]. Following [36], we fine-tuned SciBERT using a linear classifier head across different data points. For parameter update-free methods, we use the GPT-3.5-turbo model. We use OpenAI API call services for accessing the model [13]. All experiments are evaluated using macro and micro f-scores. The final scores for all experiments are the average of four runs using different seed values. We use promptless fine-tuning methods as the baseline for both datasets.

5.2 Prompt Selection. All PET experiments using varying prompt IDs and training instances are based on the experimental settings from [18] [14]. Both manual and null prompt evaluation is inspired by the prompt template engineering selection by [25, 37]. Dynamic context extraction for the dynamic prompt selection utilises the experimental settings from [30] [15]. Finally, for the Prompt + LM Fine-Tuning experiments, we used GPT-3.5-turbo and SciGPT2 [26] [16] for generating the sentence describing the relationship between the citing and the cited papers.

5.3 Additional Data. For TAPT, we use 3,000 supplementary unlabeled data (2,750 training instances and 250 validation set) derived from the full ACT dataset [35]. We re-use the source code from Hugging Face [17] for further pre-training SciBERT using task-specific data.

6 RESULTS

6.1 Parameter Updating

Fine-tuning SciBERT using citing sentences alone attains the highest macro f-score of 0.2554, given 1,500 training instances from [26]. We also evaluated methods using SPECTER [7]. However, the classifier performance was not as good as SciBERT.

https://github.com/Kel-Lu/SciGen
https://openai.com/blog/gpt-3-1-5b-release/
ACT2 and ACL-ARC macro f-scores on (a), (d) Fixed-Prompt LM Tuning; (b), (e) Dynamic Context-Prompt LM Tuning and (c), (f) Prompt + LM Fine-Tuning. The x-axis shows the number of training instances used.

Table 2: zero-shot parameter update-free method with GPT3.5

<table>
<thead>
<tr>
<th>Model</th>
<th>Prompt Type</th>
<th>Dataset</th>
<th>Prompt ID</th>
<th>macro f-score</th>
<th>micro f-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT3.5</td>
<td>zero-shot</td>
<td>ACL-ARC</td>
<td>prompt_QAP_1</td>
<td>0.4000</td>
<td>0.5000</td>
</tr>
<tr>
<td></td>
<td>Tuning-free</td>
<td></td>
<td>prompt_QAP_2</td>
<td>0.3808</td>
<td>0.5352</td>
</tr>
<tr>
<td></td>
<td>Prompting</td>
<td></td>
<td>prompt_QAP_3</td>
<td>0.3471</td>
<td>0.4354</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>prompt_IP_4</td>
<td>0.3611</td>
<td>0.3908</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACT2</td>
<td>prompt_QAP_1</td>
<td>0.2448</td>
<td>0.3720</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>prompt_QAP_2</td>
<td>0.2449</td>
<td>0.3720</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>prompt_QAP_3</td>
<td>0.2876</td>
<td>0.4510</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>prompt_IP_4</td>
<td>0.1828</td>
<td>0.3490</td>
</tr>
</tbody>
</table>

ACT2. The computational linguistics-based ACL-ARC dataset benefits from the domain distribution similarity with SciBERT pre-training corpus. This is evident from the zero-shot performance of SciBERT on the corpus, achieving as high as 0.4331 and 0.5537, respectively, for macro and micro f-score. Applying TAPT, followed by fine-tuning on ACT2 and ACL-ARC slightly improves performance by an average of 3.09% and 3.30%, respectively, which supports the findings of Gururangan et al. [11]. However, the performance improvement is the result of the inclusion of an extra multi-disciplinary dataset besides the train, test and validation sets.

The plots presented in Figure 3 illustrate the macro f-scores obtained by ACT2 and ACL-ARC for different parameter updating prompting methods across varying data points. Except for the prompt + LM fine-tuning, prompt-based methods deliver better performances when compared to the baseline promptless fine-tuned classifiers for both ACT2 and ACL-ARC. As the number of training instances increases, there is an overall improvement in the evaluation score in the case of ACL-ARC for all methods. However, with ACT2, the graphs fluctuates more often. This might be the result of the varying domain distribution in the selected training data points. Furthermore, considerable overlap between the different prompt methods evaluated under fixed-prompt and the dynamic-prompt LM tuning approaches for both datasets indicates that the choice of prompts did not make a substantial difference in the model performances.

Table 3 shows the highest macro and micro f-scores for the evaluated parameter updating method. One of our proposed methods, the Dynamic Context-prompt LM tuning method using fixed citation context as prompts, attained the highest macro f-scores for both datasets. For ACL-ARC, the method obtained the highest micro f-score of 0.7333 as well. This suggests that prompting benefits from additional information in the form of related context to citing
Table 3: Macro and micro f-scores obtained for different parameter and no parameter updating methods. Results indicate an average of 4 runs.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Prompt Type</th>
<th>Model</th>
<th>Prompt ID</th>
<th>macro f-score</th>
<th>micro f-score</th>
<th>#Data Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACL-ARC</td>
<td>Prompt-less</td>
<td>SciBERT</td>
<td>Zero-shot</td>
<td>0.4331</td>
<td>0.5537</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Fine-Tuning</td>
<td>SciBERT</td>
<td>Fine-Tuning</td>
<td>0.5547</td>
<td>0.7368</td>
<td>1,372</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TAPT + Fine-Tuning</td>
<td>prompt_null_1</td>
<td>0.6145</td>
<td>0.7183</td>
<td>1,372</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>prompt_manual_1</td>
<td>0.6266</td>
<td>0.7280</td>
<td>1,372</td>
</tr>
<tr>
<td></td>
<td>Fixed-Prompt</td>
<td>SciBERT</td>
<td>prompt_dynamic_fc</td>
<td>0.6415</td>
<td>0.7333</td>
<td>1,372</td>
</tr>
<tr>
<td></td>
<td>LM Tuning</td>
<td>SciBERT</td>
<td>prompt_generative_2</td>
<td>0.5240</td>
<td>0.7000</td>
<td>1,372</td>
</tr>
<tr>
<td></td>
<td>Prompt + LM</td>
<td>SciBERT + SPECTER</td>
<td>prompt_QAP_1</td>
<td>0.4000</td>
<td>0.5000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Fine-Tuning</td>
<td>SciBERT + GPT3.5</td>
<td>prompt_QAP_3</td>
<td>0.2876</td>
<td>0.4510</td>
<td>0</td>
</tr>
<tr>
<td>ACT2</td>
<td>Prompt-less</td>
<td>SciBERT</td>
<td>Zero-shot</td>
<td>0.1806</td>
<td>0.4718</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Fine-Tuning</td>
<td>SciBERT</td>
<td>Fine-Tuning</td>
<td>0.2554</td>
<td>0.5508</td>
<td>1,500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TAPT + Fine-Tuning</td>
<td>prompt_null_1</td>
<td>0.2700</td>
<td>0.5467</td>
<td>1,000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>prompt_manual_3</td>
<td>0.2702</td>
<td>0.5395</td>
<td>1,000</td>
</tr>
<tr>
<td></td>
<td>Fixed-Prompt</td>
<td>SciBERT</td>
<td>prompt_dynamic_fc</td>
<td>0.2843</td>
<td>0.5407</td>
<td>2,000</td>
</tr>
<tr>
<td></td>
<td>LM Tuning</td>
<td>SciBERT + SPECTER</td>
<td>prompt_generative_1</td>
<td>0.2505</td>
<td>0.5302</td>
<td>2,500</td>
</tr>
<tr>
<td></td>
<td>Prompt + LM</td>
<td>SciBERT + GPT3.5</td>
<td>prompt_QAP_3</td>
<td>0.2876</td>
<td>0.4510</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Fine-Tuning</td>
<td>SciBERT + SPECTER</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Zero-shot Tuning-free Prompting</td>
<td>GPT3.5</td>
<td>prompt_QAP_1</td>
<td>0.3770</td>
<td>0.4541</td>
<td>0</td>
</tr>
</tbody>
</table>

* Number of training instances used to attain the highest macro f-score.

sentences. Fixed-prompt LM tuning, including manual and null prompts, attained comparable results. However, as observed by [18], it could be possible that prompt-based methods might learn from the manual patterns as more and more instances are used for training.

Although the proposed prompt + LM fine-tuning method using citing title, citation context and the cited title for prompt generation using GPT3.5 attained comparable results for ACT2, the performances of these methods are far below the baseline models for both datasets [18].

6.2 No Parameter Updating

Table 2 shows the performance of GPT3.5 with zero-shot tuning-free prompting. ACT2 demonstrated strong performance with the question answer prompting method, particularly the prompt_QAP_3 variant, which outperformed the baseline model scores. The two other zero-shot QAP prompts achieved macro f-scores comparable to that of SciBERT when fine-tuned with the ACT2 training set.

Despite successfully outperforming baseline models for ACT2 dataset, QAP does not produce considerable performance improvement for ACL-ARC. Even the highest performing zero-shot tuning-free model only achieves a macro and micro f-score lower than the SciBERT’s zero-shot fine-tuning model.

Furthermore, we tested adding CoT prompts, as shown in Table 4, to the highest-performing zero-shot prompts for both datasets, but it did not improve results.

**Table 4:** Highest performing zero-shot parameter update-free method for ACL-ARC and ACT2 with CoT

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>Prompt ID</th>
<th>macro f-score</th>
<th>micro f-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT3.5</td>
<td>ACL-ARC</td>
<td>prompt_QAP_1 + CoT</td>
<td>0.3770</td>
<td>0.4541</td>
</tr>
<tr>
<td></td>
<td>ACT2</td>
<td>prompt_QAP_3 + CoT</td>
<td>0.2424</td>
<td>0.4270</td>
</tr>
</tbody>
</table>

Reformulating citation classification as an annotation experiment did not perform well for both datasets. This confirms the findings by [43] that conversational agents like ChatGPT is incapable of handling data annotation tasks at the moment. Moreover, the model needed more context occasionally to make a prediction. However, despite providing additional context, the model could not make considerable performance improvement. Perhaps adding more information in the form of annotation guidelines and class descriptions might help the model to predict the correct annotation.

7 DISCUSSION

Performance of promptless fine-tuning on ACT2: Some of the most highly performing methods for citation classification tend to be fine-tuned and evaluated on specific domains. However, as citation practices differ across disciplines, the popular SciBERT
Table 5: Predictions and rationales generated by GPT3.5 for prompt_QAP_3 on ACT2 citation contexts

<table>
<thead>
<tr>
<th>Citation Context</th>
<th>Rationale</th>
<th>Predicted</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>In contrast, bees use a graded process, whereby scouts initially discovering a new nest-site almost always recruit #CITATION_TAG, but the duration and rate of the recruiting waggle-dances are dependent on nest quality [42].</td>
<td>#CITATION_TAG is being cited to highlight a contrast between the graded recruitment process used by bees and the uniform recruitment process used by ants</td>
<td>COMPARES_CONTRASTS</td>
<td>COMPARES_CONTRASTS</td>
</tr>
<tr>
<td>Based on the results of recent searches #CITATION_TAG, we will take an SNR of 7 in each of the LIGO and Virgo detectors to be the approximate amplitude where a binary coalescence signal would stand above the noise background.</td>
<td>#CITATION_TAG is directly motivating the decision to take an SNR of 7 in each of the LIGO and Virgo detectors</td>
<td>MOTIVATION</td>
<td>MOTIVATION</td>
</tr>
<tr>
<td>Sugar production was agriculturally appropriate and, at the time, &quot;a good economic prospect&quot; (#CITATION_TAG &amp; Cook, 2000, p.1).</td>
<td>The author is using the cited source to support their statement about sugar production being a good economic prospect at the time</td>
<td>MOTIVATION</td>
<td>BACKGROUND</td>
</tr>
<tr>
<td>Its applications are diverse, including secret communication, copyright protection, digital watermarking, and tamper proofing #CITATION_TAG.</td>
<td>The paper is citing #CITATION_TAG because they are referring to its applications and how it is used in different areas</td>
<td>USES</td>
<td>BACKGROUND</td>
</tr>
</tbody>
</table>

Figure 4: Confusion Matrix for the highest performing systems - (1) ACT2 prompt_dynamic_fc, (2) ACT2 prompt_QAP_3, (3) ACL-ARC prompt_dynamic_fc and (4) ACL-ARC prompt_QAP_1

model is not particularly successful on a multi-disciplinary dataset, even when TAPT is used. It is likely that the domain distributional discrepancies between SciBERT’s pre-trained corpora and the ACT2 dataset are responsible for this.

Parameter Updating vs No Parameter updating methods: While the GPT3.5 zero-shot method achieved top performance on the ACT2 dataset, it performed the worst on the ACL-ARC dataset. It remains to be determined whether this might be related to the fact that SciBERT was trained on similar highly-specialised corpora as the one contained in the ACL-ARC dataset, while GPT3.5 has been trained on general-purpose corpora. However, it is remarkable to see that GPT3.5 achieves comparable performance to the dynamic prompt-based method for ACT2 in a zero-shot setup, which needed training on at least 2,000 instances to attain the highest score (Table 3).

Commonly misclassified citation functions: Figure 4 shows the confusion matrix of the highest-performing systems from both methods. For ACT2, the dynamic context prompt-based model is unable to identify the EXTENSION class. There is also a general tendency for it to incorrectly classify BACKGROUND as MOTIVATION and USES. A similar trend is observed even with GPT3.5, despite providing class descriptions as a hint to the model. Similarly, both models fail to classify instances to COMPARES_CONTRASTS, where citation context usually contains strong cue words associated with the author’s agreement or disagreement with the cited article. This is more evident in the case of ACL-ARC when used with GPT3.5.
Well as dealing with a higher frequency of co-citations and mathematical notations in citation contexts compared to corpora focused on computational linguistics [31].

Table 5 shows sample predictions from the highest performing zero-shot GPT3.5. Our results show that the model finds it hard to separate between BACKGROUND and MOTIVATION. Previously, Lauscher et al. [17] noted a high pointwise mutual information between these two classes while developing the MultiCite, multi-intent citation classification dataset, suggesting they might be naturally difficult to distinguish. Also, one of the common rationales generated by GPT3.5 in such cases is “the cited paper is used to support the claim made in the citing sentence”, which is not very helpful for identifying the correct class.

Overall the significant differences in scores obtained for ACT2 and the ACL-ARC datasets on all the methods indicate the challenges involved in citation classification under a multi-disciplinary setting. These challenges encompass a range of issues such as document parsing, acquiring metadata for both citing and cited articles, accommodating diverse citation styles across different fields, as well as dealing with a higher frequency of co-citations and mathematical notations in citation contexts compared to corpora focused on computational linguistics [31].

8 CONCLUSION

This paper provides the first systematic evaluation of prompting strategies for citation function classification. Unlike previous studies, we assess all models on both domain-specific and multi-disciplinary datasets and test them within both zero and few-shot settings. Our results show significant performance improvement of parameter updating methods across a variety of prompting strategies over promptless fine-tuning. The models using dynamic context-based prompts significantly improve model scores (figure 5) for both datasets and surpass the performance on the 3C shared task benchmark [29]. Although GPT3.5 achieved top performance on the ACT2 dataset in a zero-shot setup, i.e. comparable performance to our dynamic context-prompting method, which required 2,000 examples, it failed to reach even baseline performance on ACL-ARC.

ACKNOWLEDGMENTS

This research has been partially funded by Jisc under Grant Reference: 4133, OU Scientometrics PhD Studentship. Additional funding has been provided by CORE (core.ac.uk).

REFERENCES


