Meta In-Context Learning Makes Large Language Models Better Zero and Few-Shot Relation Extractors

Guozheng Li¹, Peng Wang¹,², Jiajun Liu¹, Yiakai Guo³, Ke Ji¹, Ziyu Shang¹ and Zijie Xu¹
¹School of Computer Science and Engineering, Southeast University
²Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education
³Beijing Institute of Computer Technology and Application
{gzli, pwang, jiajliu, keji, ziyus1999, zijiexu}@seu.edu.cn

Abstract

Relation extraction (RE) is an important task that aims to identify the relationships between entities in texts. While large language models (LLMs) have revealed remarkable in-context learning (ICL) capability for general zero and few-shot learning, recent studies indicate that current LLMs still struggle with zero and few-shot RE. Previous studies are mainly dedicated to design prompt formats and select good examples for improving ICL-based RE. Although both factors are vital for ICL, if one can fundamentally boost the ICL capability of LLMs in RE, the zero and few-shot RE performance via ICL would be significantly improved. To this end, we introduce MCRE (Meta In-Context learning of LLMs for Relation Extraction), a new meta-training framework for zero and few-shot RE where an LLM is tuned to do ICL on a diverse collection of RE datasets (i.e., learning to learn in context for RE). Through meta-training, the model becomes more effectively to learn a new RE task in context by conditioning on a few training examples with no parameter updates or task-specific templates at inference time, enabling better zero and few-shot task generalization. We experiment MCRE on various LLMs with different model scales and 12 public RE datasets, and then evaluate it on unseen RE benchmarks under zero and few-shot settings. MCRE delivers comparable or superior performance compared to a range of baselines including supervised fine-tuning and typical in-context learning methods. We find that the gains are particular significant for larger model scales, and using a diverse set of the meta-training RE datasets is key to improvements. Empirically, we show that MCRE can transfer the relation semantic knowledge via relation label name during inference on target RE datasets.

1 Introduction

Relation extraction (RE) [Wang and Lu, 2020; Wei et al., 2020; Li et al., 2022; Wang et al., 2023b] aims to identify the relationships between entities in texts, and plays an important role in natural language processing (NLP). Existing RE methods with pre-trained language models (PLMs) [Devlin et al., 2019; Liu et al., 2019] have achieved outstanding performance by fully supervised fine-tuning. However, such a supervised paradigm heavily depends on large-scale annotated data. Hence, in real-world scenarios, existing methods tend to struggle when recognizing new relations with insufficient annotation resources (low-shot) which is coined as zero and few-shot RE [Han et al., 2018; Chen and Li, 2021].

Two popular paradigm of methods are emerged to alleviate the challenge of zero and few-shot learning, including meta learning [Finn et al., 2017; Snell et al., 2017] and in-context learning [Brown et al., 2020; Wei et al., 2022b]. Meta-learning provides a framework for learning to learn, which addresses the challenges of zero and few-shot learning by training models on multiple tasks with limited labeled data, enabling them to generalize to new, unseen classes or tasks. However, existing meta-learning-based RE studies [Zhao et al., 2023; Zhang et al., 2023] still typically focus on specific tasks, datasets and settings, lacking of flexibility and generalization to more general low-shot scenarios. Thereafter, in-context learning (ICL) which concatenates a query and few-shot demonstrations to prompt LLMs for prediction is proposed, where it is plug-and-play and does not require additional inductive bias learning or sophisticated template design. And recent studies [Wei et al., 2022b; Kojima et al., 2022] on large language models (LLMs), such as GPT-3 [Brown et al., 2020] and ChatGPT [OpenAI, 2022], demonstrate that LLMs perform well in various downstream tasks without any training or fine-tuning but only with ICL. However, some recent researches [Ma et al., 2023b; Wang et al., 2023c] have revealed a significant performance gap in LLMs when it comes to apply ICL to the low-shot RE tasks, while others [Agrawal et al., 2022; Li et al., 2023] believe that LLMs deliver promising performances in low-shot RE, because of the influence of query forms and selected demonstrations. But we argue that fundamentally improving the ICL capability of LLMs in RE is substantially important.

In this work, we borrow and combine the ideas of meta learning and in-context learning in RE, introduce a new low-shot RE framework with meta in-context learning [Min et al., 2022; Chen et al., 2022] called MCRE: Meta In-Context
learning of LLMs for Relation Extraction. Specifically, we reformulate the RE task as a natural language generation problem. Unlike previous approaches that fine-tune the models with task-specific augmentation, MICRE tunes an LLM on a collection of RE datasets to learn how to in-context learning, and is evaluated on strictly new unseen RE datasets with zero and few-shot prompting [Brown et al., 2020; Kojima et al., 2022]. Each meta-training examples matches the inference setup where it includes several examples in training sets from one RE dataset that will be concatenated together as a single sequence to the LLM, and the output of the final example is used to calculate the cross-entropy training loss. Tuning the LLM in this manner directly leads to better ICL and low-shot RE, where the LLM learns to recover the semantics of the RE task from the given examples at inference time. This method is related to recent work [Wei et al., 2022a; Sanh et al., 2022] that uses multi-task learning for better zero-shot performance at inference time. However, MICRE allows adapting to new RE datasets and domains from several examples alone under few-shot scenarios, without relying on a task reformattting (e.g., reducing the RE process to input-output format) or task-specific templates (e.g., converting the RE task to a natural language generation problem). For zero-shot cases, we induce the complete prediction results from LLMs through prompting known entities and relations, keeping consistent prompt formats with meta-training phases.

To unify the prompt formats of zero and few-shot RE, we design a simple tabular prompting [Li et al., 2024] for extracting subjects, objects and predicates from texts. Specifically, a table header “| Predicate | Subject | Object |” is provided as part of the prompt and the LLMs automatically generate a table, where “|” is the recognizable delimiter of tables. This enables the effective zero-shot entity and relation extraction even without the participation of few-shot examples. We experiment MICRE with tabular prompting on a collection of publicly available RE datasets and evaluate it on unseen RE benchmarks to ensure no overlap between meta-training and target datasets. Experimental results show that MICRE consistently outperforms baselines including ICL without meta-training and task-specific zero and few-shot RE with fine-tuning. This demonstrates MICRE enables LLMs to recover the semantics of the entities and relations in context during inference. In summary, our contributions are three-fold:

- We introduce MICRE, a new meta in-context training framework based on LLMs for zero and few-shot RE to learn how to do in-context learning, resulting in better low-shot prompting performance at new unseen RE tasks. We also devise a effective tabular prompting to unify the prompt formats of zero and few-shot RE.
- By meta in-context training MICRE on various LLMs with different model scales and a collection of RE datasets, MICRE delivers comparable or superior performance compared to state-of-the-art zero and few-shot RE baselines on unseen datasets.
- Further analysis shows that the gains are particular significant for larger model scales and diverse datasets. And MICRE achieves good ICL results during inference due to its relation semantic knowledge transferring.

2 Related Work

Meta in-context learning. Large language models (LLMs) perform well in various downstream tasks without any training or fine-tuning but only with a few examples as instructions, which is called in-context learning [Brown et al., 2020]. ICL has been further improved by later work and shows promising results on a variety of tasks [Zhao et al., 2021; Min et al., 2022; Ji et al., 2023; Shang et al., 2024; Liu et al., 2024]. However, ICL with LLMs achieves poor performance when the target task is very different from language modeling in nature such as entity and relation extraction. While prior work has shown that multi-task learning on a large collection of tasks leads to better performance on a new task when tested zero-shot [Mishra et al., 2022; Wei et al., 2022a, recent studies [Min et al., 2022; Chen et al., 2022] propose to explicitly train models on an in-context learning objective via multi-task learning, where this learning to learn in-context paradigm is typically called meta in-context learning. Our work is based on the core idea of meta in-context training regarding entity and relation extraction via multi-task learning, showing MICRE achieves substantial improvements on zero and few-shot RE tasks.

Zero and few-shot relation extraction. Few-shot relation extraction [Han et al., 2018; Yu et al., 2020; Wang et al., 2023a] aims to predict novel relations by exploring a few labeled examples. MAML [Finn et al., 2017] and prototypical networks [Snell et al., 2017] are widely used and combined with pre-trained language models [Devlin et al., 2019; Liu et al., 2019] in few-shot settings to achieve impressive results. To be capable of extracting relations that were not specified in advance, zero-shot relation extraction [Levy et al., 2017; Chen and Li, 2021; Chia et al., 2022] is proposed to invent new models to predict new relations. Besides the fine-tuned small language models for RE, recent studies [Ma et al., 2023b; Li et al., 2023; Wan et al., 2023] leverage the LLMs with zero and few-shot prompting to extract entities and relations from texts through ICL. However, the current primary research of ICL for RE only consider two directions: query forms [Li et al., 2023; Li et al., 2024] and demonstration retrieval techniques [Ma et al., 2023b; Wan et al., 2023]. Although both factors are vital for achieving better ICL for RE performance, we argue that fundamentally improving the ICL capability of LLMs in RE task is substantially important. Therefore, we propose to train LLMs with in-context learning objective to improve the ICL ability of LLMs for better in-context RE. The main insight is to conduct a multi-task learning scheme over a collection of meta-training RE datasets, in order to learn how to condition on a small set of training examples, recover the semantics of a new RE task, and predict the output based on it.

3 Methodology

3.1 Task Formulations

We consider two RE tasks: relation classification (RC) and relational triple extraction (RTE), under two low-shot settings: zero-shot setting and few-shot setting. (1) Relation classification: Given a sentence $S_i$ and an entity pair $(s_i, o_i)$ where
Each iteration $i$, input and train the model to generate $y_{k+1}$ using a negative log likelihood objective in the final step.

$D_i$ is the subject and $o_i$ is the object. This task aims to identify a relation $r_i$ from pre-defined relation set $R$ that satisfy the relationship between subject $s_i$ and object $o_i$ expressed by the sentence $S_i$. (2) **Relational triple extraction**: Given a sentence $S_i$, this task aims to jointly extract the relational triple $z_i = (s_i, r_i, o_i)$ from sentence $S_i$ where $s_i$ and $o_i$ are entities and $r_i$ is the corresponding relation from relation set $R$. (3) **Low-shot settings**: For zero-shot setting, the model are expected to identify novel entities and relations at the inference time without any training examples. For few-shot setting, we formulate it in the typical N-way-K-shot form. Given the training set $D$ and the target set $T$, for $N$ way (relation types), the model utilizes randomly $K$ examples for each relation type from $D$ to form the support set $S = \{(x_i, y_i)\}_{i=1}^{K}$, where $x_i$ denotes the sentence $S_i$ and $y_i$ denotes the relation $r_i$ in RC or relational triple $z_i$ in RTE. The model are expected to utilize only $N \times K$ examples from $S$ for prediction and outputs recognized relations in RC and triples in RTE for each target sentence in $T$.

### 3.2 Meta In-Context Training

Following previous literature [Brown et al., 2020; Min et al., 2022; Chen et al., 2022], the training examples are concatenated and provided as a single input to the model, which is feasible for $k$-shot learning. At test time, the model is evaluated on an unseen target RE task that comes with $N \times K$ training examples (i.e. few-shot learning) or no training examples (i.e. zero-shot learning), and inference directly follows the similar prompting format as in meta-training phases.

Figure 1 provides an overview of the meta-training work flow of MCRE which consists of five steps. The model is meta-trained on a collection of RE datasets which we call meta-training datasets. Specifically, for each meta-training iteration $i$ in step 1, we sample a dataset $D_i$ from $C$ meta-training datasets in step 2. Then $k + 1$ training examples $(x_{1}, y_{1}), \ldots, (x_{k+1}, y_{k+1})$ are sampled from the training examples of the chosen dataset in step 3. To unify different tasks (i.e. RC and RTE) and settings (i.e. zero and few-shot) for convenient transferring, we transform these $k + 1$ samples into the tabular prompting format (introduce it later) in step 4. We then supervise the model (tune the LLMs) by feeding the concatenation of $x_{1}, y_{1}, \ldots, x_{k}, y_{k}, x_{k+1}$ to the model as an input and train the model to generate $y_{k+1}$ using a negative log likelihood objective in the final step.

### 3.3 Tabular Prompting

We adopt a tabular prompting for unifying two different tasks and settings that generates organized and concise outputs in ICL. Specifically, a table header “| Predicate | Subject | Object |” is provided to prompt the LLMs to automatically generate a table, where “|” is the recognizable delimiter of tables. This strategy is suitable for both zero and few-shot settings compared to solely text-to-text prompting format, as it provides precise instructed signals by table header for zero-shot prompting. During meta-training, we utilize two orders of table header “| Predicate | Subject | Object |” and “| Subject | Object | Predicate |” to improve the robustness of models and unify the RC and RTE tasks at inference time.

### 3.4 Zero and Few-shot Inference

For a new target RE task, the model is given $N \times K$ training examples $(x_{1}, y_{1}), \ldots, (x_{N \times K}, y_{N \times K})$ for few-shot prompting or no training examples for zero-shot prompting as well as a test input $x$. It is also given a set of candidates $C$ which is either a set of relation labels (RC) or relational triples (RTE). For few-shot setting, as in meta-training, the model takes a concatenation of $x_{1}, y_{1}, \ldots, x_{N \times K}, y_{N \times K}$ as the input, and compute the probability of each candidate $c_i \in C$. The candidate with the maximum conditional probability is returned as a prediction. For zero-shot setting, the model only takes $x$ as the input with no training examples to get the prediction.

The specific low-shot prompting forms in RC and RTE are illustrated in Figure 1. For few-shot prompting, the models easily recover the output of test input $x$ conditioning on $N \times K$ training examples. For zero-shot prompting, the situation becomes a little tricky. Since no training examples are available for in-context learning, the models are unaware of relation schema in a specific RE dataset. We transform the multi-classification form into multiple binary classification forms. Specifically, we prompt the models with each relation label $r$ to generate the corresponding subject $s$ and object $o$, then we select the correct relation $r$ or triple $z$ from $C$ candidates. In RC task, given relation $r$ and subject $s$, we consider relation $r$ as the zero-shot prediction if its probability of output object $o$ of the original annotation (for RC, we know both annotated subject and object) is the maximum. In RTE task, we prompt the models with each relation $r$ to generate multiple candidate relational triples, then the triple $z$ with the maximum conditional probability is selected as the zero-shot prediction.
4 Experiments

4.1 Experimental Design

Datasets. We use a collection of publicly RE datasets taken from [Wang et al., 2023c] and widely considered in RE research community. We have 12 unique RE datasets in total, covering general, news, disease and science domains. All these RE datasets are in English and we provide the statistics of these datasets in Table 1. Note that in meta-training phases, we only use the training set of 12 RE datasets. To balance the meta-training datasets, we sample 10,000 examples for each training set and include all examples for training sets with fewer than 10,000 samples [Wang et al., 2023c].

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Train</th>
<th>#Dev</th>
<th>#Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADE [Gurulingappa et al., 2012]</td>
<td>3,417</td>
<td>427</td>
<td>428</td>
</tr>
<tr>
<td>CoNLL2004 [Roth and Yih, 2004]</td>
<td>922</td>
<td>231</td>
<td>288</td>
</tr>
<tr>
<td>GIDS [Jat et al., 2018]</td>
<td>8,526</td>
<td>1,417</td>
<td>4,307</td>
</tr>
<tr>
<td>KBP37 [Zhang and Wang, 2015]</td>
<td>15,917</td>
<td>1,724</td>
<td>3,405</td>
</tr>
<tr>
<td>NYT24 [Riedel et al., 2010]</td>
<td>56,196</td>
<td>5,000</td>
<td>6,000</td>
</tr>
<tr>
<td>NYT11 [Takanobu et al., 2019]</td>
<td>62,648</td>
<td>149</td>
<td>369</td>
</tr>
<tr>
<td>SciERC [Luan et al., 2018]</td>
<td>1,366</td>
<td>187</td>
<td>397</td>
</tr>
<tr>
<td>SemEval [Hendrickx et al., 2019]</td>
<td>6,507</td>
<td>1,493</td>
<td>2,717</td>
</tr>
<tr>
<td>TACRED [Zhang et al., 2017]</td>
<td>68,124</td>
<td>22,631</td>
<td>15,509</td>
</tr>
<tr>
<td>ACE2004 [Doddington et al., 2004]</td>
<td>6,946</td>
<td>868</td>
<td>868</td>
</tr>
<tr>
<td>ACE2005 [Walker et al., 2005]</td>
<td>10,051</td>
<td>2,424</td>
<td>2,050</td>
</tr>
<tr>
<td>WebNLG [Gardent et al., 2017]</td>
<td>5,019</td>
<td>500</td>
<td>703</td>
</tr>
</tbody>
</table>

Table 1: Statistics of meta-training datasets.

We experiment on FewRel [Han et al., 2018] and Wiki-ZSL [Chen and Li, 2021] for low-shot experiments 1. To ensure no overlap between meta-training and target datasets, for each relation label names in meta-training datasets, we discard it if it overlaps with a relation label name in target RE datasets (i.e., two identical phrases appear in two names).

Zero-shot settings. We randomly select $m$ relations from FewRel and Wiki-ZSL as zero-shot relations [Chen and Li, 2021]. We repeat the experiment 5 times for random selection of $m$ relations, and report the average results. We also vary $m$ to examine how performance is affected. We use Precision (P), Recall (R), and Macro-F1 as the evaluation metrics for RC. For RTE, evaluating single triplet extraction involves only one possible triplet for each sentence, hence the metric used is Accuracy (Acc.) [Chia et al., 2022]. Note that the randomly sampled zero-shot relations may share similar semantics with some relations appearing in the meta-training datasets. However, since the relation schemas of meta-training and target datasets are different, we can still evaluate the zero-shot transfer learning ability of the models. In other words, the models are expected to understand the zero-shot relation semantics solely based on relation names.

Few-shot settings. We conduct experiments on the public benchmark dataset FewRel [Han et al., 2018], which releases 80 relations and each relation owns 700 triple instances in total. For RC, following the standard configuration of FewRel [Han et al., 2018], we conducted experiments in these settings: 5-way-1-shot, 5-way-5-shot, 10-way-1-shot and 10-way-5-shot. For RTE, we follow previous work [Yu et al., 2020] to adopt the 5-way-5-shot and 10-way-10-shot settings. Concretely, a relational triple is correct if and only if the spans of the head and tail entity are correctly identified and the associated relation is also predicted correctly. We adopt the standard Micro F1 score to evaluate the results and report the averages over 5 randomly initialized runs. Because the maximum length limitation of LLMs restricts to put too many in-context examples at once, we concatenate the maximum number of training examples satisfying the input length and discard the rest examples specially in 10-way-10-shot setting. And we should point out that this is one of the limitations of MicRE. More generally, in-context leaning paradigm is suitable for very few examples, making it difficult to better utilize more training examples compared to traditional methods.

Baselines. We consider both supervised fine-tuned methods and zero/few-shot prompting methods. For Zero-shot RC, we make comparisons with state-of-the-art matching-based methods ESIM [Lyu et al., 2017], ZS-BERT [Chen and Li, 2021], PromptMatch [Sainz et al., 2021] and RE-Matching [Zhao et al., 2023]. We also compare a seq2seq-based method RelationPrompt [Chia et al., 2022] and two zero-shot prompting methods Vanilla and SumAsk [Li et al., 2023] with GPT-3.5 [OpenAI, 2022]. For Zero-shot RTE, we provide three baseline methods for comparison: TableSequence [Wang and Lu, 2020], RelationPrompt [Chia et al., 2022] and ZETT [Kim et al., 2023]. For Few-shot RC, we make comparisons with the following state-of-the-art baselines including traditional few-shot learning methods ProtoNet [Snell et al., 2017] and MAML [Finn et al., 2017], besides the pre-training enhanced methods CP [Peng et al., 2020], HCRP [Han et al., 2021], LPD [Zhang and Lu, 2022], HDN [Zhang et al., 2023] and DeepStruct [Wang et al., 2022]. We also compare a recent promising chain of thought and few-shot prompting method CoT-ER [Ma et al., 2023a]. For Few-shot RTE, we select supervised learning methods FT-BERT [Devlin et al., 2019], FastRE [Li et al., 2022] and CasRel [Wei et al., 2020], besides the few-shot learning methods MPE [Yu et al., 2020], StructShot [Yang and Katiyar, 2020], PA-CRF [Cong et al., 2021], RelATE [Cong et al., 2022] and MG-FTE [Yang et al., 2023].

Experiment Details. For meta-training, we use a batch size of 4, learning rate of 1e-4 and a block size of 512, training the model for 100,000 max steps with 16-shot learning. To save memory during meta-training, we use deepspeed [Rasley et al., 2020] and adopt parameter efficient tuning technique LoRA [Hu et al., 2022] for model training with the rank $r$ to 8 and the merging ratio $\alpha$ to 32. As for the base LLMs, we use popular open-source models such as GPT-2 [Radford et al., 2019], T5 [Raffel et al., 2020] and LLaMA [Touvron et al., 2022]. Specifically, we adopt GPT-2 (770M), GPT-2-3B and LLaMA-7B for experiments. We should note that these LLMs without fine-tuning cannot perform ICL in RE. We empirically discover that they are unable to understand the structural sentences in ICL paradigm and recover the relation labels of test examples in low-shot settings.

1https://github.com/liguozheng/MicRE
### 4.2 Main Results

#### Zero-shot RC results

The main results of zero-shot RC are summarized in Table 2, where LLMs with meta in-context training achieve competitive results compared to supervised fine-tuned RC methods and zero-shot prompting RC methods over two datasets when varying numbers of unseen relations. We have two findings about the model parameter scales and overall performances. First, the larger the model scale, the more notable the enhancement in performance. With small model scale, MCRE with GPT-2 even underperforms ESIM. Second, encoder-decoder models seem to achieve better ICL performance than decoder-only models with similar model scales (e.g., GPT-2-large and T5-large), which is due to the positive role of encoders in language understanding tasks. For the zero-shot RC task, as $m$ increases, it is straightforward that models are difficult to predict the right relation since the possible choices have increased. Notably, the proposed MCRE with LLaMA delivers superior results compared to state-of-the-art method RE-Matching when dealing with more unseen relations. Such results not only validate the effectiveness of meta in-context training, but indicate MCRE is less sensitive to the number of relations compared to baselines. Another finding is that the recall of MCRE with LLaMA achieves the best or second best results in 5 out of 6 settings, which may be related to our zero-shot prompting strategy. Because we enumerate each relation and subject to prompt MCRE to generate its corresponding object, ensuring the final recall but slightly harming the final precision.

#### Zero-shot RTE results

The main results of zero-shot RTE by varying $m$ unseen relations on FewRel and Wiki-ZSL are summarized in Table 3, where MCRE with foundation models that have more than 1B parameters consistently outperforms existing methods across different settings. LLaMA achieves up to 9.45 and 6.87 higher accuracy than the existing state-of-the-art model, ZETT on Wiki-ZSL and FewRel datasets, respectively. Because extracting relational triples from texts is a challenging structure prediction task, the meta-training makes MCRE recover the semantics of the RTE task during zero-shot inference, where the model output is very similar with the meta-training output. However, we note that with the same foundation models, MCRE achieves less satisfied performance compared to existing baselines. Specifically, the average scores of MCRE with GPT-2 and T5-base are 10.22 and 18.55, respectively, where RelationPrompt uses GPT-2 and ZETT uses T5-base but they all delivers better results. This indicates that achieving noticeable ICL results in LLMs requires the relatively large-scale model parameters. As the model scale increases, the advantages of MCRE become more apparent, where T5-3B and LLaMA both show much better performances than previous methods.

### Few-shot RC results

The main results of few-shot RC are summarized in Table 4 (left). We observe strong few-shot RC performance of MCRE on FewRel. This suggests that the meta-in-context training is beneficial in low-resource regimes via transferring knowledge from similar tasks. First, compared to few-shot and pre-training enhanced RC methods
Table 4: Few-shot RC results (left) and few-shot RTE results (right) on FewRel. Best results are in **bold** and the second best results are marked with underline. Avg. denotes the average of all the Micro-F1 scores.

(e.g., LPD and HDN), MIRC with LLaMA achieve more superior performances, and MIRC with smaller base models can show competitive results. Second, compared to CoT-ER which is based on elaborated few-shot prompting and GPT-3, MIRC with LLaMA basically delivers similar average scores, which indicates that meta in-context training successfully boosts the ICL abilities of open-source LLMs in RE. Third, MIRC lags behind DeepStruct which is based on a pre-trained 10B parameter encoder-decoder language model GLM [Du et al., 2022] and is pre-trained on a collection of large-scale RE corpus. Despite its excellent performance with multi-task fine-tuning, DeepStruct produces unattractive zero-shot transferring ability after pre-training [Wang et al., 2022]. In contrast, MIRC appears to be able to adapt to new data, presenting a fair and strong performance with in-context learning. Another explanation is that in-context training examples help the model better understand the new RE tasks, such as the concrete output format of each task. Finally, MIRC with GPT-2 and T5-base still cannot surpass classic few-shot methods ProtoNet and MAML. Compared to pre-train then fine-tune paradigm, showing in-context learning ability tends to require large-scale model parameters. As the exceptional performance of 10B GLM based DeepStruct, meta-training on larger base models may bring better in-context learning results and is worth exploring in the future.

### Few-shot RTE results

Table 4 (right) reports the few-shot RTE results of MIRC against other baseline models on FewRel. It can be seen that, overall, MIRC with LLMs significantly outperforms all competitive methods and achieves new state-of-the-art in two few-shot settings, which highlights the pivot role of meta in-context training. Notably, even with GPT-2 and T5-base, MIRC still outperforms most strong baselines such as PA-CRF and RelATE. Similar with zero-shot RTE results, the few-shot RTE results seem to prove that generative methods are more effective in handling the low-shot RTE task. Moreover, current few-shot RTE methods first meta-train on the subset of entire dataset and then are evaluated on the test set. Thus the training and testing data are in the similar distribution and same domain. But MIRC showcases notable distribution and domain adaptation ability.

#### 4.3 Discussions

**Number of in-context training examples.** We vary the number of training examples $k$ from 0, 4, 8, 16 to 32. In-context learning with $k=0$ is equivalent to the zero-shot method. Results are shown in Figure 2. Generally, increasing $k$ helps across all tasks and settings. Besides, the few-shot setting seems to have higher variance than zero-shot setting. And increasing $k$ consistently reduces the average performance variance. Especially, for zero-shot cases, performance tends to stabilize when $k$ is greater than 8. This indicates that zero-shot prompting is less sensitive to the number of in-context training examples during meta-training as no training example is provided at inference time. In contrast, meta-training with more examples brings significant improvements for few-shot learning. However, we additionally find that the performance tends to saturate when $k$ is closer to 16 [Min et al., 2022]. The saturate phenomenon is likely because the sequence length limit of the language model makes it hard to encode many training examples, which is one of the limitations of the attention technique [Vaswani et al., 2017].

**Number of meta-training datasets.** To see the impact of the number of meta-training datasets, we subsample 1, 4, 8 meta-training datasets out of 12 in four experimental settings. For each, we use three different random seeds to additionally see the impact of the choice of meta-training datasets. Figure 2 shows the results. On average, low-shot performances generally increase as the number of datasets increase, which is consistent with results in previous work [Mishra et al., 2022; Wei et al., 2022a; Min et al., 2022]. Nonetheless, different choices of meta-training datasets brings nonnegligible variance, indicating that a choice of meta-training gives substantial impact in performance. This can be attributed to multiple reasons. On the one hand, the varying amount of training data in different datasets leads to the training effectiveness of the model. On the other hand, the data domains
and distributions between meta-training datasets and target datasets also play a key role in model performances.

**Semantic hints of similar relations.** Although we ensure that the data distributions and relation schemas during the meta-training and inference phases are different, in fact, there are inherently similar semantic relations between meta-training and inference relation sets. We use relation label words taken from the original datasets, which contain semantic hints that express what each relation label is supposed to mean. If the model is truly learning the relation in-context, it should generalize when label words are replaced with other English words, e.g., `date of birth` is replaced with token `R1`, thus not giving any hints about the relation semantics. To this end, we substitute each relation label in meta-training with `Ri`, `i ∈ [1, I]` where `I` is the total number of relations in meta-training datasets. At inference time, we also perform similar operations to evaluate the in-context relation learning ability. The results are summarized in Table 5. Note that when test labels are replaced, MiCRE is unable to perform zero-shot RC and RTE using tabular prompting. First, with test labels be replaced, the overall results suffer grave declines. This indicates that having semantic hints from relation label words is a necessary condition for LLMs to perform low-shot RE tasks. Compared to original LLaMA, meta-training on replaced training labels consistently delivers considerable improvements in few-shot RC and RTE, where MiCRE actually benefits from training on the replaced data and improves its in-context learning ability on new RE tasks. Still, overall performance is relatively poor compared to training on original labels, which implies that learning from relation label words helps the model better capture semantic differences between various relations. And the model can utilize the relation semantic knowledge during inference on target RE datasets.

**Error analysis.** We categorize three types of incorrectly predicted unseen relations for analysis and provide an example illustrated in Figure 3. (1) The true relation is not appropriate because it comes from distant supervision. It shows the noise originated from distant labeling. That is, we cannot identify the relation between `Elizbeth` and `English is language of work or name` in this specific sentence. They just happened to appear together and their relation recorded in Wikidata is `language of work or name`. (2) The predicted relation is ambiguous because it is hard to identify the order of subject and object. The golden relation and predicted relation have very similar semantics because they are reciprocal in FewRel. Unfortunately, MiCRE frequently reverses the subject and object corresponding to these two relations because it treats the relational triple (`Lisava river, tributary, Natra river`) as (`Lisava river, tributary of, Natra river`). This indicates that relying solely on relation label names may bring ambiguity. (3) The predicted relation is not precise for the targeted entity pair but may be suitable for other entities that also appear in the sentence. The targeted entities are `Anaconda` and `Minaj` as the prediction, which is actually correct if the targeted entities are `Anaconda` and `Minaj`. This shows MiCRE is able to infer the possible relation for entities in the given sentence. When we prompt the model with relation `publisher`, the model output the subject `Anaconda` and object `Minaj` with the max probability. As they are all valid spans in original sentence, we consider `publisher` as the true relation. This also hinders the capability of MiCRE in extracting overlapping relational triples. More general methods are worth exploring in the future.

### Table 5: Ablation about semantic hints of similar relations. Original and Replaced indicate original label words and labels that are replaced to special tokens, respectively. The first row denotes the LLaMA performance without meta-training. We report Avg. scores.

<table>
<thead>
<tr>
<th>Train labels</th>
<th>Test labels</th>
<th>Zero-shot RC</th>
<th>Zero-shot RTE</th>
<th>Few-shot RC</th>
<th>Few-shot RTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>Original</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Original</td>
<td>Original</td>
<td>77.65</td>
<td>29.89</td>
<td>94.31</td>
<td>58.29</td>
</tr>
<tr>
<td>Original</td>
<td>Replaced</td>
<td>-</td>
<td>-</td>
<td>68.64</td>
<td>22.11</td>
</tr>
<tr>
<td>Replaced</td>
<td>Original</td>
<td>28.62</td>
<td>1.74</td>
<td>84.45</td>
<td>45.44</td>
</tr>
<tr>
<td>Replaced</td>
<td>Replaced</td>
<td>-</td>
<td>-</td>
<td>58.43</td>
<td>15.38</td>
</tr>
</tbody>
</table>

Figure 2: Ablation on the number of training examples `k` (left) and meta-training datasets `C` (right) in four settings. The metric of each setting corresponds to its Avg. score.

Figure 3: List of three cases. The entities are highlighted in color.

### 5 Conclusion

In this work, we introduce MiCRE, a new zero and few-shot learning method where an LLM is meta-trained to learn to in-context learn relations, i.e. condition on training examples to recover the new relation semantics and make predictions. MiCRE outperforms a range of strong baselines including supervised fine-tuning and in-context learning without meta-training methods. Besides, it achieves competitive results compared to current state-of-the-art task-specific models. We also analyze the advantages and limitations of MiCRE, encouraging more effective methods in the future research.
Acknowledgments

We thank the reviewers for their insightful comments. This work was supported by National Science Foundation of China (Grant Nos.62376057) and the Start-up Research Fund of Southeast University (RF1028623234). All opinions are of the authors and do not reflect the view of sponsors.

References


[Cong et al., 2022] Xin Cong, Jiawei Sheng, et al. Relation-guided few-shot relational triple extraction. In SIGIR, 2022.


[Han et al., 2021] Jiale Han, Bo Cheng, and Wei Lu. Exploring task difficulty for few-shot relation extraction. In EMNLP, 2021.


[Ma et al., 2023b] Yubo Ma, Yixin Cao, et al. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In Findings of EMNLP, 2023.


