Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.084290

ARTICLE

COPA: Confidence-Guided Orthogonality-Constrained Prompt Adaptation for Few-Shot Relation Classification

Shunran Duan, Meijuan Yin*, Xiangyang Luo, Lunchong Cui and Chenyu Wang

Henan Provincial Key Laboratory of Cyberspace Situational Awareness, Information Engineering University, Zhengzhou, China

*Corresponding Author: Meijuan Yin. Email: raindot_ymj@163.com

Received: 20 April 2026; Accepted: 16 June 2026; Published: 13 August 2026

ABSTRACT: Few-shot relation classification aims to identify semantic relations between entity pairs under limited annotated data. Although recent prompt learning-based methods have achieved promising performance, they often rely on manually crafted, domain-specific prompt templates, which restrict their transferability across domains. In this paper, building upon the multi-task prompt transfer paradigm of MPT, we propose a Confidence-guided Orthogonality-constraint Prompt Adaptation framework for few-shot relation classification, named COPA. The proposed framework learns a shared domain-invariant prompt matrix together with domain-specific low-rank prompt matrices via multi-domain soft prompt tuning, enabling the transfer of domain-invariant relational knowledge across domains. Unlike MPT, to explicitly disentangle domain-invariant and domain-specific information, we further introduce an orthogonality constraint that encourages the shared prompt to capture invariant relational semantics while forcing the domain-specific prompt to model complementary domain residuals. For target domain adaptation, we reuse the shared prompt as prior knowledge and combine it with a target domain-specific low-rank matrix, followed by a confidence-guided prompt adaptation strategy that encourages domain-invariant knowledge preservation while facilitating efficient adaptation to the target domain. Experiments on four public Chinese datasets demonstrate that COPA achieves consistently competitive and stable results, with 1.09%–2.08% absolute Micro-F1 improvements over the strongest MPT-based baseline in representative settings and overall gains ranging from 0.83% to 8.72% across all compared baselines.

KEYWORDS: Natural language processing; relation classification; orthogonality constraint; confidence-guided prompt adaptation

1 Introduction

Few-shot Relation Classification (RC) aims to identify the semantic relation between a pair of entities from a given text under limited labeled data. Early RC studies mainly relied on Deep Neural Networks(DNNs), including Convolutional Neural Networks [1–3], Recurrent Neural Networks [4–6], and Graph Neural Networks [7–9]. Although these methods have achieved promising results, they generally depend on sufficient supervised data and task-specific architectural design, which limits their effectiveness in few-shot settings.

In practical few-shot scenarios, traditional DNN-based methods face several challenges. They often rely heavily on labeled data to learn robust relation patterns, resulting in poor performance under severe data scarcity. Their learned representations are usually domain-dependent, which weakens cross-domain generalization. In addition, such models often struggle to capture long-range dependencies and implicit

background knowledge required for relation inference. These limitations motivate the development of more sample-efficient approaches that can better exploit large-scale pre-trained knowledge.

With the success of transformer-based pre-trained language models, BERT [10] and its variants have substantially advanced RC. R-BERT [11], the first BERT-based RC model, introduces entity markers to highlight target entity positions and has inspired many extensions in specialized domains. Subsequent studies further improve entity representation and semantic interaction modeling, demonstrating the effectiveness of pre-trained contextual encoders for RC.

More recently, Large Language Models (LLMs) have promoted prompt learning as a promising paradigm for few-shot RC [12–14]. Existing prompt-based methods can be broadly divided into discrete prompts and continuous prompts. Discrete prompts leverage manually designed templates but are highly dependent on domain expertise and sensitive to prompt wording. Continuous prompts alleviate the cost of manual template design by learning soft prompt embeddings in a data-driven manner. Nevertheless, most existing continuous prompt methods for RC optimize a unified prompt representation, without explicitly modeling the distinction between domain-invariant relational knowledge and domain-specific linguistic patterns. Consequently, they may adapt insufficiently to the target domain under few-shot supervision, resulting in a trade-off between transferability and domain adaptability.

One of the representative prompt learning studies is MPT [15], which introduces multi-task prompt learning to facilitate knowledge sharing among different tasks. MPT shows that shared prompt representations can effectively capture common knowledge while retaining task-specific information. However, the setting studied in this paper is different from the original multi-task scenario of MPT. We focus on single-task multi-domain few-shot relation classification, where all domains share the same relation classification objective but exhibit different linguistic distributions, entity contexts, and domain-specific relational expressions.

Inspired by MPT [15], to mitigate aforementioned issues, we propose a few-shot relation classification approach that integrates orthogonality-constrained disentangled prompt learning with confidence-guided prompt adaptation. The proposed method mainly differs from MPT in two aspects. Specifically, the orthogonality constraint explicitly separates domain-invariant relational knowledge from domain-specific linguistic variations, thereby improving cross-domain adaptability, while the confidence-guided prompt adaptation strategy dynamically allocates optimization strength to enhance discriminative capacity within the target domain under few-shot supervision. Through this design, the proposed method alleviates the dependence of existing methods on manually designed discrete prompts and enhances the robustness of continuous prompt learning in cross-domain scenarios.

The main contributions of this paper are as follows:

(1) We propose a disentangled prompt learning framework for relation classification, in which the prompt representation is decomposed into a domain-invariant matrix and a domain-specific low-rank matrix. An orthogonality constraint is further introduced to encourage the two components to capture complementary information, thereby improving both parameter efficiency and cross-domain adaptability.

(2) We develop a confidence-guided prompt adaptation strategy for few-shot cross-domain transfer. By dynamically adjusting the optimization strength of the domain-invariant and domain-specific prompts according to model uncertainty, the proposed method preserves transferable knowledge while enabling efficient adaptation to target-domain characteristics.

(3) Extensive experiments on four public datasets demonstrate that the proposed method consistently outperforms competitive baselines in terms of Micro-F1, while also showing lower performance variance.

2 Related Work

Domain-invariant knowledge transfer has long been utilized for few-shot RC task [16–18]. Additionally, with the rise of large language models (LLMs), prompt learning has emerged as a new solution paradigm for few-shot RC task, gradually gaining favor among researchers [13,14,19–21].

2.1 Domain-Invariant Knowledge Transfer Based Methods

Domain-invariant knowledge transfer methods assume a distribution over domains of a certain task, $D \sim p(D)$, where each domain D is associated with a domain-specific data distribution $p_D(x, y)$, a label space $\mathcal{Y}_D$, a support set $\mathcal{S}_D$ and a query set $\mathcal{Q}_D$. The objective is to learn meta-parameters θ that encode domain-invariant knowledge, and then to optimize the annotation-scarce target domain-specific parameters $\phi_{D_*} = A(\mathcal{S}_{D_*}; \theta)$ with an adaptation operator A so that the model performs well on Q_{D_*}. Formally, we can proceed episodically with annotation-sufficient domains $\{D_i\}_{i=1}^{M}$ and pairs $(\mathcal{S}_i, \mathcal{Q}_i)$ to obtain the optimal $\hat{\theta}$ as follows:

$$\hat{\theta} = \min_{\theta} \sum_{i=1}^{M} \mathcal{L}(\mathcal{Q}_i; \theta, A(\mathcal{S}_i; \theta)), \quad \mathcal{L}(\mathcal{Q}_D; \theta, A(\mathcal{S}_D; \theta)) = \frac{1}{|\mathcal{Q}_D|} \sum_{(x,y)\in\mathcal{Q}_D} \ell(f(x; \theta, A(\mathcal{S}_D; \theta)), y), \tag{1}$$

where $\ell(f(x; \theta, A(\mathcal{S}_D; \theta)), y)$ represents the prediction loss of the model on the data pair (x, y).

Transfer to a target domain D_* uses the learned $\hat{\theta}$ to optimize $\phi_* = A(\mathcal{S}_*; \hat{\theta})$, and then evaluates the model $f(\cdot; \hat{\theta}, \phi_*)$ on the target data. In short, the mechanism distills a compact task summary from various annotation-sufficient domains $\{D_i\}_{i=1}^{M}$ of a certain task and conditions inference on this summary so that knowledge aggregated for the task can be used to new annotation-scarce domains.

There are some representative domain-invariant knowledge transfer based methods for few-shot RC. MIML [16] leverages class semantic concepts to guide optimizing process, enabling more effective initialization and faster adaptation by connecting instance-based and semantic information. Geng et al. [17] propose a meta-learning framework for few-shot RC that leverages support instance knowledge and cross-domain task enrichment to learn better representations from limited data. Obamuyide and Vlachos [18] propose a model-agnostic optimizing protocol that trains relation classifiers to acquire parameters that are easily adaptable for both well-supervised and low-resource relations.

2.2 Prompt Learning-Based Methods

Prompt learning reformulates downstream tasks as language modeling problems by bridging pre-training and fine-tuning, thereby enabling effective utilization of pre-trained knowledge under limited supervision. It typically consists of three components: a pre-trained language model parameterized by θ_{PLM}, a prompt template function $\mathcal{P}(\cdot)$, and a verbalizer $\mathcal{V}(\cdot)$. Given an input instance x with entities e_1 and e_2, the template function transforms it into a prompted sequence, e.g., $x_{prompt} = \mathcal{P}(x) =$ "$[x]$ The relation between $[e_1]$ and $[e_2]$ is $[\text{MASK}]$". The Pre-trained Language Model (PLM) then predicts the probability distribution over the vocabulary at the mask position, denoted as $p(w \mid x_{prompt}, \theta_{PLM})$. The verbalizer maps these word-level probabilities to label probabilities: $p(y \mid x) = \sum_{w\in\mathcal{V}(y)} p(w \mid x_{prompt}, \theta_{PLM})$, where $\mathcal{V}(y)$ is the set of label words associated with class y. Model parameters are optimized by minimizing the cross-entropy loss over the training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$:

$$\theta^{\star}_{PLM} = \arg\min_{\theta_{PLM}} \sum_{i=1}^{N} \mathcal{L}_{CE}(y_i, p(y \mid x_i, \theta_{PLM})) \tag{2}$$

Prompt learning reformulates RC as a masked token prediction or text generation task. By introducing natural language prompt templates to align RC with the pre-training objective, it enables models to exploit implicit semantic, syntactic, and world knowledge encoded in pre-trained parameters. This alignment reduces reliance on labeled data, emphasizes informative entity–context interactions, and alleviates overfitting in low-resource settings, thereby improving sample efficiency, robustness, and cross-domain transferability.

There are some representative prompt learning-based methods for few-shot RC. Zhang et al. [14] propose a prompt-based method for relation classification that employs label-representing tokens, an entity-aware contrastive module, and an attention query strategy to enhance performance in low-resource settings. Xie et al. [20] propose a novel inductive model that leverages hard prompts to help construct relevant subgraphs based on PLMs and enriches relation embeddings with textual descriptions for enhanced semantic representation. Jiang et al. [13] propose RCBP which employs bidirectional prompt templates to capture entity order and aligns the probability distributions of both prompt learning and fine-tuning, enhanced by LLM-based data augmentation. Moslemnejad and Reed [21] propose a method that leverages Frame Semantic Parsing to extract semantic frames and employs a RoBERTa PLM model to train dual prompt templates within a Siamese network architecture to enhance relation classification. Liu et al. [19] provide a comprehensive survey on prompt-based learning paradigms for various NLP tasks including relation classification.

2.3 Challenges and Limitations

Despite their advantages, both domain-invariant knowledge transfer and prompt learning methods exhibit notable limitations in few-shot relation classification. Domain-invariant transfer methods, though capable of learning domain-invariant initialization via multi-domain episodic training, often suffer from degraded adaptability under significant distribution shifts between $\mathcal{D}_{train}$ and $\mathcal{D}_{target}$. The learned meta-parameters $\hat{\theta}$ may retain domain-specific biases, limiting adaptation when the target support set $\mathcal{S}_{target}$ is scarce. Prompt learning methods face two key challenges. (1) *Optimization difficulty*: In few-shot settings, continuous prompt learning is under-constrained, leading to overfitting and unstable optimization. (2) *Domain adaptation bottleneck*: Although prompts effectively elicit pre-trained knowledge, they lack explicit mechanisms for transferring domain-invariant relational patterns. Consequently, when adapting from $\mathcal{D}_{source}$ to $\mathcal{D}_{target}$, learned prompts may fail to generalize across domains, resulting in suboptimal performance.

2.4 Relation to Orthogonality Constraint and Uncertainty-Based Weighting

Inspired by the prompt decomposition mechanism of MPT [15], we further incorporate orthogonality constraint and confidence-guided prompt adaptation. In this section, we explain the motivation for these modifications and highlight the key differences between our method and traditional practice.

Orthogonality constraint has been widely used in domain adaptation and representation learning to encourage different latent components to capture non-redundant information. Existing orthogonality-based disentanglement methods usually impose such constraints on feature representations, latent factors, or domain-specific encoders, with the goal of separating shared and private representations across domains. In contrast, COPA applies orthogonality constraint to the prompt space of a frozen LLM for few-shot relation classification. Specifically, COPA constrains the shared domain-invariant prompt and the domain-specific low-rank prompt to be complementary, so that transferable relational semantics and domain-dependent residual information can be separated within a parameter-efficient prompt adaptation framework.

Uncertainty-based weighting has been studied in curriculum learning, self-paced learning, and domain adaptation, where instance-level confidence is often used to select reliable pseudo-labels, down-weight noisy samples, or adjust training difficulty. For instance, Litrico et al. [22] estimate pseudo-label uncertainty in source-free unsupervised domain adaptation and use it to reweight the classification loss, reducing the influence of noisy pseudo-labels. Different from these methods, COPA uses confidence to regulate prompt adaptation in a labeled few-shot target domain. The confidence score is computed from the teacher-forced probability of the gold output sequence and is used to weight the per-example adaptation loss and asymmetrically scale the learning rates of the domain-invariant and domain-specific prompt components. This design allows COPA to preserve transferable prompt knowledge on high-confidence instances while assigning stronger domain-specific updates to low-confidence instances.

The combination of these components is motivated by the transfer dilemma in single-task multi-domain few-shot RC. Prompt decomposition provides a compact parameterization for separating shared and domain-specific prompt knowledge, but decomposition alone does not ensure that the two components encode non-redundant information, especially for single task. Orthogonality constraint is therefore introduced to explicitly encourage complementarity between the domain-invariant and domain-specific prompts. During target-domain adaptation, confidence-guided weighting further determines how strongly each target instance should influence the two prompt components. Thus, prompt decomposition, orthogonality constraint, and confidence-guided adaptation address parameterization, disentanglement, and adaptation control, respectively.

3 Methodology

3.1 Task Definition

Given an input sentence $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ with two marked entities, namely the head entity e_h and the tail entity e_t, the goal of relation classification is to predict the semantic relation $r \in \mathcal{R}$ between them, where $\mathcal{R} = \{r_1, r_2, \ldots, r_m\}$ denotes the predefined relation set. The entities may span one or multiple consecutive tokens, and their corresponding entity types are denoted by t_h and t_t.

To align symbolic relation labels with the natural language output space of large language models (LLMs), we adopt *label verbalization*. For each relation type r, a natural language verbalization $\bar{\boldsymbol{r}} = (\bar{r}_1, \bar{r}_2, \ldots, \bar{r}_{|\bar{\boldsymbol{r}}|})$ is defined as its textual expression. This verbalization provides a linguistically natural form of the relation label and facilitates generation-based prediction within the LLM framework.

For example, given the sentence *"The squirrel has a fluffy tail."*, the relation between e_h = "$squirrel$" and e_t = "$fluffy\ tail$" can be classified as *Part-Whole*, whose verbalization may be defined as *"is part of"*. During training, the model learns to generate such verbalizations for corresponding relation types, and during inference, the generation probability of each verbalization is used to determine the predicted relation.

3.2 The Overall Framework

To alleviate the challenges discussed in Section 2.3, we propose COPA (Confidence-guided Orthogonality-constrained Prompt Adaptation), which operates through three sequential and interconnected stages as shown in Fig. 1, each designed to address specific challenges in few-shot cross-domain relation classification:

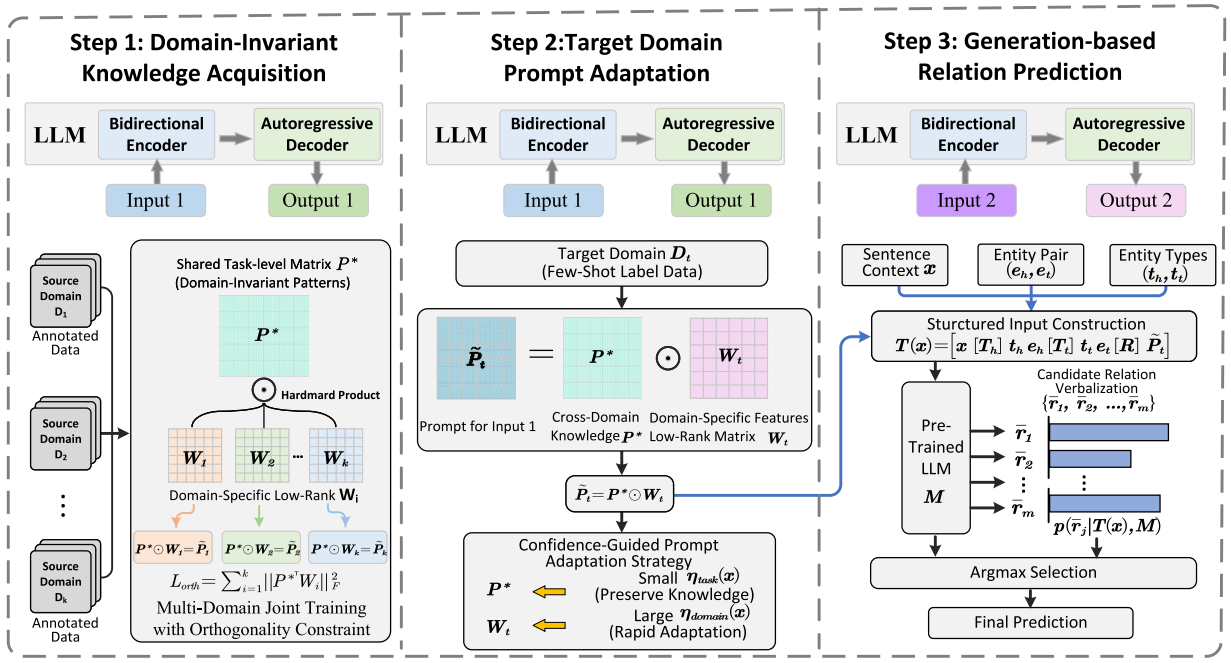

Figure 1: The framework of COPA (confidence-guided orthogonality-constrained prompt adaptation). COPA consists of three stages: (1) domain-invariant knowledge acquisition, which jointly learns a domain-invariant prompt and domain-specific low-rank prompts with orthogonality constraints to disentangle invariant and domain-specific knowledge; (2) target domain prompt adaptation, which performs confidence-guided prompt adaptation strategy to adapt the prompt to the target domain while preserving transferable knowledge; and (3) generation-based relation prediction, which uses the adapted prompt to guide the LLM in generating relation verbalizations for final prediction.

(1) Domain-Invariant Knowledge Acquisition. We jointly optimize across multiple source domains to learn a domain-invariant prompt P^* and domain-specific low-rank prompts $\{W_i\}_{i=1}^{k}$. For each domain $\mathcal{D}_i$, the effective prompt is defined as $\widetilde{P}_i = P^* \odot W_i$, where W_i is low-rank parameterized. An orthogonality constraint is imposed between P^* and W_i to disentangle domain-invariant relational knowledge from domain-specific variations, enabling parameter-efficient modeling with improved cross-domain adaptability.

(2) Target Domain Prompt Adaptation. For a target domain, a domain-specific prompt W_t is initialized from averaged source parameters and combined with P^* to form $\widetilde{P}_t = P^* \odot W_t$. We adopt a confidence-guided prompt adaptation strategy that encourages domain-invariant knowledge preservation in P^* while enabling efficient adaptation of W_t under few-shot supervision.

(3) Generation-Based Relation Prediction. Given a target input, we construct a structured prompt and use $\widetilde{P}_t$ to guide the LLM in generating relation verbalizations. The confidence of each relation is computed via the average token-level generation probability of its verbalization. Entity type constraints are further applied to filter incompatible candidates, and the relation with the highest confidence is selected as the final prediction.

3.3 Domain-Invariant Knowledge Acquisition

We formulate relation classification as a conditional generation task, where an LLM is prompted to generate the verbalized relation for a given entity pair. To this end, we design a unified prompt template as Eq. (3) (indicated by Input 1 format in Fig. 2):

$$T(\boldsymbol{x}) = \boldsymbol{x}\ [T_h]\ e_h\ [T_t]\ e_t\ [R]\ \widetilde{P} \tag{3}$$

where $[T_h]$ and $[T_t]$ denote the type sentinel tokens of the head and tail entities, respectively, $[R]$ is the relation sentinel token, and $\widetilde{P}$ is the tunable prompt. We use sentinel tokens to indicate the locations of the corresponding information to be generated in the target sentence. The corresponding target sequence is defined as Eq. (4) (indicated by Output 1 format in Fig. 2):

$$\boldsymbol{y} = [T_h]\ t_h\ [T_t]\ t_t\ [R]\ \bar{\boldsymbol{r}}\ [E] \tag{4}$$

where $\bar{\boldsymbol{r}}$ is the label verbalization of relation r, and $[E]$ is the end-of-sequence token.

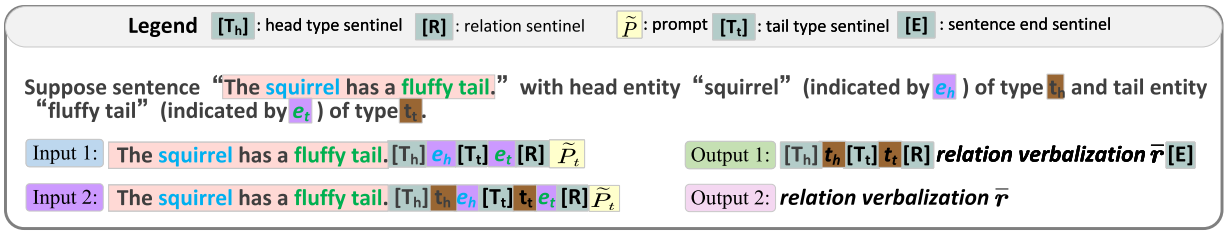

Figure 2: The input and output format example of COPA.

For simplicity, we denote the LLM input as the concatenation of the text sequence and the prompt $\widetilde{P}$. In practice, the text sequence is tokenized and mapped to embeddings, while $\widetilde{P}$ is a learnable matrix directly defined in the embedding space. The concatenation is thus performed at the embedding level, and the resulting sequence is fed into the LLM. We use "text_sequence $\widetilde{P}$" to represent this input for brevity.

To capture both domain-invariant relational semantics and domain-specific linguistic patterns, we decompose the prompt into a shared domain-invariant matrix P^* and a domain-specific low-rank matrix W_i. For domain $\mathcal{D}_i$, the effective prompt is defined as a Hadamard product of two components following Wang et al. [15], i.e., $\widetilde{P}_i = P^* \odot W_i$, where W_i is parameterized in low-rank form for parameter-efficient adaptation. The Hadamard product $\odot$ is applied element-wise over the entire prompt matrix. If we suppose the rank of W_i is 1, so $W_i = u_i v_i^T$, where $u_i \in \mathbb{R}^d, v_i \in \mathbb{R}^l, \{W_i, P^*, \widetilde{P}_i\} \in \mathbb{R}^{d \times l}$, d denotes the number of prompt tokens and l denotes the backbone LLM's internal embedding dimension. This design is motivated by the observation that relation classification involves both invariant relational knowledge and domain-dependent surface realizations.

To explicitly disentangle these two factors, we impose an orthogonality constraint between P^* and W_i by minimizing the Frobenius norm:

$$\mathcal{L}_{orth} = \sum_{i=1}^{k} \left\| P^{*\top} W_i \right\|_F^2 \tag{5}$$

which encourages W_i to encode complementary domain-specific residuals rather than redundant invariant information.

Fig. 3 illustrates the decomposition of the prompt into a shared matrix P^* and a domain-specific rank-1 matrix W_i as an example. The shared component P^* is jointly optimized across all domains to capture domain-invariant relational knowledge. In contrast, W_i is optimized independently for each domain and is parameterized as a rank-1 matrix, i.e., $W_i = u_i v_i^\top$, enabling parameter-efficient modeling of domain-specific variations. This joint formulation allows P^* to encode transferable knowledge, while W_i provides flexibility for domain adaptation.

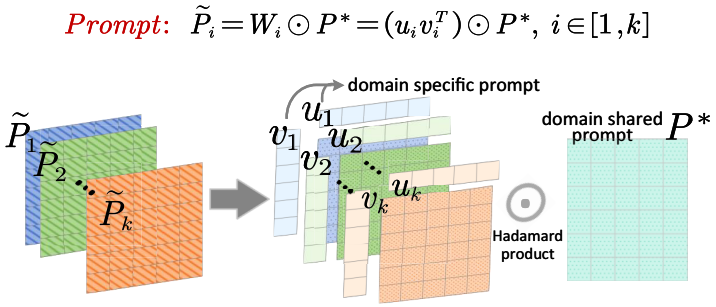

Figure 3: Prompt decomposition diagram with rank-1 matrix W_i as an example.

Given an LLM $\mathcal{M}$, the overall multi-domain training objective is defined as the weighted sum of the loss function of each individual domain, which is the negative log-likelihood of the target sequence $\boldsymbol{y}$:

$$\mathcal{L} = \sum_{i=1}^{k} \lambda_i \mathcal{L}_i + \gamma \mathcal{L}_{orth}, \quad \mathcal{L}_i = -\sum_{j=1}^{|\boldsymbol{y}|} \log p\left(y_j \mid y_{<j}, T(\boldsymbol{x}), \mathcal{M}\right) \tag{6}$$

where λ_i is the weight for domain $\mathcal{D}_i$ with $\sum_{i=1}^{k} \lambda_i = 1$, γ controls the strength of orthogonality constraint, and $y_{<j}$ denotes previously generated tokens.

3.4 Target Domain Prompt Adaptation

After multi-domain training, the shared prompt P^* encodes transferable domain-invariant relational knowledge and serves as a strong prior for target-domain adaptation. To incorporate domain-specific characteristics under few-shot supervision, we construct the target prompt as $\widetilde{P}_t = P^* \odot W_t$, where we suppose W_t is a rank-1 matrix, so $\widetilde{P}_t = P^* \odot W_t = P^* \odot \left(u_t v_t^T\right)$.

To facilitate stable adaptation, u_t and v_t are initialized by averaging k source domain-specific rank-1 matrices $\{W_i = u_i v_i^T\}_{i=1}^{k}$:

$$u_t = \frac{1}{k}\sum_{i=1}^{k} u_i, \quad v_t = \frac{1}{k}\sum_{i=1}^{k} v_i \tag{7}$$

which provides a consensus initialization and improves convergence under limited data.

Given the structured input $T(\boldsymbol{x})$, we freeze all parameters of the backbone model $\mathcal{M}$ and only adapt the prompt parameters $\widetilde{P}_t$ in the target domain. To emphasize informative hard instances for target-domain adaptation, we introduce a **confidence-guided prompt adaptation** strategy. For each labeled target-domain instance, the confidence score is computed using the teacher-forced probability of the gold output sequence, rather than the probability of a greedily decoded prediction. Specifically, at the beginning of the e-th adaptation epoch, we estimate the confidence score as

$$s^{(e)}(\boldsymbol{x}) = \frac{1}{|\boldsymbol{y}|}\sum_{j=1}^{|\boldsymbol{y}|} p_{\Theta_t^{(e-1)}}\left(y_j \mid y_{<j}, T(\boldsymbol{x}), \mathcal{M}\right), \quad \omega^{(e)}(\boldsymbol{x}) = 1 - s^{(e)}(\boldsymbol{x}), \tag{8}$$

where $y_{<j}$ denotes the gold prefix under teacher forcing, and $p_{\Theta_t^{(e-1)}}(\cdot)$ is computed by the frozen LLM's decoder through the softmax operation over the vocabulary conditioned on the target prompt parameters from the previous epoch. Thus, $s^{(e)}(\boldsymbol{x})$ measures how confidently the current prompt-adapted model assigns probability to the ground-truth relation description.

To avoid introducing a coupled optimization between the confidence estimator and the prompt parameters within each gradient step, $\omega^{(e)}(\boldsymbol{x})$ is recomputed only at epoch boundaries and is kept fixed during the e-th epoch. Moreover, the uncertainty weight is treated as a constant coefficient when optimizing the prompt parameters, i.e., gradients are not propagated through $\omega^{(e)}(\boldsymbol{x})$. The weighted adaptation objective is therefore defined as

$$\mathcal{L}_{adapt}^{(e)} = -\sum_{\boldsymbol{x}\in\mathcal{D}_t} \mathrm{sg}\left[\omega^{(e)}(\boldsymbol{x})\right] \sum_{j=1}^{|\boldsymbol{y}|} \log p_{\Theta_t}\left(y_j \mid y_{<j}, T(\boldsymbol{x}), \mathcal{M}\right), \tag{9}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operation.

The uncertainty-weighted objective assigns larger weights to low-confidence instances, which are more likely to reflect target-domain-specific patterns insufficiently captured by the transferred domain-invariant prompt. Therefore, the model focuses more on informative and difficult target-domain samples during adaptation. We further modulate the learning rates of the domain-invariant and domain-specific prompt:

$$\eta_{task}^{(e)}(\boldsymbol{x}) = \eta_{task}\left(1 - \mathrm{sg}\left[\omega^{(e)}(\boldsymbol{x})\right]\right), \quad \eta_{domain}^{(e)}(\boldsymbol{x}) = \eta_{domain}\mathrm{sg}\left[\omega^{(e)}(\boldsymbol{x})\right), \tag{10}$$

where η_{task} and η_{domain} are base learning rates for the invariant prompt and the domain-specific prompt, respectively. This formulation ensures that: (1) For high-confidence instances, $\omega^{(e)}(\boldsymbol{x})$ is small, and the adaptation mainly preserves and slightly refines the invariant prompt initialized from P^*. (2) For low-confidence instances, $\omega^{(e)}(\boldsymbol{x})$ becomes larger, assigning stronger updates to the domain-specific prompt matrix W_t so that the model can better capture target domain-specific patterns.

3.5 Generation-Based Relation Prediction

In the target-domain inference stage, we apply the adapted prompt $\widetilde{P}_t$ to predict relations for test instances. Unlike the training stages, entity types are explicitly provided as auxiliary inputs, allowing the model to focus on relation prediction under type constraints. For an input sentence $\boldsymbol{x}$ with head entity e_h of type t_h and tail entity e_t of type t_t, the structured input is defined as

$$T(\boldsymbol{x}) = \boldsymbol{x}\ [T_h]\ t_h\ e_h\ [T_t]\ t_t\ e_t\ [R]\ \widetilde{P} \tag{11}$$

where $[T_h]$, $[T_t]$, and $[R]$ are sentinel tokens, and $\widetilde{P}$ denotes the adapted prompt.

We adopt a generation-based scoring strategy for relation prediction. For each candidate relation $r \in \mathcal{R}$, we obtain a label verbalization $\bar{\boldsymbol{r}} = [\bar{r}_1, \ldots, \bar{r}_{|\bar{\boldsymbol{r}}|}]$ via a generation-and-selection strategy. Specifically, the LLM used in COPA first generates multiple candidate verbalizations, and then selects the most suitable one based on semantic consistency, clarity, and conciseness. The confidence score of relation r is computed as the average token-level generation probability:

$$c_r = \frac{1}{|\bar{\boldsymbol{r}}|}\sum_{j=1}^{|\bar{\boldsymbol{r}}|} p_{\bar{r}_j} \tag{12}$$

where $p_{\bar{r}_j} = p(\bar{r}_j \mid T(\boldsymbol{x}), \bar{r}_1, \ldots, \bar{r}_{j-1}, \mathcal{M})$. This yields a stable sequence-level confidence estimate for each candidate relation.

To improve both efficiency and accuracy, we further introduce entity type compatibility constraints. Specifically, only relations compatible with the observed entity types (t_h, t_t) are retained:

$$\mathcal{R}_{compatible} = \{r \in \mathcal{R} \mid \mathcal{TCM}[t_h, t_t, r] = 1\} \tag{13}$$

where $\mathcal{TCM}$ is a type compatibility matrix. For instance, the relation "married_to" is only meaningful when both entities are of type "Person", so $\mathcal{TCM}[Per, Per, married\_to] = 1$. The final prediction is then obtained by

$$\hat{r} = \arg \max_{r \in \mathcal{R}_{compatible}} c_r \tag{14}$$

This generation-based paradigm effectively leverages the generative knowledge of LLMs, while the adapted prompt and type constraints further improve prediction interpretability, efficiency, and accuracy.

4 Experiments

4.1 Datasets and Experimental Settings

To ensure fair evaluation, the datasets used for domain-invariant knowledge acquisition are disjoint from those used for validation. The training domains include SKE[1] (50 relations, from general domain Baidu Baike), FR2KG [23] (19 relations, from company research reports), and IPRE [24] (34 relations, from general domain for inter-personal relationship extraction). To mitigate domain imbalance, 1000 samples are uniformly selected from each dataset, evenly covering all relation types.

For evaluation, we use four public Chinese datasets: SanWen [25] (9 relations, from Chinese literature text), DuIE [26] (49 relations, from news, entertainment, etc.), ACE 2005 Chinese Corpus[2] (18 relations, from newswire, broadcast news, weblog), and FinRE [27] (44 relations, from financial and economic domain). Few-shot settings are simulated by sampling $K \in \{8, 16, 32\}$ instances per relation only from the training split of each target dataset. Thus, for a dataset with R relation types, the training set contains $R \times K$ labeled relation instances. This follows a low-resource per-relation evaluation protocol rather than an conventional N-way-K-shot setting. For each K, sampling is repeated 5 times, and each split is trained 5 times, resulting in 25 runs. The test split is kept fixed and strictly disjoint from the sampled few-shot training instances in all replications. No data point used for target-domain prompt adaptation is included in the subsequent test set, thereby avoiding data leakage. Performance is reported as the average Micro-F1.

We report the statistics and relation overlap in Appendix A. It should be noted that the source datasets used for multi-domain training consist of heterogeneous texts from broad domains, and their general topical categories may partially overlap with those of the evaluation datasets. However, this does not invalidate the cross-domain evaluation setting. The source and target datasets are strictly disjoint at both the dataset and instance levels. Moreover, these datasets differ in relation schemas, entity distributions, annotation criteria, and textual styles, resulting in clear distributional shifts. Our goal is not to require source and target domains to be completely unrelated, but to examine whether domain-invariant relational knowledge learned from heterogeneous source domains can be transferred to unseen target datasets under few-shot supervision.

All experiments are conducted on Ubuntu 22.04.4 with two NVIDIA Quadro GV100 GPUs (32 GB). The model is implemented with PyTorch [28] and HuggingFace Transformers. We adopt Qwen-7B[3] [29] and DeepSeek-7B[4] [30] as backbone LLMs for COPA, while prompt-based baselines use Qwen-7B. Hyperparameter settings are provided in Table 1.

[1] https://aistudio.baidu.com/datasetdetail/18801.
[2] https://catalog.ldc.upenn.edu/LDC2006T06.
[3] https://huggingface.co/Qwen/Qwen2-7B.
[4] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B.

Table 1: The relevant hyperparameters for experiments.

Hyperparameter	Value
The rank of the domain-specific low-rank matrix	1
The strength of orthogonality constraint γ	0.05
Loss function weight for each domain λ_i, $i \in \{1, 2, 3\}$	1/3
Learning rate for domain-invariant knowledge acquisition	0.1
The proportion of warmup steps to the total training steps	10%
Weight decay for learning rate of domain-invariant knowledge acquisition	1e−4
Learning rate for shared matrix $\boldsymbol{P}^*$ during target-domain adaptation η_{task}	5e−4
Learning rate for low-rank matrix $\boldsymbol{W}_t$ during target-domain adaptation η_{domain}	5e−3
Number of prompt tokens d	100
Hidden size of backbone Qwen-7B	3584
Hidden size of backbone DeepSeek-7B	3584

4.2 Validation Experiments

4.2.1 Baseline Methods

We compare COPA with several representative few-shot relation classification baselines from two categories.

Neural network-based methods. SRE-HGNN [7] enhances sample representations via heterogeneous graph modeling and adaptive neighbor selection. HCRP [31] improves prototypical learning with relation-prototype contrastive learning and task-adaptive strategies. Proto-MCE [3] adopts multichannel encoding and dependency tree attention, further augmented by pseudo-labeling and entropy-based instance selection.

Prompt learning-based methods. MRC-SP [32] is a multilingual prompt-based relation classification method. GenPT [12] reformulates relation classification as a generative infilling task with entity-guided decoding. LabelPrompt [14] models labels as learnable tokens and introduces entity-aware contrastive learning. Prompt Tuning [33] updates task-specific soft prompts while freezing pretrained model parameters. Multitask Prompt Tuning [15] learns a transferable prompt across domains, but does not explicitly enforce disentanglement or confidence-guided adaptation. For Prompt Tuning and Multitask Prompt Tuning, we adopt the same generation-based relation prediction as COPA, denoted as PT-GRP and MPT-GRP, respectively.

4.2.2 Validation Experimental Results and Analysis

The experimental results are summarized in Table 2, which reports the average Micro-F1 scores with standard deviations. The best and second-best results are highlighted in bold and italics, respectively. Fig. 4 further visualizes the performance distribution across 25 runs using box plots, where the dots within the box plots indicate the mean value across the 25 runs. The line chart reflects the trend of micro-F1 scores under different data scales.

Table 2: Micro-F1 (%) values of methods on different datasets.

	Model	SanWen			DuIE			ACE 2005			FinRE		
		K = 8	K = 16	K = 32	K = 8	K = 16	K = 32	K = 8	K = 16	K = 32	K = 8	K = 16	K = 32
Neural Network	SRE-HGNN [7]	51.15 (±1.15)	55.17 (±1.42)	58.63 (±1.70)	65.34 (±1.27)	68.16 (±1.56)	70.32 (±1.69)	63.21 (±1.37)	65.54 (±1.49)	68.75 (±1.56)	36.25 (±1.32)	38.63 (±1.60)	40.17 (±1.95)
	HCRP [31]	56.37 (±1.75)	59.82 (±1.73)	60.43 (±1.65)	68.67 (±1.82)	71.25 (±1.62)	73.86 (±1.37)	67.49 (±1.23)	69.84 (±1.46)	71.79 (±1.59)	39.18 (±1.60)	41.76 (±1.41)	43.28 (±2.06)
	Proto-MCE [3]	52.49 (±1.32)	56.63 (±1.45)	59.17 (±1.51)	66.31 (±1.47)	70.25 (±1.71)	71.39 (±1.56)	64.73 (±1.91)	68.53 (±1.62)	70.41 (±1.86)	38.54 (±1.56)	40.73 (±1.44)	41.81 (±1.65)
Prompt Tuning	MRC-SP [32]	53.76 (±1.54)	57.51 (±1.46)	60.37 (±1.34)	67.61 (±1.80)	70.83 (±1.41)	71.46 (±1.79)	65.71 (±1.43)	69.06 (±0.95)	70.86 (±2.11)	38.17 (±1.59)	40.27 (±1.42)	41.06 (±1.37)
	GenPT [12]	56.67 (±1.37)	60.54 (±1.21)	61.38 (±1.43)	68.64 (±1.36)	71.53 (±0.86)	72.86 (±1.26)	68.39 (±0.81)	70.12 (±1.48)	71.57 (±1.27)	39.67 (±1.34)	40.83 (±1.24)	41.95 (±0.85)
	LabelPrompt [14]	56.27 (±1.25)	60.37 (±1.19)	61.25 (±0.89)	68.07 (±0.97)	71.76 (±1.26)	73.49 (±1.33)	68.74 (±1.03)	70.06 (±1.26)	71.94 (±1.14)	40.18 (±0.83)	*43.03* (±1.27)	*45.25* (±0.88)
	PT-GRP [33]	58.27 (±1.15)	61.03 (±1.09)	61.87 (±0.92)	69.52 (±1.25)	72.02 (±1.14)	73.86 (±1.12)	68.79 (±1.17)	71.88 (±1.24)	73.17 (±1.16)	40.15 (±1.19)	42.01 (±1.21)	44.65 (±1.08)
	MPT-GRP [15]	58.63 (±0.81)	61.83 (±0.79)	62.64 (±0.75)	69.83 (±0.75)	72.18 (±0.84)	*74.27* (±0.73)	68.84 (±0.86)	72.04 (±0.73)	73.58 (±0.85)	41.26 (±0.79)	42.11 (±0.71)	44.18 (±0.61)
	COPA (deepseek)	*59.15* (±0.62)	*62.12* (±0.77)	*63.81* (±0.71)	**71.69** (±0.84)	*72.87* (±0.91)	74.15 (±0.75)	*69.05* (±0.81)	**73.53** (±0.87)	*74.16* (±0.77)	*41.53* (±0.83)	42.54 (±0.79)	44.83 (±0.72)
	COPA (Qwen)	**59.87** (±0.87)	**62.93** (±0.65)	**64.12** (±0.58)	*71.53* (±0.76)	**73.92** (±0.81)	**75.66** (±0.54)	**70.22** (±0.73)	*73.38* (±0.69)	**74.67** (±1.08)	**42.73** (±0.68)	**43.86** (±0.63)	**46.26** (±0.57)

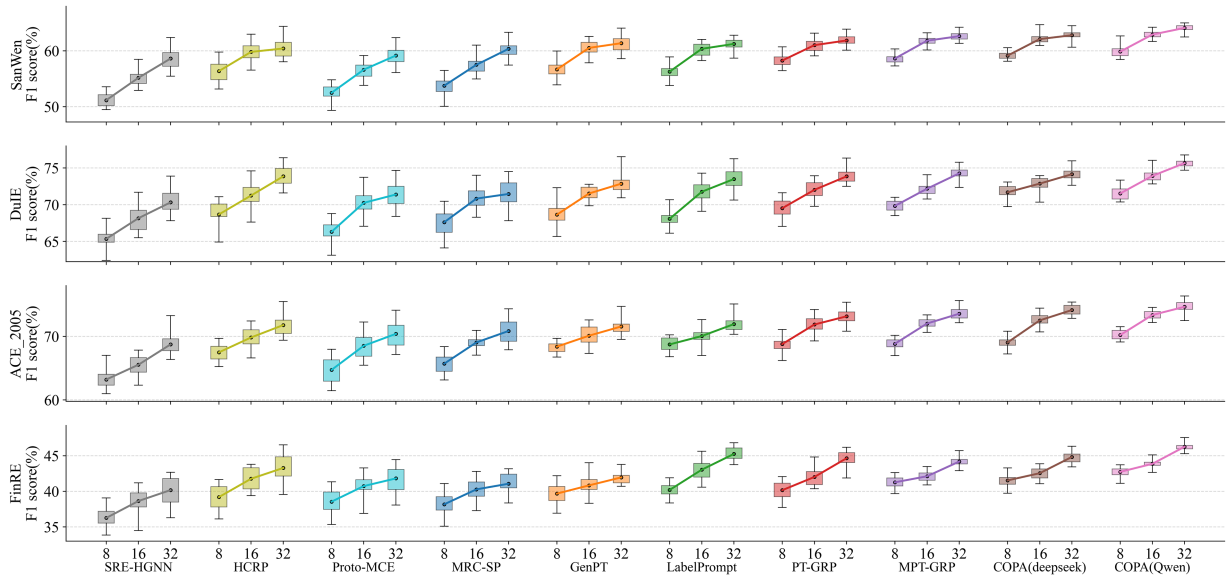

Figure 4: The box plots and trend of Micro-F1 values of different methods on $K(8,16,32)$ annotated samples (horizontal axis) across 4 datasets (vertical axis).

As shown in Table 2 and Fig. 4:

(1) **Overall performance.** COPA consistently outperforms baselines, achieving improvements of 0.83%–8.72% in Micro-F1. This gain is mainly attributed to two factors. First, the orthogonality-constrained prompt disentanglement separates domain-invariant semantics from domain-specific variations, enabling more effective knowledge transfer while avoiding redundant encoding. Second, the confidence-guided prompt

adaptation dynamically allocates optimization strength based on prediction uncertainty, preserving general knowledge while improving discrimination on target-domain instances.

(2) **Stability.** COPA achieves relatively lower standard deviations across all settings, indicating enhanced robustness. This stability is attributed to the domain-invariant knowledge encoded in the shared prompt P^* and the confidence-guided prompt adaptation mechanism, which together reduce overfitting and ensure more consistent performance under limited supervision.

(3) **Advantage of prompt-based methods over Deep Neural Network (DNN) methods.** Most prompt-based methods outperform DNN-based approaches across datasets and shot settings, indicating their superior suitability for few-shot relation classification. This advantage stems from their ability to better exploit the rich semantic knowledge encoded in pre-trained language models, whereas DNN-based methods are more constrained by limited labeled data and thus exhibit inferior performance.

4.2.3 Comparison with Other LLM-Based Adaptation Baselines

To further compare COPA with stronger LLM-based adaptation strategies, we add three baselines using the same Qwen-7B and DeepSeek-7B backbones. First, Zero-shot In-Context Learning(ICL) directly prompts the frozen LLM with the relation classification template and candidate relation verbalizers. Second, Few-shot ICL augments the prompt with K labeled demonstrations selected from the target-domain training set for each test instance. Since our K setting denotes K labeled instances per relation class, including all $R \times K$ instances as demonstrations is infeasible for datasets with many relations; therefore, the ICL baseline uses K demonstrations per query. Third, LoRA [34] fine-tunes the same LLM on K samples per relation setting using the LLaMA-Factory[5] framework, providing a stronger but more expensive adaptation baseline.

For few-shot ICL, we construct the prompt using the template shown below, which includes the task instruction, candidate relation labels, labeled demonstrations, and the test instance. For zero-shot ICL, we use the same template while removing the demonstration section.

ICL Prompt Template (English Translation Version)
You are an expert in Chinese relation classification. Given a sentence and two marked entities, your task is to identify the semantic relation between the head entity and the tail entity.
Candidate relation labels: {relation_label_list}
Example 1: Sentence: {demo_sentence_1}; Head entity: {demo_head_1}; Head entity type: {demo_head_type_1}; Tail entity: {demo_tail_1}; Tail entity type: {demo_tail_type_1}; Relation: {demo_relation_1}. Example 2: .....
Now classify the following instance.
Sentence: {test_sentence}; Head entity: {test_head_entity}; Head entity type: {test_head_type}; Tail entity: {test_tail_entity}; Tail entity type: {test_tail_type}.
Please choose the most appropriate relation label from the candidate relation labels. Only output the relation label. Do not output any explanation.
Relation:

As shown in Table 3, with the best results in bold and the second-best in italics:

(1) Zero-shot and few-shot ICL perform worse than trainable adaptation methods, and few-shot ICL shows larger variance, indicating its sensitivity to demonstration selection. This may suggest that ICL may require sufficiently large and capable LLMs to fully exploit the information provided by demonstrations. In our

[5] https://github.com/hiyouga/LLaMAFactory.

7B-scale backbone setting, demonstration-based prompting alone may be insufficient for stable few-shot relation classification.

(2) LoRA achieves strong results on DuIE and FinRE, as it updates additional LLM parameters and have relatively more training data for these two datasets, thus provides a stronger but more expensive adaptation baseline. Nevertheless, COPA remains highly competitive and achieves the best overall average performance, with particularly strong results on SanWen and ACE 2005. These results show that COPA offers an effective and parameter-efficient alternative to more expensive LLM adaptation methods.

Table 3: Micro-F1 (%) values of LLM-based adaptation methods

Method	SanWen			DuIE			ACE 2005			FinRE		
	K = 8	K = 16	K = 32	K = 8	K = 16	K = 32	K = 8	K = 16	K = 32	K = 8	K = 16	K = 32
Zero-shot (deepseek)	53.34 (±0.14)			67.53 (±0.17)			66.44 (±0.21)			35.31 (±0.16)		
Zero-shot (Qwen)	54.22 (±0.11)			68.35 (±0.14)			65.97 (±0.18)			34.86 (±0.15)		
Few-shot (deepseek)	55.41 (±1.81)	57.73 (±1.62)	58.45 (±1.63)	69.85 (±2.56)	70.52 (±2.35)	70.97 (±1.96)	67.71 (±1.97)	68.25 (±1.68)	68.57 (±1.49)	36.82 (±2.63)	37.51 (±2.41)	37.94 (±2.07)
Few-shot (Qwen)	55.64 (±1.77)	58.23 (±1.67)	58.64 (±1.41)	69.73 (±2.34)	70.86 (±2.47)	71.25 (±2.04)	67.83 (±1.88)	68.38 (±1.72)	68.51 (±1.45)	37.12 (±2.75)	37.81 (±2.48)	38.14 (±2.13)
Lora (deepseek)	54.82 (±1.06)	57.87 (±1.12)	61.38 (±1.13)	**72.21** (±1.12)	*74.18* (±0.97)	*76.65* (±1.13)	67.75 (±1.03)	69.73 (±1.15)	72.85 (±0.93)	42.17 (±1.05)	*44.26* (±0.87)	*47.83* (±0.94)
Lora (Qwen)	54.75 (±1.02)	57.69 (±0.93)	61.71 (±1.08)	71.57 (±1.05)	**74.35** (±1.07)	**77.16** (±0.85)	67.92 (±1.13)	71.19 (±0.91)	73.63 (±0.94)	*42.41* (±1.13)	**44.83** (±0.96)	**48.02** (±0.87)
COPA (deepseek)	*59.15* (±0.62)	*62.12* (±0.77)	*63.81* (±0.71)	*71.69* (±0.84)	72.87 (±0.91)	74.15 (±0.75)	*69.05* (±0.81)	**73.53** (±0.87)	*74.16* (±0.77)	41.53 (±0.83)	42.54 (±0.79)	44.83 (±0.72)
COPA (Qwen)	**59.87** (±0.87)	**62.93** (±0.65)	**64.12** (±0.58)	71.53 (±0.76)	73.92 (±0.81)	75.66 (±0.54)	**70.22** (±0.73)	*73.38* (±0.69)	**74.67** (±1.08)	**42.73** (±0.68)	43.86 (±0.63)	46.26 (±0.57)

4.3 Parameter Analysis

4.3.1 Effects of Rank of Domain-Specific Matrix

To investigate the impact of the rank of the domain-specific low-rank matrix, we conduct a sensitivity analysis on COPA (Qwen) by varying the rank from 1 to 3 across multiple datasets under $K \in \{8, 16, 32\}$ settings. The results are reported in Table 4, where the best results are highlighted in bold. Fig. 5 illustrates the corresponding Micro-F1 trends.

Table 4: Parameter sensitivity results (Micro-F1%) of rank across datasets with different K settings.

Rank	SanWen			DuIE			ACE 2005			FinRE		
	K = 8	K = 16	K = 32	K = 8	K = 16	K = 32	K = 8	K = 16	K = 32	K = 8	K = 16	K = 32
1	**59.87**	**62.93**	64.12	**71.53**	73.92	75.66	70.22	**73.38**	74.67	**42.73**	**43.86**	**46.26**
2	59.62	62.88	**64.31**	71.45	**74.06**	**75.88**	**70.26**	73.33	**75.16**	41.24	42.37	46.14
3	58.36	60.59	63.94	71.18	73.27	75.07	69.14	72.76	74.87	40.68	42.14	45.87

Table 4 and Fig. 5 show that increasing the rank does not consistently improve performance. Although rank = 2 yields slight gains in some cases, the improvement is marginal relative to the increased parameter cost (The number of learnable parameters for low-rank matrix doubles). When the rank is further increased to 3, performance generally degrades, likely due to optimization difficulty under limited supervision for more parameters. Therefore, we adopt rank = 1 in experiments for a better trade-off between effectiveness and parameter efficiency.

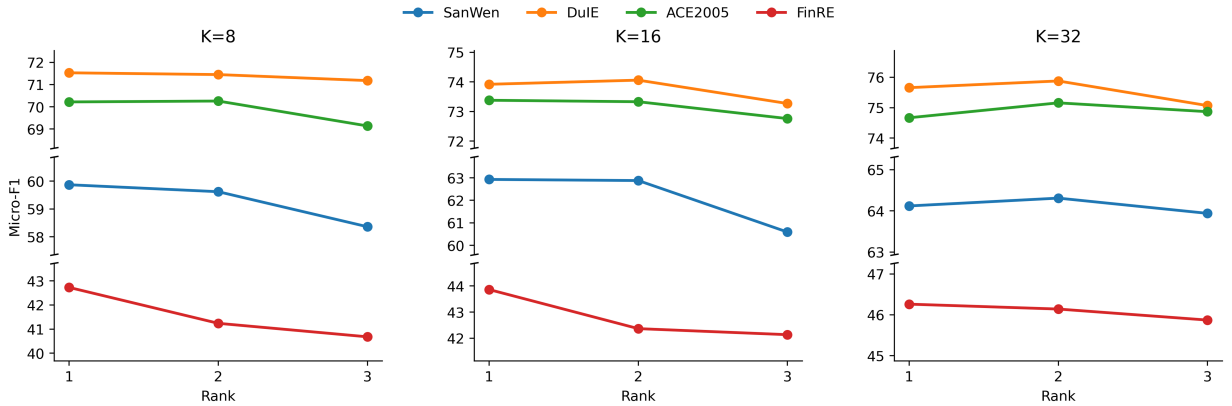

Figure 5: Micro-F1 scores with respect to rank under different K settings across datasets. Broken y-axes are applied for better visualization.

4.3.2 Effects of Learning Rates

To examine whether the performance of COPA(Qwen) is dominated by the asymmetric learning rates of domain-invariant and domain-specific prompt components, we conduct a sensitivity analysis on the learning-rate ratio. Specifically, we fix the domain-invariant learning rate η_{task} as 5e−4 and vary the domain-specific learning rate η_{domain} from 5e−4 to 1e−2, corresponding to ratios of 1×, 2×, 5×, 10×, and 20×. We also include a symmetric high-learning-rate setting where both learning rates are set to 5e−3. The results are shown in Table 5 and the best and second-best results are highlighted in bold and italics, respectively.

Table 5: Results (Micro-F1%) of sensitivity analysis of task/domain learning-rate ratios.

η_{task}	η_{domain}	Ratio	SanWen			DuIE			ACE 2005			FinRE		
			K = 8	K = 16	K = 32	K = 8	K = 16	K = 32	K = 8	K = 16	K = 32	K = 8	K = 16	K = 32
5e−4	5e−4	Symmetric	58.44	61.37	62.72	70.87	72.54	74.68	68.54	71.83	73.15	41.22	42.47	44.91
5e−4	1e−3	2×	58.74	61.93	63.18	71.25	72.89	75.15	68.94	72.37	73.72	41.36	42.84	45.68
5e−4	2.5e−3	5×	59.27	62.14	63.34	**72.14**	**74.23**	**76.04**	69.89	72.67	74.08	**42.87**	**44.26**	**46.97**
5e−4	5e−3	10×	**59.87**	**62.93**	**64.12**	71.53	73.92	75.66	**70.22**	**73.38**	**74.67**	42.73	43.86	46.26
5e−4	1e−2	20×	58.13	60.81	62.34	69.58	72.11	74.16	68.35	72.01	73.38	40.92	42.35	44.12
5e−3	5e−3	Symmetric	57.78	59.86	62.41	69.19	71.54	73.45	67.93	71.58	73.17	40.34	41.87	43.85

Table 5 shows that moderate learning-rate asymmetry is beneficial. The 5× and 10× ratios achieve the best average performance. In contrast, the symmetric settings and the overly large 20× ratio perform worse, indicating that domain-specific prompts require faster adaptation, but excessive domain-side updates may hurt stability. These results show that COPA is not dominated by a single 10× learning-rate choice; instead, it remains robust under moderate asymmetric ratios.

4.4 Ablation Experiments

4.4.1 Ablation Study on Model Components

We conduct ablation studies to evaluate the contributions of orthogonality constraint in Domain-Invariant Knowledge Acquisition stage and confidence-guided prompt adaptation strategy in Target Domain Prompt Adaptation stage, and denoted method without orthogonality constraint as "-OC", method without confidence-guided prompt adaptation strategy as "-CGPA", method without both fall back to MPT-GRP.

The ablation results in Table 6 demonstrate the effectiveness of both orthogonality constraint (OC) and confidence-guided prompt adaptation (CGPA). Removing OC leads to consistent but moderate performance drops, indicating its role in learning more discriminative domain-invariant representations. In contrast, removing CGPA results in more significant degradation, particularly under higher-shot settings, highlighting its importance in effective target-domain adaptation. When both components are removed (i.e., MPT-GRP), performance declines most substantially, suggesting that OC and CGPA are complementary and jointly contribute to the overall performance.

Table 6: Ablation study results(Micro-F1%) of orthogonality constraint and confidence-guided prompt adaptation strategy.

Method	SanWen			DuIE			ACE 2005			FinRE		
	K = 8	K = 16	K = 32	K = 8	K = 16	K = 32	K = 8	K = 16	K = 32	K = 8	K = 16	K = 32
COPA(Qwen)	59.87	62.93	64.12	71.53	73.92	75.66	70.22	73.38	74.67	42.73	43.86	46.26
-OC	59.24	62.34	63.56	70.82	73.31	75.14	69.85	73.06	74.26	42.32	43.22	45.82
	(−0.63)	(−0.59)	(−0.56)	(−0.71)	(−0.61)	(−0.52)	(−0.37)	(−0.32)	(−0.41)	(−0.41)	(−0.64)	(−0.44)
-CGPA	59.03	62.15	62.87	70.16	72.93	74.71	69.07	72.59	73.87	41.71	42.64	44.58
	(−0.84)	(−0.78)	(−1.25)	(−1.37)	(−0.99)	(−0.95)	(−1.15)	(−0.79)	(−0.80)	(−1.02)	(−1.22)	(−1.68)
MPT-GRP	58.63	61.83	62.64	69.83	72.18	74.27	68.84	72.04	73.58	41.26	42.11	44.18
	(−1.24)	(−1.10)	(−1.48)	(−1.70)	(−1.74)	(−1.39)	(−1.38)	(−1.34)	(−1.09)	(−1.47)	(−1.75)	(−2.08)

4.4.2 Ablation Study on Prompt Structure

To further validate the effectiveness of each component in the prompt structure, we conduct detailed ablation studies on four datasets using COPA based on Qwen. The Micro-F1 results are reported in Tables 7 and 8.

Table 7: Micro-F1 (%) of Ablation experiments on DuIE and SanWen based on COPA (Qwen).

No.	Inputs	Targets	DuIE			SanWen		
			K = 8	K = 16	K = 32	K = 8	K = 16	K = 32
1	$\boldsymbol{x}\ [T_h]\ e_h\ [T_t]\ e_t\ [R]\ \widetilde{P}$	$[T_h]\ t_h\ [T_t]\ t_t\ [R]\ \bar{\boldsymbol{r}}\ [E]$	70.28	73.62	75.17	59.31	62.27	63.74
2	$\boldsymbol{x}\ hard\_prompt^*$	$[T_h]\ t_h\ [T_t]\ t_t\ [R]\ \bar{\boldsymbol{r}}\ [E]$	69.34	72.14	73.95	58.17	60.74	62.04
3	$\boldsymbol{x}\ t_h\ e_h\ t_t\ e_t\ [R]\ \widetilde{P}$	$[R]\ \bar{\boldsymbol{r}}\ [E]$	68.06	70.21	72.43	57.23	58.19	60.86
4	$\boldsymbol{x}\ e_h\ e_t\ [R]\ \widetilde{P}$	$[R]\ \bar{\boldsymbol{r}}\ [E]$	67.84	70.02	72.15	56.64	57.41	59.82
5	$\boldsymbol{x}\ \widetilde{P}$	$[R]\ \bar{\boldsymbol{r}}\ [E]$	60.17	62.28	64.12	50.85	52.16	54.43
6	$\boldsymbol{x}$	$\bar{\boldsymbol{r}}$	56.83	60.53	61.87	49.48	51.17	52.86

Note: $^*hard\_prompt$: $the\ relation\ between\ [T_h]\ e_h\ and\ [T_t]\ e_t\ is\ [R]$.

Table 8: Micro-F1 (%) of Ablation experiments on FinRE and ACE 2005 based on COPA (Qwen).

No.	Inputs	Targets	FinRE			ACE 2005		
			K = 8	K = 16	K = 32	K = 8	K = 16	K = 32
1	$\boldsymbol{x}\ [T_h]\ e_h\ [T_t]\ e_t\ [R]\ \widetilde{P}$	$[T_h]\ t_h\ [T_t]\ t_t\ [R]\ \bar{\boldsymbol{r}}\ [E]$	41.26	43.21	45.18	69.84	72.18	73.58
2	$\boldsymbol{x}\ hard\_prompt^*$	$[T_h]\ t_h\ [T_t]\ t_t\ [R]\ \bar{\boldsymbol{r}}\ [E]$	40.61	41.73	44.05	69.17	71.35	72.61
3	$\boldsymbol{x}\ t_h\ e_h\ t_t\ e_t\ [R]\ \widetilde{P}$	$[R]\ \bar{\boldsymbol{r}}\ [E]$	40.18	40.77	43.22	68.25	70.43	72.13
4	$\boldsymbol{x}\ e_h\ e_t\ [R]\ \widetilde{P}$	$[R]\ \bar{\boldsymbol{r}}\ [E]$	39.66	39.92	41.39	67.23	68.61	69.88
5	$\boldsymbol{x}\ \widetilde{P}$	$[R]\ \bar{\boldsymbol{r}}\ [E]$	36.71	38.76	39.55	61.75	63.83	64.76
6	$\boldsymbol{x}$	$\bar{\boldsymbol{r}}$	36.46	37.53	39.62	61.06	62.38	64.12

Note: $^*hard\_prompt$: $the\ relation\ between\ [T_h]\ e_h\ and\ [T_t]\ e_t\ is\ [R]$.

Tables 7 and 8 present the ablation results of different prompt structure variants. The full prompt design (Row 1), which uses entity type sentinel tokens as input and generates both entity types and relation verbalizations, achieves the best performance, indicating that this formulation better aligns with autoregressive decoding and promotes deeper semantic understanding.

Replacing the soft prompt with a manually designed hard prompt (Row 2) leads to inferior results, highlighting the effectiveness of learnable prompts over manual template design. Providing entity types explicitly as input (Row 3) degrades performance, suggesting reduced reliance on contextual inference. Further removing entity type information (Row 4) and both entity and type information (Row 5) results in additional performance drops, with the latter showing the largest decline due to substantial information loss.

Finally, the sequence-to-sequence fine-tuning baseline (Row 6) performs worst, demonstrating the superiority of prompt-based learning for relation classification with LLMs.

4.5 Significance Tests

Since the gains of COPA (Qwen) over -OC and COPA (deepseek) are relatively small(0.3%–0.6%) and within one standard deviation, we conduct paired two-sided t-tests between COPA (Qwen) and these two comparisons and mark significant improvements at $p < 0.05$ in Table 9.

Table 9: Paired t-test results for COPA (Qwen) against -OC and COPA (deepseek).

Dataset	K	-OC			COPA (deepseek)		
		t-Stat	p-Value	Significance	t-Stat	p-Value	Significance
ACE_2005	8	1.71831307	0.09861590	FALSE	4.64061532	0.00010356	TRUE
	16	1.51581560	0.14262740	FALSE	−0.64057314	0.52786943	FALSE
	32	1.55459905	0.13313076	FALSE	2.05217273	0.05121412	FALSE
DuIE	8	3.42525275	0.00221559	TRUE	−0.75731806	0.45623085	FALSE
	16	2.90430464	0.00778103	TRUE	4.27007297	0.00026576	TRUE
	32	2.71074885	0.01220202	TRUE	6.91305943	3.77892739e−07	TRUE
FinRE	8	2.67564083	0.01322380	TRUE	5.79595334	5.63903147e−06	TRUE
	16	3.39200259	0.00240438	TRUE	6.09097186	2.72548448e−06	TRUE
	32	2.20740933	0.03709226	TRUE	8.55053224	9.52583202e−09	TRUE

(Continued)

Table 9 (continued)

Dataset	K	-OC			COPA (deepseek)		
		t-Stat	p-Value	Significance	t-Stat	p-Value	Significance
SanWen	8	2.24615915	0.03416676	TRUE	3.02145189	0.00589703	TRUE
	16	3.70584324	0.00110406	TRUE	4.37514553	0.00020347	TRUE
	32	2.60978015	0.01536079	TRUE	1.42840704	0.16606039	FALSE

The results show that COPA (Qwen) significantly outperforms COPA (deepseek) in most settings. However, the differences are not significant on ACE_2005 with $K = 16$ and $K = 32$, DuIE with $K = 8$, and SanWen with $K = 32$, indicating that the advantage of COPA (Qwen) over COPA (deepseek) is not uniform across all datasets and K settings. For the comparison between full COPA (Qwen) and -OC, the improvements are statistically significant on SanWen, DuIE, and FinRE, while none of the gains on ACE_2005 reach significance, suggesting that the contribution of the orthogonality constraint may be relatively moderate in some cases.

5 Conclusion

In this paper, we propose COPA, a confidence-guided orthogonality-constrained prompt adaptation framework for few-shot relation classification. By combining orthogonality-constrained prompt disentanglement with confidence-guided prompt adaptation, COPA effectively captures transferable relational knowledge while enabling stable target-domain adaptation under limited supervision. Experimental results on four public datasets demonstrate that COPA consistently outperforms strong baselines in terms of Micro-F1, with lower variance. In future work, we will explore extending COPA to cross-lingual relation classification for low-resource cross-domain and cross-lingual transfer.

Acknowledgement: None.

Funding Statement: This research was funded by Key Program of the National Natural Science Foundation of China (No. U23A20305) and Zhongyuan Scholars Project (No. 254000510007).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Shunran Duan, Meijuan Yin, Xiangyang Luo; data collection: Shunran Duan; analysis and interpretation of results: Shunran Duan; baseline methods reproduction: Shunran Duan, Lunchong Cui, Chenyu Wang; draft manuscript preparation: Shunran Duan; visualization: Shunran Duan. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The involved datasets that support the findings of this study include: SKE (https://aistudio.baidu.com/datasetdetail/18801), FR2KG [23] and IPRE [24], SanWen [25], DuIE [26], ACE 2005 Chinese Corpus (https://catalog.ldc.upenn.edu/LDC2006T06) and FinRE [27].

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A Dataset Characterization

To better substantiate the cross-domain setting, we further characterize the source and target datasets from three aspects: text register, relevant statistics, and relation-set overlap. Specifically, we report text

register and domain, the number of relations (#Rel.), average sentence length (Avg. Sent. Len.), average entity length (Avg. Ent. Len.), average entity distance (Ave. Ent. Dis.), normalized relation overlap (normalized) Overlap $(S, T) = \frac{|\mathcal{R}_S \cap \mathcal{R}_T|}{|\mathcal{R}_T|}$ and Jaccard similarity coefficient (Jaccard) Jaccard$(S, T) = \frac{|\mathcal{R}_S \cap \mathcal{R}_T|}{|\mathcal{R}_S \cup \mathcal{R}_T|}$ in Table A1. For FR2KG, since it is a document-level relation extraction dataset, 55.88 denotes the average sentence length, while the value in parentheses, 2213.83, denotes the average document length. Entity distance is defined as the absolute character distance between the start positions of the head and tail entity mentions.

Table A1: Dataset statistics and relation overlap comparison.

Dataset	Split Role	Register/Domain	#Rel.	Avg. Sent. Len.	Avg. Ent. Len.	Avg. Ent. Dis.	Normalized	Jaccard
SKE*	Source	Encyclopedia/General	50	54.63	4.50	21.82	—	—
FR2KG	Source	Company Research Reports/Financial	19	55.88 (2213.83)	8.54	724.68	—	—
IPRE	Source	Inter-personal Relation-ship/General	34	53.94	2.62	14.99	—	—
SanWen	Target	Chinese Litera-ture/Literary	9	62.56	3.02	15.27	0.030928	0.029126
DuIE	Target	News, Entertainment, etc./General	49	54.63	4.48	21.74	0.505155	0.500000
ACE 2005	Target	Newswire, Broadcast News, Weblog/News	18	28.24	4.45	7.40	0.061856	0.053571
FinRE	Target	Economic Text/Financial	44	54.80	4.78	15.15	0.020619	0.014286

Note: *SKE refers to Schema-based Knowledge Extraction dataset (https://aistudio.baidu.com/datasetdetail/18801).

References

1. dos Santos C, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural networks. In: Zong C, Strube M, editors. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg, PA, USA: ACL; 2015. p. 626–34.
2. Wang L, Cao Z, de Melo G, Liu Z. Relation classification via multi-level attention CNNs. In: Erk K, Smith NA, editors. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL; 2016. p. 1298–307.
3. Yin G, Wang X, Zhang H, Wang J. Cost-effective CNNs-based prototypical networks for few-shot relation classification across domains. Knowl Based Syst. 2022;253(1):109470. doi:10.1016/j.knosys.2022.109470.
4. Zhang S, Zheng D, Hu X, Yang M. Bidirectional long short-term memory networks for relation classification. In: Zhao H, editor. Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation; 2015 Oct 30–Nov 1; Shanghai, China. p. 73–8.
5. Wang Z, Yang B. Attention-based bidirectional long short-term memory networks for relation classification using knowledge distillation from BERT. In: 2020 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). Piscataway, NJ, USA: IEEE; 2020. p. 562–8.
6. Cai R, Zhang X, Wang H. Bidirectional recurrent convolutional neural network for relation classification. In: Erk K, Smith NA, editors. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL; 2016. p. 756–65.

7. Xing Z, Ye Y, Song R, Teng Y, Li Z, Liu J. Sample feature enhancement model based on heterogeneous graph representation learning for few-shot relation classification. Inf Sci. 2025;690:121583. doi:10.1016/j.ins.2024.121583.
8. Wei C, Li J, Wang Z, Wan S, Guo M. Graph convolutional networks embedding textual structure information for relation extraction. Comput, Mater Continua. 2024;79(2):3299–314. doi:10.32604/cmc.2024.047811.
9. Yang R, Chen Y, Yan J, Qin Y. A graph with adaptive adjacency matrix for relation extraction. Comput, Mater Continua. 2024;80(3):4129–47. doi:10.32604/cmc.2024.051675.
10. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019. Stroudsburg, PA, USA: ACL; 2019. p. 4171–86. doi:10.18653/v1/n19-1423.
11. Wu S, He Y. Enriching pre-trained language model with entity information for relation classification. In: Zhu W, Tao D, Cheng X, Cui P, Rundensteiner EA, Carmel D, et al., editors. Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019; 2019 Nov 3–7; Beijing, China. p. 2361–4. doi:10.1145/3357384.3358119.
12. Han J, Zhao S, Cheng B, Ma S, Lu W. Generative prompt tuning for relation classification. In: Goldberg Y, Kozareva Z, Zhang Y, editors. Findings of the Association for Computational Linguistics: EMNLP 2022. Stroudsburg, PA, USA: ACL; 2022. p. 3170–85.
13. Jiang Y, Li J, Chen H. Relation classification via bidirectional prompt learning with data augmentation by large language model. In: Calzolari N, Kan MY, Hoste V, Lenci A, Sakti S, Xue N, editors. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italia: ELRA and ICCL; 2024. p. 13885–97.
14. Zhang W, Song X, Feng Z, Xu T, Wu X. LabelPrompt: effective prompt-based learning for relation classification. In: Nguyen V, Lin H, editors. Asian Conference on Machine Learning. Vol. 260 of Proceedings of Machine Learning Research. Cambridge, MA, USA: PMLR; 2024. p. 1304–19.
15. Wang Z, Panda R, Karlinsky L, Feris R, Sun H, Kim Y. Multitask prompt tuning enables parameter-efficient transfer learning. In: The Eleventh International Conference on Learning Representations, ICLR 2023; 2023 May 1–5; Kigali, Rwanda.
16. Dong B, Yao Y, Xie R, Gao T, Han X, Liu Z, et al. Meta-information guided meta-learning for few-shot relation classification. In: Scott D, Bel N, Zong C, editors. Proceedings of the 28th International Conference on Computational Linguistics. Stroudsburg, PA, USA: ACL; 2020. p. 1594–605.
17. Geng X, Chen X, Zhu KQ, Shen L, Zhao Y. MICK: a meta-learning framework for few-shot relation classification with small training data. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 415–24. doi:10.1145/3340531.3411858.
18. Obamuyide A, Vlachos A. Model-agnostic meta-learning for relation classification with limited supervision. In: Korhonen A, Traum D, Màrquez L, editors. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL; 2019. p. 5873–9.
19. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv. 2023;55(9):195. doi:10.1145/3560815.
20. Xie S, Pan Q, Wang X, Luo X, Sugumaran V. Combining prompt learning with contextual semantics for inductive relation prediction. Expert Syst Appl. 2024;238(Part D):121669. doi:10.1016/j.eswa.2023.121669.
21. Moslemnejad S, Reed C. Prompt templates for argument relation classification using frame semantic parsing. Knowl Inf Syst. 2025;67(10):9189–219. doi:10.1007/s10115-025-02500-8.
22. Litrico M, Bue AD, Morerio P. Guiding Pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023. Piscataway, NJ, USA: IEEE; 2023. p. 7640–50. doi:10.1109/CVPR52729.2023.00738.
23. Wang W, Xu Y, Du C, Chen Y, Wang Y, Wen H. Data set and evaluation of automated construction of financial knowledge graph. Data Intell. 2021;3(3):418–43. doi:10.1162/dint_a_00108.

24. Wang H, He Z, Ma J, Chen W, Zhang M. IPRE: a dataset for inter-personal relationship extraction. In: Tang J, Kan M, Zhao D, Li S, Zan H, editors. Natural Language Processing and Chinese Computing-8th CCF International Conference, NLPCC 2019. Cham, Switzerland: Springer; 2019. p. 103–15. doi:10.1007/978-3-030-32236-6_9.
25. Xu J, Wen J, Sun X, Su Q. A discourse-level named entity recognition and relation extraction dataset for Chinese literature text. arXiv:1711.07010. 2017.
26. Li S, He W, Shi Y, Jiang W, Liang H, Jiang Y, et al. DuIE: a large-scale chinese dataset for information extraction. In: Tang J, Kan M, Zhao D, Li S, Zan H, editors. Natural Language Processing and Chinese Computing-8th CCF International Conference, NLPCC 2019. Cham, Switzerland: Springer; 2019. p. 791–800. doi:10.1007/978-3-030-32236-6_72.
27. Li Z, Ding N, Liu Z, Zheng H, Shen Y. Chinese relation extraction with multi-grained information and external linguistic knowledge. In: Korhonen A, Traum D, Màrquez L, editors. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL; 2019. p. 4377–86.
28. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 2019 Dec 8–14; Vancouver, BC, Canada. p. 8024–35.
29. Yang A, Yang B, Hui B, Zheng B, Yu B, Zhou C, et al. Qwen2 technical report. arXiv:2407.10671. 2024.
30. DeepSeek-AI, Guo D, Yang D, Zhang H, Song J, Wang P, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. arXiv:2501.12948. 2025.
31. Han J, Cheng B, Wan Z, Lu W. Towards hard few-shot relation classification. IEEE Trans Knowl Data Eng. 2023;35(9):9476–89. doi:10.1109/TKDE.2023.3240851.
32. Chen Y, Harbecke D, Hennig L. Multilingual relation classification via efficient and effective prompting. In: Goldberg Y, Kozareva Z, Zhang Y, editors. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022. Stroudsburg, PA, USA: ACL; 2022. p. 1059–75. doi:10.18653/v1/2022.emnlp-main.69.
33. Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. In: Moens M, Huang X, Specia L, Yih SW, editors. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021. Stroudsburg, PA, USA: ACL; 2021. p. 3045–59. doi:10.18653/v1/2021.emnlp-main.243.
34. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. In: The Tenth International Conference on Learning Representations, ICLR 2022; 2022 Apr 25–29; Virtual Event.