Prototype Tuning: A Meta-Learning Approach for Few-Shot Document-Level Relation Extraction with Large Language Models

Dinghao Pan, Yuanyuan Sun*, Bo Xu, Jiru Li, Zhihao Yang,
Ling Luo, Hongfei Lin, Jian Wang
School of Computer Science and Technology, Dalian University of Technology, China
{dinghaopan, jiruli}@mail.dlut.edu.cn
{syuan, xubo, yangzh, lingluo, hflin, wangjian}@dlut.edu.cn

Abstract

Few-Shot Document-Level Relation Extraction (FSDLRE) aims to develop models capable of generalizing to new categories with minimal support examples. Although Large Language Models (LLMs) demonstrate exceptional In-Context Learning (ICL) capabilities on many few-shot tasks, their performance on FSDLRE tasks remains suboptimal due to the significant gap between the task format and the intrinsic capabilities of language models, coupled with the complexity of ICL prompts for document-level text. To address these challenges, we introduce a novel meta-training approach for LLMs termed Prototype Tuning. We construct simulated episodes using data with relation types that do not overlap with the test corpus, fundamentally enhancing the ICL capabilities of LLMs in FSDLRE through meta-learning. To further enhance the effects of meta-learning, we innovatively integrate the concept of prototype into the fine-tuning process of LLMs. This involves aggregating entity pairs from support documents into prototypes within the prompts and altering the way of determining relation categories to identifying the closest prototype. Experimental results demonstrate that our LLMs trained with this approach outperform all baselines. Our proposed approach markedly improves the ICL capabilities of LLMs in FSDLRE and mitigates the impact of relation semantic discrepancies between the training corpus and the test corpus on model performance.

1 Introduction

Document-level Relation Extraction (DocRE) (Yao et al., 2019; Xu et al., 2021a) aims to extract structured knowledge from unstructured documents. This task is more complex than sentence-level relation extraction due to phenomena like coreference and cross-sentence relationships, but it more closely resembles real-world scenarios. The development of DocRE is crucial for downstream applications such as knowledge graph construction and question-answering systems. However, fully-supervised DocRE tasks face challenges such as high annotation costs, lack of domain-specific training data and long-tail distributions of labels. These issues spur many researchers to shift towards Few-Shot Document-Level Relation Extraction (FSDLRE) (Popovic and Färber, 2022; Meng et al., 2023), aiming to train models that can better utilize a small number of annotated documents to adapt to new categories and new domains compared to supervised models.

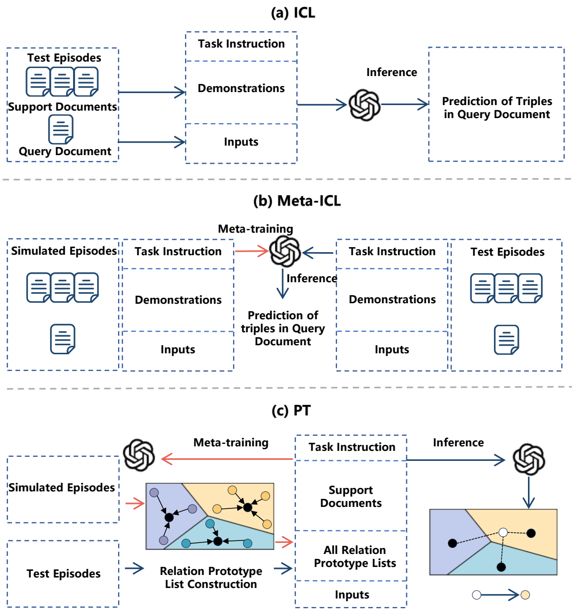


Figure 1: Three approaches for FSDLRE: In-Context Learning (ICL), MetaICL, and Prototype Tuning (PT). Red arrows depict the meta-training process, while blue arrows show the inference process in testing episodes.

Large language models (LLMs) demonstrate outstanding performance across various few-shot tasks (Brown et al., 2020; Wei et al., 2022). Previous studies explore the application of LLMs in few-shot information extraction tasks (Ma et al., 2023a;

[*] Corresponding author

Findings of the Association for Computational Linguistics: NAACL 2025, pages 1112–1128
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

Wadhwa et al., 2023; Ma et al., 2023c). However, LLMs without specific training perform suboptimally on FSDLRE (Meng et al., 2023). Earlier research often employs In-Context Learning (ICL) (Min et al., 2022b) to leverage annotated information from support samples. For document-level texts, this approach constructs lengthy inputs, increasing the model's difficulty in understanding instructions and context. Moreover, merely relying on the model's contextual understanding ability and limited annotations in few-shot support documents does not fully exploit the potential of LLMs in FSDLRE tasks. We aim to design a training approach that enables models to learn how to select truly useful information from complex ICL prompts based on the implicit patterns of the task, this would fundamentally enhance the ICL ability of LLMs in FSDLRE.

To address the aforementioned challenge, we introduce meta-learning to enhance the ICL capabilities of LLMs in FSDLRE task. We construct simulated ICL episodes using training corpus composed of relation categories that do not overlap with those in the test corpus. This approach is named MetaICL, enables the model to learn how to extract genuinely useful knowledge from the context based on existing annotations, thereby boosting its ICL abilities. However, while this approach effectively acquaints models with task-specific patterns, it predisposes them to generate responses biased towards specific relation categories encountered during training. The conflict in relational semantics between the training and testing corpus limits the effectiveness of this approach in fully enhancing the ICL capabilities of LLMs.

Prototypical Networks (Snell et al., 2017) is a classic metric-based method used in few-shot classification tasks. Instead of the traditional method of adding neurons to a classifier and training with minimal data, this method calculates the similarity between query samples and class prototypes. These prototypes are constructed from support samples of the same category. By doing this, Prototypical Networks effectively reduce overfitting issues that are specific to certain categories and improve performance of models in few-shot scenarios. Although this method relies on sample vectorization and is incompatible with language models that target token generation probabilities from a vocabulary, it offers heuristic value in addressing semantic conflicts of relation labels within meta-training.

In this paper, we propose a new few-shot LLMs fine-tuning approach based on meta-learning, named Prototype Tuning, which applies meta-learning and prototype matching concepts to LLMs. This approach enables LLMs to learn how to better perform ICL in FSDLRE tasks based on the sparse annotations of support samples within an episode. Specifically, we first construct simulated episodes with support and query samples based on a training corpus that does not overlap with the test set categories. Then we incorporate the concept of prototypes in prompt construction by grouping triples from support documents with the same target relation category into a relation prototype set. We also modify traditional task instructions and labeling formats for relation extraction, guiding LLMs to output entity pairs that belong to a specific prototype based on the similarity between the candidate entity pairs in the query documents and the relation category prototypes. By using Prototypical ICL prompts and meta-training, LLMs demonstrate enhanced generalization capabilities to new categories and domains, as well as superior ICL performance on the FSDLRE task. In Figure 1, we illustrate the elements of the FSDLRE task and the brief workflows of ICL, MetaICL, and Prototype Tuning for addressing this few-shot task.

In summary, our main contributions are as follows: (1) We propose Prototype Tuning, which uses meta-learning to enhance the ICL ability of LLMs for FSDLRE. This approach fully leverages the potential of LLMs to learn from complex prompts constructed from few-shot annotated documents to extract relations in target documents. (2) In Prototype Tuning, we introduce the concept of prototypes into the meta-training of generative models. By aggregating instances with the same relationship type from support samples and changing the way candidate entity pairs are classified in the responses, we effectively mitigate semantic conflicts between training and testing data relationships in meta-training. (3) Extensive experiments demonstrate that our approach consistently outperforms both meta-trained and non-meta-trained LLMs using the ICL approach, and significantly surpasses the state-of-the-art models for FSDLRE tasks. Further analysis shows that our approach exhibits strong robustness.

2 Related Work

Few-shot Document Level Relation Extraction. Current research on DocRE primarily focuses on supervised learning models, utilizing graph-based (Zeng et al., 2020; Xu et al., 2021b; Duan et al., 2022; Lu et al., 2023) and transformer-based (Xu et al., 2021a; Tan et al., 2022a; Xiao et al., 2022; Xie et al., 2022; Ma et al., 2023b) approaches to handle complex interactions between entities. While these methods perform well on large annotated datasets, they struggle in low-resource environments with scarce data (Li et al., 2023; Hu et al., 2023). To address the issue of data scarcity in real-world DocRE scenarios, a previous work (Popovic and Färber, 2022) reformulates the DocRE task as a few-shot learning problem, introducing several metric-based prototypical network models. Subsequent research (Meng et al., 2023) proposes a relation-aware prototypical method, constructing instance-level prototypes to better capture the semantic relationships in various contexts. Although these studies achieve some performance improvements over fully supervised models, the limitations of pre-trained language models prevent them from reaching optimal performance. This motivates us to explore ways to better leverage the potential of LLMs for FSDLRE, aiming to develop more effective models for the task.

Meta In-Context Learning. Without any fine-tuning, LLMs can exhibit strong performance across various downstream tasks by simply adding a few instances to the prompt, which is known as the In-Context Learning (Min et al., 2022b). However, due to the fundamental differences between information extraction and language modeling, LLMs' ICL perform poorly on few-shot information extraction tasks (Meng et al., 2023; Ma et al., 2023c). Previous research (Min et al., 2022a) propose meta-training LLMs on a large number of tasks using annotations from existing data, allowing the model to learn how to perform ICL. In the field of relation extraction, There is a study (Li et al., 2024) explores using table prompts and multi-task meta-learning to improve sentence-level relation extraction. However, these methods often lead to overfitting to specific tasks and relation types. Additionally, sentence-level prompt templates do not suitable for complex prompts constructed from annotations in documents. We propose for the first time the use of meta-learning to enhance the ICL capabilities of LLMs in FSDLRE. Additionally, we introduce the notion of prototypes into the prompting templates of LLMs, organizing and aggregating annotated information from support documents into category prototypes. By directing LLMs to identify similar prototypes, our approach alleviates the semantic conflict between the training and test corpus, enabling robust meta-training of LLMs in FSDLRE task.

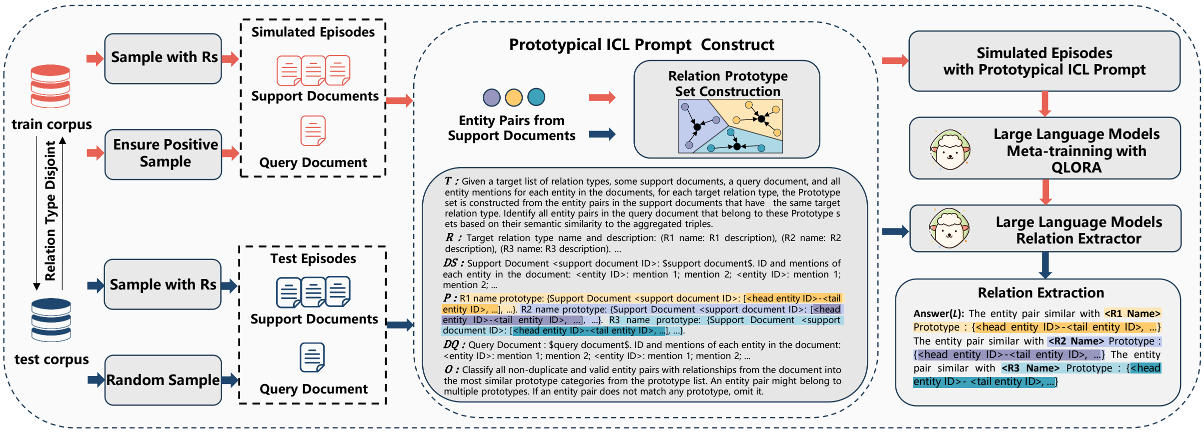
3 Methodology

Figure 2 illustrates the overall architecture of the prototype tuning. Initially, we describe the sampling process of simulated episodes used during training. Subsequently, we detail the conceptual framework for constructing prototype prompts. Finally, we discuss the meta-training and few-shot inference processes based on LLMs.

3.1 Problem Definition

In few-shot learning, each training/testing step is referred to as an episode (also known as a task). Each episode includes a support set comprising M documents and a query document. We conduct fine-tuning of LLMs under the classical meta-learning setup. The sets of relation types in the training corpus R_{train}, the validation corpus R_{dev}, and the test corpus R_{test} are pairwise disjoint. We adhere to the N-DOC setting (Popovic and Färber, 2022) for FSDLRE. As each document contains triplets with various relation types, the number of target categories and the number of samples for each relation type vary from episode to episode. Each support document contains a set T_S of all available triplets (e_h, r, e_t), where e_h represents the head entity, r denotes the relation type, and e_t indicates the tail entity. The entity e may appear multiple times in the document, referred to as the entity's mentions m. All relations contained in the support documents are denoted as $R_{episode}$, which are the target relation types that need to be identified for all entity pairs within the query document. If there is no relation between the entity pairs or if the relation is not part of $R_{episode}$, it is designated as NOTA (None-Of-The-Above) relation type.

For the FSDLRE task, entity mentions in both support and query documents are pre-annotated, and the goal is to predict the set of triplets T_Q in the query document based on the provided information. Specific details regarding the division of relation categories can be found in dataset paper (Popovic and Färber, 2022). The testing episodes for FSDLRE are sampled in two steps: First, from the set R_{test}, the relation type R_s that is currently least selected in the test corpus is chosen. If there


train corpus
Relation Type Disjoint
test corpus
Sample with Rs
Ensure Positive Sample
Sample with Rs
Random Sample
Simulated Episodes
Support Documents
Query Document
Test Episodes
Support Documents
Query Document
Prototypical ICL Prompt Construct
Entity Pairs from Support Documents
Relation Prototype Set Construction
Simulated Episodes with Prototypical ICL Prompt
Large Language Models Meta-trainning with QLORA
Large Language Models Relation Extractor
Relation Extraction

***T* :** Given a target list of relation types, some support documents, a query document, and all entity mentions for each entity in the documents, for each target relation type, the Prototype set is constructed from the entity pairs in the support documents that have   the same target relation type. Identify all entity pairs in the query document that belong to these Prototype s ets based on their semantic similarity to the aggregated triples.

***R* :** Target relation type name and description: (R1 name: R1 description), (R2 name: R2 description), (R3 name: R3 description). ...

***DS* :** Support Document <support document ID>: $support document$. ID and mentions of each entity in the document: <entity ID>: mention 1; mention 2; <entity ID>: mention 1; mention 2; ...

***P* :** R1 name prototype: {Support Document <support document ID>: [<head entity ID>-<tail entity ID>, ...], ...}. R2 name prototype: {Support Document <support document ID>: [<head entity ID>-<tail entity ID>, ...], ...}. R3 name prototype: {Support Document <support document ID>: [<head entity ID>-<tail entity ID>, ...], ...}.

***DQ* :** Query Document : $query document$. ID and mentions of each entity in the document: <entity ID>: mention 1; mention 2; <entity ID>: mention 1; mention 2; ...

***O* :** Classify all non-duplicate and valid entity pairs with relationships from the document into the most similar prototype categories from the prototype list. An entity pair might belong to multiple prototypes. If an entity pair does not match any prototype, omit it.

Answer(*L*): The entity pair similar with **<R1 Name>** Prototype : {<head entity ID>-<tail entity ID>, ...} The entity pair similar with **<R2 Name>** Prototype : {<head entity ID>-<tail entity ID>, ...} The entity pair similar with **<R3 Name>** Prototype : {<head entity ID>- <tail entity ID>, ...}

Figure 2: Overview of Prototype Tuning. The red arrows indicate the meta-training process and blue arrows indicate the testing process. Entities and prototypes of the same relation type are highlighted with the same background color. $Document$ denotes the specific content of the document.

are multiple such relation types, one is randomly selected. For this relation type, support documents are sampled, each containing at least one instance of R_s. Since the selected support documents may contain instances of other relation types from R_{test}, all relation types in the support documents that belong to R_{test} are added to $R_{episode}$. The query documents for test episodes are randomly drawn from the test corpus to authentically represent the NOTA distribution of the entire corpus. Our objective is to address the shortcomings of the existing ICL approach and develop a meta-learning approach that fully exploits the potential of LLMs in the FS-DLRE task to better adapt to new relation types with limited annotated data.

3.2 Simulated Episodes Sampling

We sample data from the training and development corpus with visible relation types to construct simulated episodes for our meta-training. The sampling of support documents for these episodes is identical to that described for test episodes in the problem definition section. The target relation category for each episode is determined by the support documents selected.

For query document sampling, we adopt an Ensure Positive strategy (Popovic and Färber, 2022) rather than random sampling. Specifically, we make sure that each training episode's query document contains at least one triplet with the target relation R_s. In scenarios where NOTA entity pairs are prevalent (Meng et al., 2023), this strategy allows the model to encounter more non-NOTA samples during training. Additionally, due to the high inference costs of LLMs, this configuration allows us to gather sufficient samples of each category with fewer development documents, thereby supporting macro-average evaluations and reducing computational costs.

3.3 Prototypical ICL Prompt

We organize the support documents and query documents in the episode into input prompts for the model. By incorporating the concept of prototypes in this process, the model obtains more abstract class representations, reducing the impact of semantic conflicts between meta-training and inference on model performance. The complete prompts for each part are presented in Figure 2.

For each episode, we first construct the task description template T for prototype tuning in FS-DLRE, which includes: (1) the task definition, (2) the prototype set definition, and (3) the initial description of the classification approach for the relationship categories of candidate entity pairs in the query document.

To constrain the range of prototypes and target relation types in each episode, we provide the names and specific descriptions of all target relation types within the episode in R. The target relation types are determined by all annotated instances contained in the support documents. All relationship types that do not belong to R or are of the None type are identified as NOTA. Additionally, to present the textual and entity information from each document clearly, we structure the content of M supporting documents into a prompt, each document has its own specific ID. We aggregate mentions of the same entity and represent them by a unique entity ID within each document. We then append this entity information to the text content of each document, forming the prompt $DS = [D_1, ..., D_M]$.

We design a specialized prompt template that aggregates pairs of entities with the same category of relationships from all supporting documents in an episode into a collection, termed as the prototype set. This template aims to create more abstract relation category representation, thereby enhancing the model's generalization capabilities for unseen categories. Specifically, all prototype sets are placed in a dictionary $P = \{R_1 : \{\}, ..., R_n : \{\}\}$ within the prompts, where the key of the dictionary is the name of the relationship type. Within each target relationship category prototype, we first specify the ID of supporting document from which the entity pair originates, then use the entity IDs from the entity information of each support document to represent the corresponding head and tail entities. Notably, due to the excessive number of NOTA type candidate entity pairs (Popovic and Färber, 2022), We do not set up a separate explicit NOTA prototype in the Prompt P, considering the constraints of prompt length and complexity. Instead, we implicitly represent the NOTA prototype by classifying all entity pairs with target relations into the corresponding prototype.

Finally, we organize the relevant information from the query documents within the prompt and construct instructions to guide the model's output. Specifically, we format the text and entity information from the query documents similarly to a support document and place it into the D_Q. And we use specail output instruction O to encourage the model to focus more on the similarity between the entity pairs in the query document and the various relationship prototypes. Notably, although our prototype ICL prompt reorganizes the input for FSDLRE, it does not significantly increase the prompt length or computational overhead compared to the ICL and MetaICL prompts. Detailed information about these prompts can be found in Appendices A and B.

3.4 Meta Training

We utilize the triplet annotations in the query document to construct the generated gold label. In the label L, each relation prototype is represented by a set, and entity pairs are classified into the set of the corresponding prototype based on their relationship type. The head and tail entities are represented by their entity IDs from the Document information prompt D_Q of the query document, while each relationship prototype collection is defined by its corresponding relationship name. Same to the prototype construction in P, the NOTA prototype in L is also implicitly represented, if a candidate entity pair from the query document does not appear in the model's response, it is automatically classified as belonging to the NOTA relationship category.

We employ Prototypical ICL Prompts I and labels L constructed from simulated episodes for Meta-training LLMs. Specifically, the input $I = [T, R, DS, P, DQ, O]$ consists of task instructions T, a description of the target relation R, supporting document text and related entity information DS, various prototype sets P, query document text and related entity information DQ, and the output command O concatenated together. The labels for the generated model are constructed from all triples T_Q in the query document. After inputting I into the model, we obtain the model's predicted probabilities y for tokens in the vocabulary and calculate the negative log likelihood loss with the labels L to update the model's parameters. During the meta-training process, we employ a parameter-efficient fine-tuning method QLORA (Dettmers et al., 2023) to fine-tune the LLMs.

3.5 Few-Shot Inference

After prototype tuning the LLMs with training and development simulated episodes constructed from corpus containing visible relationship categories, we evaluate the model's performance in a true few-shot setting on a set of episodes built from the test corpus with unseen relationship categories. Specifically, For an N-DOC FSDLRE episode that includes unseen relationship categories, the model is provided with N support documents $S = \{s_1, s_2, ..., s_n\}$ annotated with Triple sets $Y = \{y_1, y_2, ..., y_n\}$, a query document q, and the target relationship types $R_{episode}$. We organize the information from the test episode into the same input format I as used during meta-training. The meta-trained LLMs assesses the similarity between the implied candidate entities in the query document and the various relationship prototypes in P, assigning them to the most similar relationship prototype. If a candidate entity pair does not resemble any existing relationship prototype it is categorized into the implicit NOTA prototype. All responses from the meta-trained LLMs are presented in the format of label L. Based on the classification of the candidate entity pairs in the response, we categorize the entity pairs with specific head and tail entities into the relationship category corresponding to the respective prototype.

4 Experiments

4.1 Dataset and Metric

We conduct experiments on two publicly available FSDLRE benchmark, FREDo (Popovic and Färber, 2022) and ReFREDo (Meng et al., 2023). These two benchmarks provide pre-sampled fixed test episodes for each setting to ensure fairness in model performance comparisons.

FREDo. This benchmark includes two tasks: In-Domain and Cross-Domain, each with 1-DOC and 3-DOC subtasks, aiming to evaluate the model's scalability in different scenarios. For the In-Domain task, the training and testing documents both come from DocRED (Yao et al., 2019). Specifically, the relationship types in DocRED are divided into three non-overlapping subsets: training (62 types), development (16 types), and in-domain testing (18 types). FREDo uses DocRED's training set as the training and development document corpus, with its development set serving as the document corpus for in-domain testing. For the In-Domain task, document-based trained models are evaluated on 15K episodes derived from DocRED. In the Cross-Domain task, the training documents come from DocRED, while the testing documents come from SciERC (Luan et al., 2018), whose document topics, relationship types, and text styles significantly differ from those of the training documents. FREDo uses the entire SciERC dataset as the corpus for cross-domain testing. Models initially trained on DocRED samples are evaluated on 3K episodes from SciERC documents.

ReFREDo. This benchmark is a revised version of FREDo, replacing the training, development, and in-domain testing corpus with documents from Re-DocRED (Tan et al., 2022b), which extends the relationship facts in DocRED to 119,991. This expansion addresses the issue of missing labels and provides more comprehensive annotations. In ReFREDo, the division of relationship types for each dataset remains the same as in FREDo. The in-domain test sampled 15K episodes, while the cross-domain test episodes, like in FREDo, are constructed based on the entire SciERC dataset.

Metric. We use Macro-F1 as the evaluation metric for all models to thoroughly assess their generalization ability to new relationship categories. In the generated responses, we extract results based on a specified format of L. A triplet is considered correct only if the head entity, tail entity, and relationship type all match the labels exactly. This setup simulates the real-world scenario of developing models that can quickly adapt to new categories using a small number of new category annotations.

4.2 Baselines

We compare the Prototype-Tuned LLM with the LLMs that have not undergone training and those trained using the standard meta-training under the ICL paradigm. Furthermore, to validate the advanced nature of our training approach, we also evaluate against the current state-of-the-art methods for the FSDLRE task.

Prototypical Network-Based Baselines. These methods vectorize entity pairs in support documents and aggregate them into relation prototypes based on the relationships between the entities. The relationship type of candidate entity pairs in query documents is then predicted by computing their similarity to these aggregated prototypes. DL-Base encode documents using the untuned BERT-base model (Devlin et al., 2019). DL-MNAV (Popovic and Färber, 2022) extends sentence-level methodologies (Sabo et al., 2021) to document-level for few-shot relation extraction. RAPL (Meng et al., 2023) redefines relational prototypes at the instance level and introduces a relation-weighted contrastive learning approach to improve the precision of these prototypes. It also develops a task-specific strategy for generating NOTA prototypes, enhancing the ability to capture NOTA semantics in each task.

LLM-Based Baselines. We design a common prompt that adapts the conventional ICL approach to the FSDLRE task. Following the textual and entity information in each document, we present the supporting document triples as example responses, formatted as <*head entity ID*>-<*relation name*>-<*tail entity ID*>. The response labels also employ the same setup. Based on this common prompt, we construct two baselines. The first is an untrained LLMs, used to evaluate the intrinsic performance of LLMs under the classic ICL approach on the FS-DLRE task, including models like ChatGPT-3.5[1] and Llama-3-8B-Instruct[2]. The second is a Llama3 baseline that uses the MetaICL approach, further enhancing the ICL through meta-learning.

[1] openai.com/api. The version is gpt-3.5-turbo-0125.

[2] llama.meta.com.

Model/Macro F1(%)	**FREDo**				**ReFREDo**			
	In-Domain		**Cross-Domain**		**In-Domain**		**Cross-Domain**	
	1-Doc F_1	3-Doc F_1	1-Doc F_1	3-Doc F_1	1-Doc F_1	3-Doc F_1	1-Doc F_1	3-Doc F_1
Prototypical Network								
DL-Base	0.60	0.89	1.76	1.98	1.38	1.84	1.76	1.98
DL-MNAV	7.05 ± 0.18	8.42 ± 0.64	0.84 ± 0.16	0.48 ± 0.21	12.97 ± 0.88	12.43 ± 0.36	1.12 ± 0.38	2.28 ± 0.19
DL-MNAV$_{SIE}$	7.06 ± 0.15	6.77 ± 0.21	1.77 ± 0.60	2.51 ± 0.66	13.37 ± 0.98	12.00 ± 0.80	1.39 ± 0.74	2.92 ± 0.41
DL-MNAV$_{SIE+SBN}$	1.71 ± 0.24	2.79 ± 0.24	2.85 ± 0.12	3.72 ± 0.14	4.59 ± 0.30	5.43 ± 0.24	2.84 ± 0.24	3.86 ± 0.27
RAPL	8.75 ± 0.80	10.67 ± 0.77	3.33 ± 0.50	5.35 ± 0.72	15.20 ± 0.82	16.35 ± 0.60	3.51 ± 0.79	5.48 ± 0.63
LLMs								
ChatGPT$_{ICL}$	2.25	2.95	5.22	5.83	2.86	5.39	5.22	5.83
Llama-3-8B-Instruct$_{ICL}$	2.04	2.27	5.05	5.53	3.07	2.36	5.05	5.53
Llama-3-8B-Instruct$_{MetaICL}$	13.81 ± 0.32	14.67 ± 0.44	4.50 ± 0.28	5.46 ± 0.53	21.14 ± 0.97	23.89 ± 0.83	5.79 ± 0.48	5.49 ± 0.67
Llama-3-8B-Instruct$_{PT}$	**14.98 ± 0.70**	**16.83 ± 0.61**	**7.42 ± 0.55**	**7.54 ± 0.64**	**31.54 ± 0.99**	**33.12 ± 1.05**	**8.10 ± 0.85**	**8.69 ± 0.72**

Table 1: Results on FREDo and ReFREDo benchmarks. The scores of existing methods are from previous paper (Meng et al., 2023). The values of the best-performing approach's metrics are highlighted in bold. For all meta-training approaches, we report the mean and standard deviation of the Macro F1 score across five runs with different random seeds.

Model/Macro F1(%)	**In-Domain**		**Cross-Domain**	
	1-Doc	3-Doc	1-Doc	3-Doc
ICL	3.07	2.36	5.05	5.53
MetaICL	21.14	23.89	5.79	5.49
Prototype Tuning	**31.54**	**33.12**	**8.10**	**8.69**
Train w/o RS				
MetaICL	17.58	16.88	6.32	6.88
Prototype Tuning	22.65	23.08	6.34	6.55
Train and Test w/o RS				
ICL	1.63	2.13	3.93	4.07
MetaICL	11.44	12.11	3.18	3.52
Prototype Tuning	14.65	12.23	4.83	4.70

Table 2: Results of Relation Semantics Ablation on ReFREDo. In the prompts and labels, relation names and descriptions are replaced with relational identifiers. The table separately shows the performance of all approaches that ablate relation semantics only during training and during both training and testing.

4.3 Implement Details

In the meta-training of LLMs, we use a learning rate of 2e-5, a batch size of 2, and a maximum sequence length of 4096 tokens. The first 6% of steps are linearly warmed up, followed by a linear decay to zero. Each meta-training session samples 50K simulated training episodes and 1K development episodes, with early stopping based on macro F1 on the development set. For all meta-training approaches, we report the mean and standard deviation of the macro F1 score across five runs with different random seeds. We utilize Unsloth[3] to reduce memory usage without impacting training or inference, allowing all experiments to be conducted on a single RTX 4090 GPU. In the parameter-efficient fine-tuning technique QLoRA, we set the rank to 64 and the merging factor α to 16.

4.4 Main Results

The main results under various settings on FREDo and ReFREDo are shown in Table 1. We can observe that: (1) With the ICL approach for the FSDLRE task, ChatGPT and Llama-3 do not surpass the performance of methods based on traditional prototypical networks. This clearly indicates an inherent compatibility issue between LLMs and the FSDLRE task, highlighting the necessity of using meta-learning to enhance the ICL capabilities of LLMs. (2) With the MetaICL approach, LLM demonstrate significant performance improvements across all settings in the FSDLRE task compared to using the ICL approach, with an average increase of **+5.61 F1**. This suggests that meta-training helps the model better adapt to task formats and acquire general knowledge, enhancing its capability to understand complex prompts and extract relations. (3) With Prototye Tuning approach, the model show further performance improvements, with an average increase of **+4.18 F1** on top of the MetaICL, and a total increase of **+9.79 F1** compared to the ICL approach, achieving the best performance across all settings. This indicates that for LLMs, introducing prototype concepts into Meta Learning and shifting the classification approach to finding similar prototypes significantly helps the model adapt to unseen relation categories, effectively alleviating the conflict between training and testing relation semantics. (4) In cross-domain testing, the significant stylistic differences between test

[3] https://github.com/unslothai/unsloth

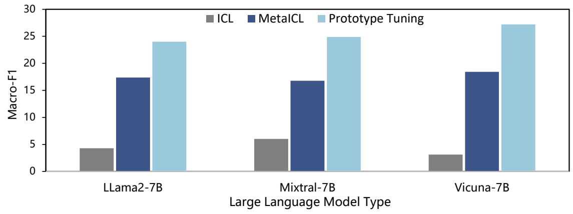


Figure 3: Results of the performance evaluation of other LLMs using three approaches under the 3-DOC configuration of ReFREDo's In-Domain Test.

and meta-training texts limited MetaICL's ability to improve model performance. Prototype Tuning, by using abstract category prototypes in prompts, better harnesses the potential of LLM to adapt to new categories in cross-domain documents. (5) Across both in-domain and cross-domain scenarios, models perform better on Re-FREDo, with more complete triple annotations, than on FREDo. This suggests that label accuracy in simulated training significantly impacts performance, and incomplete annotations in testing can affect in-domain results.

4.5 Analysis and Discussion

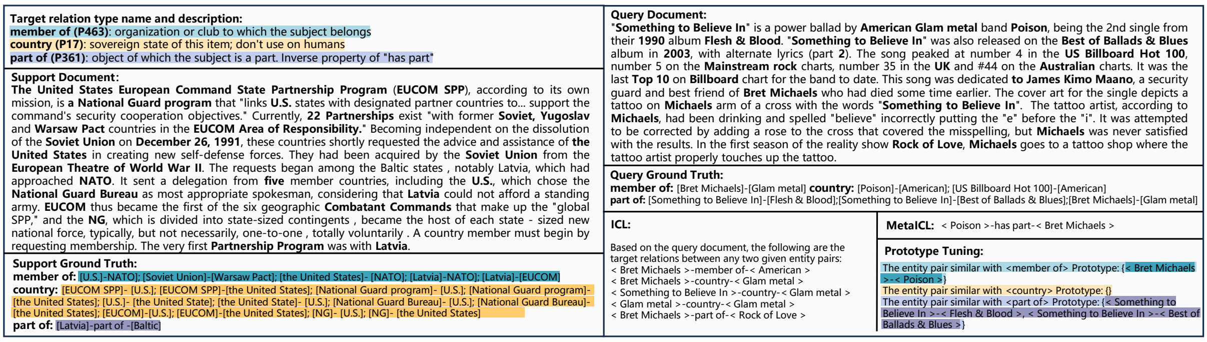
Relation Semantics Ablation. Although the training and testing corpus are divided according to relationship categories, it is difficult to completely avoid potential semantic associations and conflicts between the two datasets. To explore the impact of the given relationship semantics in prompts on model performance, we conduct ablation experiments on ReFREDo involving label names and descriptions based on Llama-3-8B-Instruct. We replace all relationship names in the prompts and labels with non-specific relationship markers $[R_1, R_2, ..., R_k]$ and remove all relationship descriptions. As shown in Table 2, we can observe that: (1) When relationship semantics are ablated during training, both meta-learning approaches generally show performance declines. However, the MetaICL approach shows improved cross-domain performance, suggesting that semantic conflicts negatively impact relationship extraction more in scenarios with greater training-testing style differences. The Prototype Tuning approach shows that it handles these conflicts between training and testing corpora better. (2) When relationship semantics are ablated during both training and inference, all three approaches experience a significant performance drop. This highlights the importance of relationship semantics within an episode for the model's understanding of the target relationships and accurate relationship extraction.

Model/Macro-F1 (%)	**In-Domain**		**Cross-Domain**	
	1-Doc	3-Doc	1-Doc	3-Doc
Ensure Positive	31.54	33.12	8.10	8.69
Random Sample	30.08	31.66	7.83	8.18

Table 3: Results of Ensure Positive Ablation on ReFREDo. The Random Sample strategy refers to randomly sampling query document when constructing simulated episodes.

Robustness of Prototype Tuning. As shown in Figure 3, we compare the performance of several LLMs using ICL, MetaICL, and Prototype Tuning frameworks on the 3-DOC setting of ReFREDo. The models used are Mistral-7B-Instruct (Jiang et al., 2023), Vicuna-7B-v1.5 (Chiang et al., 2023), and Llama-2-7B-Chat. Experimental results show that both MetaICL and Prototype Tuning have a considerable positive impact on model performance across different LLMs. This finding support a broadly applicable conclusion across all tested LLMs: compared to solely using the ICL approach, Meta Learning significantly enhances LLMs' adaptability to complex document prompts and structured information extraction tasks.Introducing the concept of prototypes further enhances meta-training by reducing the impact of semantic conflicts between the training corpus and testing corpus on model performance. These experiments underscore the robustness of our proposed approaches.

The Role of Ensure Positive. To validate the effectiveness of the Ensure Positive sampling strategy implemented during the simulated episodes in our training process, we conduct a comparative experiment using the 3-Doc scenario of ReFREDo. During the training phase, we replaced the sampling strategy in simulated episodes from Ensure Positive to random sampling, akin to our approach in testing episodes. As demonstrated in table 3, employing the Ensure Positive strategy not only reduced training costs but also significantly enhanced model performance due to an increase in effective scenarios. In contrast, random sampling often led to the generation of numerous instances with no answer (i.e., an excess of NOTA candidate entity pairs), thereby weakening the model's predictive power for new categories. This experiment robustly supports the rationale and benefits of our chosen sampling strategy in simulated episodes.



Target relation type name and description:
member of (P463): organization or club to which the subject belongs
country (P17): sovereign state of this item; don't use on humans
part of (P361): object of which the subject is a part. Inverse property of "has part"

Support Document:
The United States European Command State Partnership Program (**EUCOM SPP**), according to its own mission, is **a National Guard program** that "links **U.S.** states with designated partner countries to... support the command's security cooperation objectives." Currently, **22 Partnerships** exist "with former **Soviet, Yugoslav** and **Warsaw Pact** countries in the **EUCOM Area of Responsibility.**" Becoming independent on the dissolution of the **Soviet Union** on **December 26, 1991**, these countries shortly requested the advice and assistance of **the United States** in creating new self-defense forces. They had been acquired by the **Soviet Union** from the **European Theatre of World War II**. The requests began among the Baltic states , notably Latvia, which had approached **NATO**. It sent a delegation from **five** member countries, including the **U.S.**, which chose the **National Guard Bureau** as most appropriate spokesman, considering that **Latvia** could not afford a standing army. **EUCOM** thus became the first of the six geographic **Combatant Commands** that make up the "global SPP," and the **NG**, which is divided into state-sized contingents , became the host of each state - sized new national force, typically, but not necessarily, one-to-one , totally voluntarily . A country member must begin by requesting membership. The very first **Partnership Program** was with **Latvia**.

Support Ground Truth:
member of: [U.S.]-NATO]; [Soviet Union]-[Warsaw Pact]; [the United States]- [NATO]; [Latvia]-NATO]; [Latvia]-[EUCOM]
country: [EUCOM SPP]- [U.S.]; [EUCOM SPP]-[the United States]; [National Guard program]- [U.S.]; [National Guard program]-[the United States]; [U.S.]- [the United State]; [the United State]- [U.S.]; [National Guard Bureau]- [U.S.]; [National Guard Bureau]-[the United States]; [EUCOM]-[U.S.]; [EUCOM]-[the United States]; [NG]- [U.S.]; [NG]- [the United States]
part of: [Latvia]-part of -[Baltic]

Query Document:
"**Something to Believe In**" is a power ballad by **American Glam metal** band **Poison**, being the 2nd single from their **1990** album **Flesh & Blood**. "**Something to Believe In**" was also released on the **Best of Ballads & Blues** album in **2003**, with alternate lyrics (part **2**). The song peaked at number 4 in the **US Billboard Hot 100**, number 5 on the **Mainstream rock** charts, number 35 in the **UK** and #44 on the **Australian** charts. It was the last **Top 10** on **Billboard** chart for the band to date. This song was dedicated **to James Kimo Maano**, a security guard and best friend of **Bret Michaels** who had died some time earlier. The cover art for the single depicts a tattoo on **Michaels** arm of a cross with the words "**Something to Believe In**". The tattoo artist, according to **Michaels**, had been drinking and spelled "believe" incorrectly putting the "e" before the "i". It was attempted to be corrected by adding a rose to the cross that covered the misspelling, but **Michaels** was never satisfied with the results. In the first season of the reality show **Rock of Love**, **Michaels** goes to a tattoo shop where the tattoo artist properly touches up the tattoo.

Query Ground Truth:
member of: [Bret Michaels]-[Glam metal] **country:** [Poison]-[American]; [US Billboard Hot 100]-[American]
part of: [Something to Believe In]-[Flesh & Blood];[Something to Believe In]-[Best of Ballads & Blues];[Bret Michaels]-[Glam metal]

ICL:

Based on the query document, the following are the target relations between any two given entity pairs:
< Bret Michaels >-member of-< American >
< Bret Michaels >-country-< Glam metal >
< Something to Believe In >-country-< Glam metal >
< Glam metal >-country-< Glam metal >
< Bret Michaels >-part of-< Rock of Love >

MetaICL: < Poison >-has part-< Bret Michaels >

Prototype Tuning:
The entity pair similar with <member of> Prototype: {< Bret Michaels >-< Poison >}
The entity pair similar with <country> Prototype: {}
The entity pair similar with <part of> Prototype: {< Something to Believe In >-< Flesh & Blood >, < Something to Believe In >-< Best of Ballads & Blues >}

Figure 4: Case study of an in-domain 1-Doc episode in ReFREDo. The mentions are highlighted in bold. The specific inputs for the three approaches can be found in the appendix.

Case Study. As shown in Figure 4, we select a representative in-domain 1-doc test episode from ReFREDo for a case study, showcasing the responses of LLaMa3-8B-Instruct under three different approaches. This helps to intuitively illustrate the strengths and limitations of each approach. The specific input formats for each approach can be found in the appendix, and for readability, entity IDs in the model responses are replaced with one of their mentions. We can observe that: (1) With the ICL approach, the model struggles to accurately understand the specific definitions of each relation based only on relation descriptions and single document triplet annotation. This results in many incorrect triplet predictions. (2) With the MetaICL approach, meta-training helps the model better distinguish between different relation categories. However, the model can still be influenced by the semantic differences between the training and test data. For example, when "Poison" is the head entity and "Bret Michaels" is the tail entity, this pair should be classified as NOTA In this episode and excluded from the response. However, because the training set includes the relation "has part", the model mistakenly associates the "part of" relation in the query with "has part" from the training set, leading to incorrect predictions. (3) Prototype Tuning introduces the concept of prototypes into the meta-learning process. By constructing category prototypes from relation instances in the support documents, the model can more accurately extract relations between entity pairs in the query document. While Prototype Tuning significantly improves the performance of FSDLRE tasks using LLMs, the overall performance of these models still falls short of supervised setups. This indicates that further optimization of input prompt construction and meta-learning approaches for LLMs remains a crucial area for future research.

5 Conclusion

In this work, we introduce a new few-shot learning approach named Prototype Tuning. This approach incorporates the concept of prototypes into the meta-training process of large language models (LLMs), enabling model to better extract relations in the context of documents. Extensive experiments based on a broad range of LLMs demonstrate that Prototype Tuning consistently outperforms In Context Learning (ICL) approaches without meta-training and standard meta-trained ICL approaches. Furthermore, it achieves significant improvements compared to state-of-the-art task-specific models. We also conduct extensive analytical experiments on the Prototype Tuning approach to validate its strengths and weaknesses, inspiring future research to explore more effective few-shot document-level relation extraction approaches.

Limitations

Although we use an optimized training framework to reduce the training costs, MetaICL and Prototype Tuning inevitably introduce additional computational overhead due to the construction of simulated episodes and meta-training, compared to ICL approaches that do not require instruction tuning. In the future, we will explore using more optimized Parameter efficient fine-tuning methods (Dettmers et al., 2023) to reduce the training costs of our approach. While Prototype Tuning shows promising improvements in cross-domain settings, it is still far from practical application. We will continue to investigate techniques such as data augmentation(Sun et al., 2024) to better unlock the potential of LLMs in FSDLRE cross-domain tasks.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (No. 62276043, 62302076), the National Key Research and Development Program of China (No. 2022YFC3301801).

References

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 10088–10115.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Zhichao Duan, Xiuxing Li, Zhenyu Li, Zhuo Wang, and Jianyong Wang. 2022. Not just plain text! fuel document-level relation extraction with explicit syntax refinement and subsentence modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1941–1951.

Xuming Hu, Junzhe Chen, Shiao Meng, Lijie Wen, and Philip S Yu. 2023. Selflre: Self-refining representation learning for low-resource relation extraction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2364–2368.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Guozheng Li, Peng Wang, Jiajun Liu, Yikai Guo, Ke Ji, Ziyu Shang, and Zijie Xu. 2024. Meta in-context learning makes large language models better zero and few-shot relation extractors. *arXiv preprint arXiv:2404.17807*.

Shu'ang Li, Xuming Hu, Li Lin, Aiwei Liu, Lijie Wen, and Philip S Yu. 2023. A multi-level supervised contrastive learning framework for low-resource natural language inference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1771–1783.

Chonggang Lu, Richong Zhang, Kai Sun, Jaein Kim, Cunwang Zhang, and Yongyi Mao. 2023. Anaphor assisted document-level relation extraction. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232.

Xilai Ma, Jing Li, and Min Zhang. 2023a. Chain of thought with explicit evidence reasoning for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2334–2352.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023b. Dreeam: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1963–1975.

Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. 2023c. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! *arXiv preprint arXiv:2303.08559*.

Shiao Meng, Xuming Hu, Aiwei Liu, Fukun Ma, Yawen Yang, Lijie Wen, et al. 2023. Rapl: A relation-aware prototype learning approach for few-shot document-level relation extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5208–5226.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022a. Metaicl: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.

Nicholas Popovic and Michael Färber. 2022. Few-shot document-level relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5733–5746.

Ofer Sabo, Yanai Elazar, Yoav Goldberg, and Ido Dagan. 2021. Revisiting few-shot relation classification: Evaluation data and classification schemes. *Transactions of the Association for Computational Linguistics*, 9:691–706.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Qi Sun, Kun Huang, Xiaocui Yang, Rong Tong, Kun Zhang, and Soujanya Poria. 2024. Consistency guided knowledge retrieval and denoising in llms for zero-shot document-level relation triplet extraction. In *Proceedings of the ACM on Web Conference 2024*, pages 4407–4416.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting docred-addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. *arXiv preprint arXiv:2305.05003*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, and Jiawei Han. 2022. Sais: Supervising and augmenting intermediate steps for document-level relation extraction. In *2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, pages 2395–2409. Association for Computational Linguistics (ACL).

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 257–268.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021a. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, pages 14149–14157.

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021b. Discriminative reasoning for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1653–1663.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640.

A ICL and MetaICL Prompt

The input prompts for our model when using the ICL and MetaICL approaches are presented as follows:

Given a target list of relation types, some support Documents, a query document, and all entity mentions for each entity in the query documents, please identify the target relations between any two given entity pairs in the query document. Do not present the results of the support documents.

Target relation type name and description:

<R1 Name>: R1 Description;

<R2 Name>: R2 Description;

......

Support Documents: $Support Document$

ID and mentions of each entity in the document:

<1>: Mention 1 of Entity 1; Mention 2 of Entity 1; ......

<2>: Mention 1 of Entity 2; Mention 2 of Entity 2; ......

......

All non-duplicate valid <subject entity ID>-<target relation type>-<object entity ID>triples in the document (output format: <entity ID>-<relation type name>-<entity ID>, e.g., <1>-<relation des>-<2>; one triple per line, If there are no entities with existing relationships, return None):

<head entity ID>-relation name-<tail entity ID>

<head entity ID>-relation name-<tail entity ID>

......

Query Document: $Query Document$

ID and mentions of each entity in the document:

<1>: Mention 1 of Entity 1; Mention 2 of Entity 1; ......

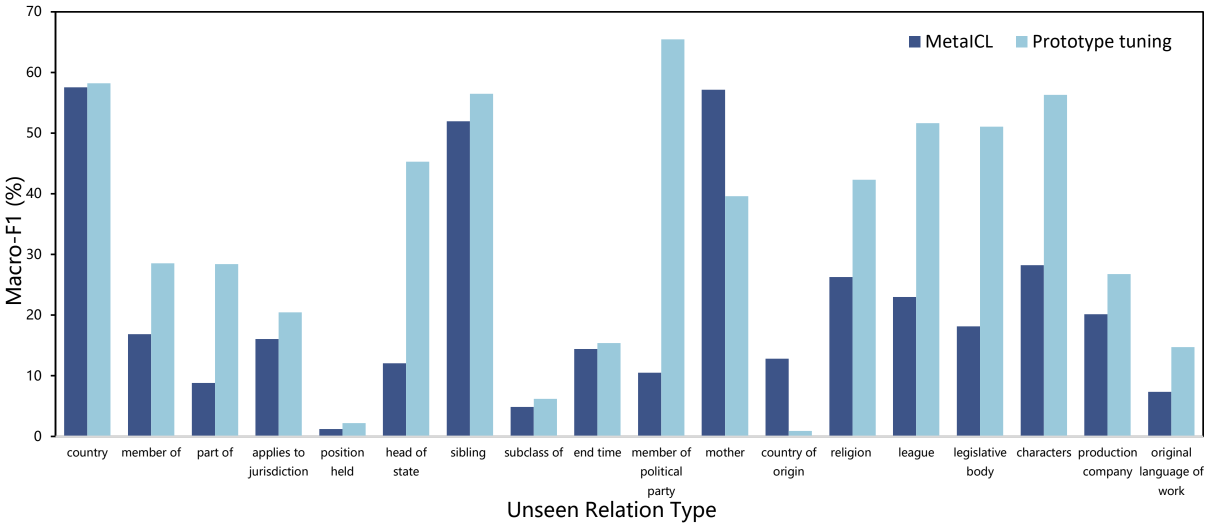


Figure 5: The performance of MetaICL and Prototype Tuning across various unseen types.

<2>: Mention 1 of Entity 2; Mention 2 of Entity 2; ......

......

All non-duplicate valid <subject entity ID>-<target relation type>-<object entity ID>triples in the document (output format: <entity ID>-<relation type name>-<entity ID>, e.g., <1>-<relation des>-<2>; one triple per line, If there are no entities with existing relationships, return None):

The format of the label L in the MetaICL approach during meta-training is as follows:

<head entity ID>-<relation name>-<tail entity ID>;

<head entity ID>-<relation name>-<tail entity ID>;

......

B Prototypical ICL Prompt

The input prompts for our model when using the Prototype Tuning approach are presented as follows:

Given a target list of relation types, some support documents, a query document, and all entity mentions for each entity in the documents, for each target relation type, the corresponding Prototype set is constructed from the entity pairs in the support documents that have the same target relation type. Identify all entity pairs in the query document that belong to these Prototype sets based on their semantic similarity to the aggregated triples.

Target relation type name and description:

<R1 Name>: R1 Description;

<R2 Name>: R2 Description;

......

Support Documents: $Support Document$

ID and mentions of each entity in the document:

<1>: Mention 1 of Entity 1; Mention 2 of Entity 1; ......

<2>: Mention 1 of Entity 2; Mention 2 of Entity 2; ......

......

R1 name prototype: {

Support Document <support document id>: [<head entity id>- <tail entity id>, ...],

Support Document <support document id>: [<head entity id>- <tail entity id>, ...]

....}

R2 name prototype: {

Support Document <support document id>: [<head entity id>- <tail entity id>, ...],

Support Document <support document id>: [<head entity id>- <tail entity id>, ...]

......}

......

Query Document: $Query Document$

ID and mentions of each entity in the document:

<1>: Mention 1 of Entity 1; Mention 2 of Entity 1; ......

<2>: Mention 1 of Entity 2; Mention 2 of Entity 2; ......

......

Classify all non-duplicate and valid entity pairs with relationships from the document into the most similar prototype types from the prototype list. An entity pair might belong to multiple prototypes. If an entity pair does not match any prototype, omit it.

The format of the label L in the Prototype Tuning approach during meta-training is as follows:

The entity pairs similar with <R1 name>Prototype: {<head entity id>-<tail entity id>, ... }

The entity pairs similar with <R2 name>Prototype: {<head entity id>-<tail entity id>, ... },

......

Benchmark	Task	N	K(micro)	K(macro)
FREDo	In-Domain 1-DOC	2.18	2.36	2.24
	In-Domain 2-DOC	3.47	4.30	4.31
ReFREDo	In-Domain 1-DOC	3.50	3.50	3.11
	In-Domain 2-DOC	5.67	6.50	5.73
FREDo and ReFREDo	Cross-Domain 1-DOC	4.26	2.73	2.40
	Cross-Domain 2-DOC	6.08	5.55	5.27

Table 4: Average values for N and K are reported across test episodes in FREDo and ReFREDo. K (micro) represents the average from all episodes, while K (macro) refers to the weighted average of the mean K values for each relation type.

C Model Performance on Unseen Relation Types

We investigate the performance of MetaICL and Prototype Tuning across various unseen types in the ReFREDo In-domain 3-DOC scenario. As shown in Figure 5, our findings indicate that Prototype Tuning generally surpasses MetaICL in most types. This advantage arises from MetaICL's reliance on general features learned during pre-training and meta-training when dealing with unseen types, whereas Prototype Tuning enhances classification accuracy by focusing more on the similarity between candidate entity pairs in the query document and those in the support document. Notably, since the high-performing types make up a significant portion of the test episodes, the overall Micro F1 score achieved using Prototype Tuning reaches 44.21. However, we prioritize the Macro-F1 score over the Micro F1 score, as the latter does not sufficiently assess the adaptability of LLMs to a diverse range of new types, especially when only a limited amount of labeled data is available.

D Supplementary Description of the Dataset

To align the FSDLRE task with the conventional N-way K-shot format typical of few-shot tasks, we outline the distribution of N and K across the test sets in Table 4

In Tables 5 to 9, we list the types of relations for training, development, in-domain testing, and cross-domain testing document corpora in FREDo and ReFREDo. We present the name and description of each relation type.

Wikidata ID	Name	Description
P6	head of government	head of the executive power of this town, city, municipality, state, country, or other governmental body
P19	place of birth	most specific known (e.g. city instead of country, or hospital instead of city) birth location of a person, animal or fictional character
P20	place of death	most specific known (e.g. city instead of country, or hospital instead of city) death location of a person, animal or fictional character
P22	father	male parent of the subject
P26	spouse	the subject has the object as their spouse (husband, wife, partner, etc.)
P30	continent	continent of which the subject is a part
P31	instance of	that class of which this subject is a particular example and member. (Subject typically an individual member with Proper Name label.)
P36	capital	primary city of a country, state or other type of administrative territorial entity
P37	official language	language designated as official by this item
P40	child	subject has the object in their family as their offspring son or daughter (independently of their age)
P54	member of sports team	sports teams or clubs that the subject currently represents or formerly represented
P58	screenwriter	author(s) of the screenplay or script for this work
P69	educated at	educational institution attended by the subject
P108	employer	person or organization for which the subject works or worked
P123	publisher	organization or person responsible for publishing books, periodicals, games or software
P127	owned by	owner of the subject
P131	located in the administrative territorial entity	the item is located on the territory of the following administrative entity
P155	follows	immediately prior item in some series of which the subject is part
P156	followed by	immediately following item in some series of which the subject is part
P159	headquarters location	specific location where an organization's headquarters is or has been situated
P161	cast member	actor performing live for a camera or audience
P162	producer	producer(s) of this film or music work (film: not executive producers, associate producers, etc.)
P166	award received	award or recognition received by a person, organisation or creative work
P170	creator	maker of a creative work or other object (where no more specific property exists)
P171	parent taxon	closest parent taxon of the taxon in question
P172	ethnic group	subject's ethnicity (consensus is that a VERY high standard of proof is needed for this field to be used. In general this means 1) the subject claims it him/herself, or 2) it is widely agreed on by scholars, or 3) is fictional and portrayed as such).
P175	performer	performer involved in the performance or the recording of a work
P178	developer	organisation or person that developed this item

Table 5: Relation types and description of training document corpus in FREDo and ReFREDo (continued on next page).

Wikidata ID	Name	Description
P190	sister city	twin towns, sister cities, twinned municipalities and other localities that have a partnership or cooperative agreement, either legally or informally acknowledged by their governments
P205	basin country	country that have drainage to/from or border the body of water
P206	located in or next to body of water	sea, lake or river
P241	military branch	branch to which this military unit, award, office, or person belongs
P264	record label	brand and trademark associated with the marketing of subject music recordings and music videos
P276	location	location of the item, physical object or event is within
P400	platform	platform for which a work has been developed or released / specific platform version of a software developed
P403	mouth of the watercourse	the body of water to which the watercourse drains
P449	original network	network(s) the radio or television show was originally aired on, including
P527	has part	part of this subject. Inverse property of "part of"
P551	residence	the place where the person is, or has been, resident
P569	date of birth	date on which the subject was born
P570	date of death	date on which the subject died
P576	dissolved, abolished or demolished	date or point in time on which an organisation was dissolved/disappeared or a building demolished
P577	publication date	date or point in time a work is first published or released
P580	start time	indicates the time an item begins to exist or a statement starts being valid
P585	point in time	time and date something took place, existed or a statement was true
P607	conflict	battles, wars or other military engagements in which the person or item participated
P676	lyrics by	author of song lyrics
P706	located on terrain feature	located on the specified landform
P710	participant	person, group of people or organization (object) that actively takes/took part in the event (subject)
P737	influenced by	this person, idea, etc. is informed by that other person, idea, etc.
P740	location of formation	location where a group or organization was formed
P749	parent organization	parent organization of an organisation, opposite of subsidiaries
P800	notable work	notable scientific, artistic or literary work, or other work of significance among subject's works
P807	separated from	subject was founded or started by separating from identified object
P840	narrative location	the narrative of the work is set in this location
P937	work location	location where persons were active
P1198	unemployment rate	portion of a workforce population that is not employed
P1336	territory claimed by	administrative divisions that claim control of a given area
P1344	participant of	event a person or an organization was a participant in, inverse of "participant"
P1365	replaces	person or item replaced
P1376	capital of	country, state, department, canton or other administrative division of which the municipality is the governmental seat
P1412	languages spoken, written or signed	language(s) that a person speaks or writes, including the native language(s)

Table 6: Relation types and description of training document corpus in FREDo and ReFREDo (continued).

Wikidata ID	Name	Description
P27	country of citizenship	the object is a country that recognizes the subject as its citizen
P150	contains administrative territorial entity	(list of) direct subdivisions of an administrative territorial entity
P571	inception	date or point in time when the organization/subject was founded/created
P50	author	main creator(s) of a written work (use on works, not humans)
P1441	present in work	work in which this fictional entity or historical person is present
P57	director	director(s) of this motion picture, TV-series, stageplay, video game or similar
P179	series	subject is part of a series, whose sum constitutes the object
P136	genre	a creative work's genre or an artist's field of work
P112	founded by	founder or co-founder of this organization, religion or place
P137	operator	person or organization that operates the equipment, facility, or service
P355	subsidiary	subsidiary of a company or organization, opposite of parent company
P176	manufacturer	manufacturer or producer of this product
P86	composer	person(s) who wrote the music
P488	chairperson	presiding member of an organization, group or body
P1056	product or material produced	material or product produced by a government agency, business, industry, facility, or process
P1366	replaced by	person or item which replaces another

Table 7: Relation types and description of development document corpus in FREDo and ReFREDo.

Wikidata ID	Name	Description
P17	country	sovereign state of this item; don't use on humans
P495	country of origin	country of origin of the creative work or subject item
P361	part of	object of which the subject is a part. Inverse property of "has part"
P3373	sibling	the subject has the object as their sibling (brother, sister, etc.)
P463	member of	organization or club to which the subject belongs
P102	member of political party	the political party of which this politician is or has been a member
P1001	applies to jurisdiction	the item (an institution, law, public office ...) belongs to or has power over or applies to the value (a territorial jurisdiction: a country, state, municipality, ...)
P140	religion	religion of a person, organization or religious building, or associated with this subject
P674	characters	characters which appear in this item (like plays, operas, operettas, books, comics, films, TV series, video games)
P194	legislative body	legislative body governing this entity; political institution with elected representatives, such as a parliament/legislature or council
P118	league	league in which team or player plays or has played in
P35	head of state	official with the highest formal authority in a country/state
P272	production company	company that produced this film, audio or performing arts work
P279	subclass of	all instances of these items are instances of those items; this item is a class (subset) of that item
P364	original language of work	language in which a film or a performance work was originally created
P582	end time	indicates the time an item ceases to exist or a statement stops being valid
P25	mother	female parent of the subject
P39	position held	subject currently or formerly holds the object position or public office

Table 8: Relation types and description of in-domain test document corpus in FREDo and ReFREDo.

Wikidata ID	**Name**	**Description**
HYPONYM-OF	hyponym of	subject is a hyponym of the object; subject is a type of the object.
PART-OF	part of	subject is a part of the object.
USED-FOR	used for	subject is used for the object; subject models the object; object is trained on the subject; subject exploits the object; object is based on the subject.
COMPARE	compare	compare two models/methods, or listing two opposing entities.
EVALUATE-FOR	evaluate for	evaluate for
FEATURE-OF	feature of	subject belongs to the object; subject is a feature of the object; subject is under the object domain.
CONJUNCTION	conjunction	function as similar role or use/incorporate with.

Table 9: Relation types and description of cross-domain test document corpus in SciERC.