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

Guozheng Li[1] , **Peng Wang**[1,2*] , **Jiajun Liu**[1] , **Yikai Guo**[3] , **Ke Ji**[1] , **Ziyu Shang**[1] and **Zijie Xu**[1]
[1]School of Computer Science and Engineering, Southeast University
[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education
[3]Beijing 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 MICRE (**M**eta **I**n-**C**ontext learning of LLMs for **R**elation **E**xtraction), 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 MICRE 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. MICRE 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 MICRE 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 MICRE: **M**eta **I**n-**C**ontext

*Corresponding author

learning of LLMs for **R**elation **E**xtraction. 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 reformatting (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 12 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

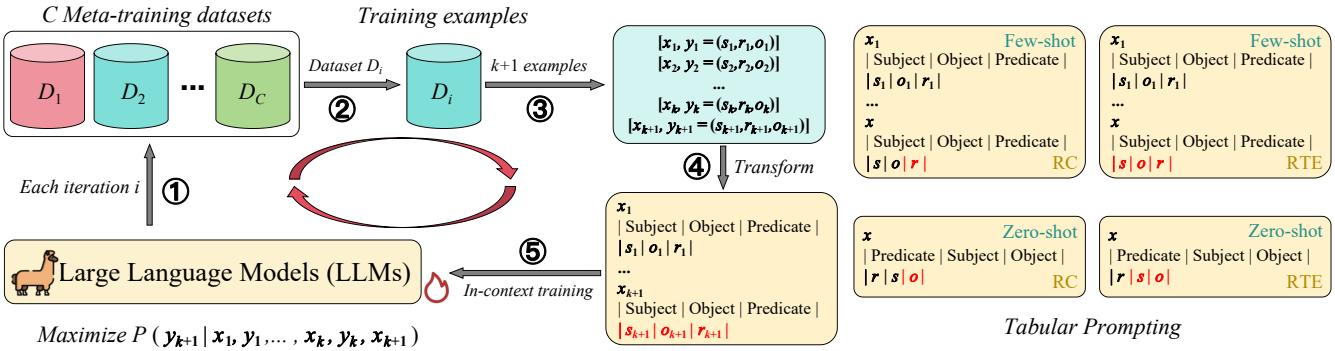

Figure 1: Overview of the meta-training work flow of MICRE. The output of LLMs are highlighted in red.

s_i is the subject and o_i is the object. This task aims to identify a relation r_i from pre-defined relation set $\mathcal{R}$ 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 $\mathcal{R}$. (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 $\mathcal{D}$ and the target sentence set $\mathcal{T}$, for N way (relation types), the model utilizes randomly K examples for each relation type from $\mathcal{D}$ to form the support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N\times K}$, 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 $\mathcal{S}$ for prediction and outputs recognized relations in RC and triples in RTE for each target sentence in $\mathcal{T}$.

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 an 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 MICRE 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 $(x_1, y_1), ..., (x_{k+1}, y_{k+1})$ 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 $x_1, y_1, ..., x_k, y_k, x_{k+1}$ 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 $(x_1, y_1), ..., (x_{N\times K}, y_{N\times K})$ 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 $\mathcal{C}$ 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 $x_1, y_1, ..., x_{N\times K}, y_{N\times K}, x$ as the input, and compute the probability of each candidate $c_i \in \mathcal{C}$. 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 unware 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 $\mathcal{C}$ 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].

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

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 [Levy *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 α 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.*, 2023]. Specifically, we adopt GPT-2 (117M), GPT-2-large (770M), GPT-2-XL (1.5B), T5-base (220M), T5-large (770M), T5-3B 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.

[1] https://github.com/1iguozheng/Micre

Method	Wiki-ZSL m=5			Wiki-ZSL m=10			Wiki-ZSL m=15			FewRel m=5			FewRel m=10			FewRel m=15			Avg.
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
ESIM	48.58	47.74	48.16	44.12	45.46	44.78	27.31	29.62	28.42	56.27	58.44	57.33	42.89	44.17	43.52	29.15	31.59	30.32	42.09
ZS-BERT	71.54	72.39	71.96	60.51	60.98	60.74	34.12	34.38	34.25	76.96	78.86	77.90	56.92	57.59	57.25	35.54	38.19	36.82	56.49
PromptMatch	77.39	75.90	76.63	71.86	71.14	71.50	62.13	61.76	61.95	91.14	90.86	91.00	83.05	82.55	82.80	72.83	72.10	72.46	76.06
RelationPrompt	70.66	83.75	76.63	68.51	74.76	71.50	63.69	67.93	65.74	90.15	88.50	89.30	80.33	79.62	79.96	**74.33**	72.51	73.40	76.09
RE-Matching	**78.19**	78.41	**78.30**	**74.39**	73.54	**73.96**	**67.31**	67.33	67.32	**92.82**	**92.34**	**92.58**	**83.21**	**82.64**	**82.93**	73.80	73.52	73.66	**78.13**
Vanilla w/ GPT-3.5	64.47	70.83	67.50	41.83	46.22	43.92	23.17	27.82	25.28	67.41	72.97	70.08	42.48	46.26	44.29	25.71	27.77	26.70	46.30
SumAsk w/ GPT-3.5	75.64	70.96	73.23	62.31	61.08	61.69	43.55	40.27	41.85	78.27	72.55	75.30	64.77	60.94	62.80	44.76	41.13	42.87	59.62
MICRE w/ GPT-2	44.63	48.63	46.54	37.18	37.88	37.53	25.74	26.62	26.17	50.52	54.94	52.64	41.83	43.24	42.52	32.64	33.52	33.07	39.75
MICRE w/ GPT-2-large	68.22	70.50	69.34	64.78	65.62	65.20	54.51	55.87	55.18	80.64	83.52	82.05	68.41	69.99	69.19	60.55	62.68	61.60	67.09
MICRE w/ GPT-2-XL	70.75	73.66	72.18	67.43	69.75	68.57	62.30	63.64	62.96	86.74	89.23	87.97	77.85	78.73	78.29	68.60	69.52	69.06	73.17
MICRE w/ T5-base	56.53	58.66	57.58	45.73	47.28	46.49	26.74	27.31	27.02	73.24	76.54	74.85	65.10	66.76	65.92	43.62	44.60	44.10	52.66
MICRE w/ T5-large	70.25	72.55	71.38	65.37	67.44	66.39	57.45	58.39	57.92	82.43	84.63	83.52	71.72	73.51	72.60	60.57	61.98	61.27	68.85
MICRE w/ T5-3B	74.75	77.36	76.03	70.64	71.89	71.30	64.11	65.43	64.76	88.23	89.77	88.99	78.82	80.53	79.67	70.28	71.69	70.98	75.29
MICRE w/ LLaMA	76.46	**78.53**	77.48	72.36	**74.88**	73.60	67.14	**68.87**	**67.99**	89.34	91.88	90.59	80.67	82.31	81.48	73.74	**75.83**	**74.77**	77.65

Table 2: Zero-shot RC results on Wiki-ZSL and FewRel datasets. Best results are in **bold** and the second best results are marked with underline. **Avg.** denotes the average of all the Macro-F1 scores in six settings.

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 datasests 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, MICRE with GPT-2 even underperforms ESIM. Second, encoder-decoder models seem to achieve better ICL performance than decoder-only models with similar model scales (eg., 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 MICRE 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 MICRE is less sensitive to the number of relations compared to baselines. Another finding is that the recall of MICRE 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 MICRE 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 MICRE 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 MICRE 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, MICRE achieves less satisfied performance compared to existing baselines. Specifically, the average scores of MICRE 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 MICRE become more apparent, where T5-3B and LLaMA both show much better performances than previous methods.

Method	Wiki-ZSL m=5	Wiki-ZSL m=10	Wiki-ZSL m=15	FewRel m=5	FewRel m=10	FewRel m=15	Avg.
TableSequence	14.47	9.61	9.20	11.82	12.54	11.65	11.55
RelationPrompt	16.74	12.13	10.47	24.36	21.45	20.24	17.57
ZETT	21.49	17.27	12.78	30.71	27.90	26.17	22.72
MICRE w/ GPT-2	14.77	10.62	6.90	12.88	9.20	6.93	10.22
MICRE w/ GPT-2-large	19.50	17.73	14.03	28.36	24.75	17.66	20.34
MICRE w/ GPT-2-XL	21.56	19.61	15.75	31.97	28.44	21.20	23.09
MICRE w/ T5-base	17.43	15.35	11.66	27.34	23.54	15.98	18.55
MICRE w/ T5-large	20.64	17.87	14.53	29.58	26.40	18.05	21.18
MICRE w/ T5-3B	25.20	23.65	21.80	36.75	33.18	30.44	28.50
MICRE w/ LLaMA	**27.74**	**24.64**	**22.23**	**37.53**	**34.77**	**32.42**	**29.89**

Table 3: Zero-shot RTE results on Wiki-ZSL and FewRel datasets. Best results are in **bold** and the second best results are marked with underline. **Avg.** denotes the average of all the accuracy scores.

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 MICRE 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

Method	5-way-1-shot	5-way-5-shot	10-way-1-shot	10-way-5-shot	Avg.	Method	5-way-5-shot	10-way-10-shot	Avg.
ProtoNet	82.92	91.32	73.24	83.68	82.79	FT-BERT	4.71	2.94	3.83
MAML	82.93	86.21	73.20	76.06	79.60	FastRE	7.73	6.82	7.28
CP	88.29	92.77	80.50	88.61	87.54	CasRel	2.11	2.04	2.08
HCRP	90.89	92.90	83.17	86.43	88.35	MPE	23.34	12.08	17.71
LPD	93.51	94.33	87.77	89.19	91.20	StructShot	25.94	20.28	23.11
HDN	95.46	96.59	89.34	92.46	93.46	PA-CRF	34.14	30.44	32.29
DeepStruct	**98.40**	**100.00**	**97.80**	**99.80**	**99.00**	RelATE	42.32	40.93	41.63
CoT-ER w/ GPT-3	97.40	97.00	92.10	94.70	95.30	MG-FTE	55.17	53.33	54.25
MICRE w/ GPT-2	71.53	72.33	65.84	66.15	68.96	-	43.44	40.64	42.04
MICRE w/ GPT-large	85.37	85.98	78.34	80.46	82.54	-	46.36	44.20	45.28
MICRE w/ GPT-XL	89.38	90.58	82.10	83.74	87.45	-	52.66	50.82	51.74
MICRE w/ T5-base	76.24	78.92	72.75	73.55	75.37	-	45.08	44.00	44.54
MICRE w/ T5-large	87.77	88.46	79.93	82.64	84.70	-	48.64	47.59	48.12
MICRE w/ T5-3B	93.77	94.30	88.45	89.56	91.52	-	56.35	53.57	54.96
MICRE w/ LLaMA	95.74	97.11	93.36	94.25	95.12	-	**59.82**	**56.75**	**58.29**

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), MICRE with LLaMA achieve more superior performances, and MICRE with smaller base models can show competitive results. Second, compared to CoT-ER which is based on elaborated few-shot prompting and GPT-3, MICRE 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, MICRE 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, MICRE 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, MICRE 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 MICRE against other baseline models on FewRel. It can be seen that, overall, MICRE 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, MICRE 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 MICRE 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

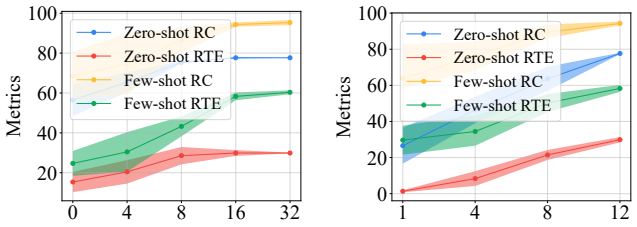

Figure 2: Ablation on the number of training examples k (left) and meta-training datasets C (right) in four settings. The metric of each setting corresponds to its **Avg.** score.

Train labels	Test labels	Zero-shot		Few-shot	
		RC	RTE	RC	RTE
-	Original	0.00	0.00	0.00	0.00
Original	Original	77.65	29.89	94.31	58.29
Original	Replaced	-	-	68.64	22.11
Replaced	Original	28.62	1.74	84.45	45.44
Replaced	Replaced	-	-	58.43	15.38

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.

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 inevitably 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 $Ri, i \in [1, I]$ 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 task. 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.

Case 1: Annabeth is a female **English** given name created from a combination of the names anna and **Elizabeth**. Ground Truth: ***language of work or name*** Prediction: ***country***
Case 2: The Natra river is a tributary of the **Lisava river** in romania. Ground Truth: ***tributary*** Prediction: ***mouth of the watercourse***
Case 3: When promoting **Anaconda**, **Minaj** confirmed plans of a tour in support of **The Pinkprint** in an interview with Carson Daly on AMP Radio. Ground Truth: ***tracklist*** Prediction: ***publisher***

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

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 *Elizabeth* 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 *The Pinkprint*, and MICRE yields *publisher* 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.

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

[Agrawal *et al.*, 2022] Monica Agrawal, Stefan Hegselmann, et al. Large language models are few-shot clinical information extractors. In *EMNLP*, 2022.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[Chen and Li, 2021] Chih-Yao Chen and Cheng-Te Li. ZS-BERT: Towards zero-shot relation extraction with attribute representation learning. In *NAACL-HLT*, 2021.

[Chen *et al.*, 2022] Yanda Chen, Ruiqi Zhong, et al. Meta-learning via language model in-context tuning. In *ACL*, 2022.

[Chia *et al.*, 2022] Yew Ken Chia, Lidong Bing, et al. Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In *Findings of ACL*, 2022.

[Cong *et al.*, 2021] Xin Cong, Shiyao Cui, et al. Few-shot event detection with prototypical amortized conditional random field. In *Findings of ACL*, 2021.

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

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[Doddington *et al.*, 2004] George R Doddington, Alexis Mitchell, et al. The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, 2004.

[Du *et al.*, 2022] Zhengxiao Du, Yujie Qian, et al. GLM: General language model pretraining with autoregressive blank infilling. In *ACL*, 2022.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[Gardent *et al.*, 2017] Claire Gardent, Anastasia Shimorina, et al. The webnlg challenge: Generating text from rdf data. In *INLG*, 2017.

[Gurulingappa *et al.*, 2012] Harsha Gurulingappa, Abdul Mateen Rajput, et al. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892, 2012.

[Han *et al.*, 2018] Xu Han, Hao Zhu, et al. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *EMNLP*, 2018.

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

[Hendrickx *et al.*, 2019] Iris Hendrickx, Su Nam Kim, et al. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *arXiv preprint arXiv:1911.10422*, 2019.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

[Jat *et al.*, 2018] Sharmistha Jat, Siddhesh Khandelwal, and Partha Talukdar. Improving distantly supervised relation extraction using word and entity based attention. *arXiv preprint arXiv:1804.06987*, 2018.

[Ji *et al.*, 2023] Ke Ji, Yixin Lian, et al. Hierarchical verbalizer for few-shot hierarchical text classification. In *ACL*, 2023.

[Kim *et al.*, 2023] Bosung Kim, Hayate Iso, et al. Zero-shot triplet extraction by template infilling. In *AACL*, 2023.

[Kojima *et al.*, 2022] Takeshi Kojima, Shixiang Shane Gu, et al. Large language models are zero-shot reasoners. In *NeurIPS*, 2022.

[Levy *et al.*, 2017] Omer Levy, Minjoon Seo, et al. Zero-shot relation extraction via reading comprehension. In *CoNLL*, 2017.

[Li *et al.*, 2022] Guozheng Li, Xu Chen, et al. Fastre: Towards fast relation extraction with convolutional encoder and improved cascade binary tagging framework. In *IJCAI*, 2022.

[Li *et al.*, 2023] Guozheng Li, Peng Wang, and Wenjun Ke. Revisiting large language models as zero-shot relation extractors. In *Findings of EMNLP*, 2023.

[Li *et al.*, 2024] Guozheng Li, Wenjun Ke, et al. Unlocking instructive in-context learning with tabular prompting for relational triple extraction. *arXiv preprint arXiv:2402.13741*, 2024.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Liu *et al.*, 2024] Jiajun Liu, Wenjun Ke, et al. Towards continual knowledge graph embedding via incremental distillation. In *AAAI*, 2024.

[Luan *et al.*, 2018] Yi Luan, Luheng He, et al. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602*, 2018.

[Ma *et al.*, 2023a] Xilai Ma, Jing Li, and Min Zhang. Chain of thought with explicit evidence reasoning for few-shot relation extraction. In *Findings of EMNLP*, 2023.

[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.

[Min *et al.*, 2022] Sewon Min, Mike Lewis, et al. Metaicl: Learning to learn in context. In *NAACL-HLT*, 2022.

[Mishra *et al.*, 2022] Swaroop Mishra, Daniel Khashabi, et al. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*, 2022.

[OpenAI, 2022] OpenAI. Introducing chatgpt, 2022.

[Peng *et al.*, 2020] Hao Peng, Tianyu Gao, et al. Learning from context or names? an empirical study on neural relation extraction. In *ACL*, 2020.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(1):5485–5551, 2020.

[Rasley *et al.*, 2020] Jeff Rasley, Samyam Rajbhandari, et al. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *SIGKDD*, 2020.

[Riedel *et al.*, 2010] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *ECML-PKDD*, 2010.

[Roth and Yih, 2004] Dan Roth and Wen-tau Yih. A linear programming formulation for global inference in natural language tasks. In *CoNLL*, 2004.

[Sainz *et al.*, 2021] Oscar Sainz, Oier Lopez de Lacalle, et al. Label verbalization and entailment for effective zero- and few-shot relation extraction. In *EMNLP*, 2021.

[Sanh *et al.*, 2022] Victor Sanh, Albert Webson, et al. Multitask prompted training enables zero-shot task generalization. In *ICLR*, 2022.

[Shang *et al.*, 2024] Ziyu Shang, Wenjun Ke, et al. Ontofact: Unveiling fantastic fact-skeleton of llms via ontology-driven reinforcement learning. In *AAAI*, 2024.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS*, 2017.

[Takanobu *et al.*, 2019] Ryuichi Takanobu, Tianyang Zhang, et al. A hierarchical framework for relation extraction with reinforcement learning. In *AAAI*, 2019.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, et al. Attention is all you need. In *NIPS*, 2017.

[Walker *et al.*, 2005] Christopher Walker, Stephanie Strassel, et al. Ace 2005 multilingual training corpus-linguistic data consortium. *URL: https://catalog. ldc. upenn. edu/LDC2006T06*, 2005.

[Wan *et al.*, 2023] Zhen Wan, Fei Cheng, et al. GPT-RE: In-context learning for relation extraction using large language models. In *EMNLP*, 2023.

[Wang and Lu, 2020] Jue Wang and Wei Lu. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *EMNLP*, 2020.

[Wang *et al.*, 2022] Chenguang Wang, Xiao Liu, et al. Deepstruct: Pretraining of language models for structure prediction. In *Findings of ACL*, 2022.

[Wang *et al.*, 2023a] Peng Wang, Tong Shao, et al. fmlre: a low-resource relation extraction model based on feature mapping similarity calculation. In *AAAI*, 2023.

[Wang *et al.*, 2023b] Peng Wang, Jiafeng Xie, et al. Pascore: a chinese overlapping relation extraction model based on global pointer annotation strategy. In *IJCAI*, 2023.

[Wang *et al.*, 2023c] Xiao Wang, Weikang Zhou, et al. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*, 2023.

[Wei *et al.*, 2020] Zhepei Wei, Jianlin Su, et al. A novel cascade binary tagging framework for relational triple extraction. In *ACL*, 2020.

[Wei *et al.*, 2022a] Jason Wei, Maarten Bosma, et al. Finetuned language models are zero-shot learners. In *ICLR*, 2022.

[Wei *et al.*, 2022b] Jason Wei, Xuezhi Wang, et al. Chain of thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

[Yang and Katiyar, 2020] Yi Yang and Arzoo Katiyar. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In *EMNLP*, 2020.

[Yang *et al.*, 2023] Chengmei Yang, Shuai Jiang, et al. Mutually guided few-shot learning for relational triple extraction. In *ICASSP*, 2023.

[Yu *et al.*, 2020] Haiyang Yu, Ningyu Zhang, et al. Bridging text and knowledge with multi-prototype embedding for few-shot relational triple extraction. In *COLING*, 2020.

[Zhang and Lu, 2022] Peiyuan Zhang and Wei Lu. Better few-shot relation extraction with label prompt dropout. In *EMNLP*, 2022.

[Zhang and Wang, 2015] Dongxu Zhang and Dong Wang. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*, 2015.

[Zhang *et al.*, 2017] Yuhao Zhang, Victor Zhong, et al. Position-aware attention and supervised data improve slot filling. In *EMNLP*, 2017.

[Zhang *et al.*, 2023] Liang Zhang, Chulun Zhou, et al. HyperNetwork-based decoupling to improve model generalization for few-shot relation extraction. In *EMNLP*, 2023.

[Zhao *et al.*, 2021] Zihao Zhao, Eric Wallace, et al. Calibrate before use: Improving few-shot performance of language models. In *ICML*, 2021.

[Zhao *et al.*, 2023] Jun Zhao, WenYu Zhan, et al. RE-matching: A fine-grained semantic matching method for zero-shot relation extraction. In *ACL*, 2023.