Bridging Generative and Discriminative Learning: Few-Shot Relation Extraction via Two-Stage Knowledge-Guided Pre-training

Quanjiang Guo[1] , **Jinchuan Zhang**[1] , **Sijie Wang**[2] , **Ling Tian**[1] , **Zhao Kang**[1*] ,
Bin Yan[3] , **Weidong Xiao**[4]

[1] University of Electronic Science and Technology of China, Chengdu, China
[2] Nanyang Technological University, Singapore
[3] Information Engineering University, Zhengzhou, China
[4] National University of Defense Technology, Changsha, China
guochance1999@163.com, jinchuanz@std.uestc.edu.cn, wang1679@e.ntu.edu.sg,
{lingtian, zkang}@uestc.edu.cn, tom.yan@gmail.com, wdxiao@nudt.edu.cn

Abstract

Few-Shot Relation Extraction (FSRE) remains a challenging task due to the scarcity of annotated data and the limited generalization capabilities of existing models. Although large language models (LLMs) have shown potential in FSRE through in-context learning, their general-purpose training objectives often result in suboptimal performance for task-specific relation extraction. To overcome these challenges, we propose TKRE (Two-Stage Knowledge-Guided Pre-training for Relation Extraction), a novel framework that synergistically integrates LLMs with traditional relation extraction models, bridging generative and discriminative learning paradigms. TKRE introduces two key innovations: (1) leveraging LLMs to generate explanation-driven knowledge and schema-constrained synthetic data, addressing the issue of data scarcity; and (2) a two-stage pre-training strategy combining Masked Span Language Modeling (MSLM) and Span-Level Contrastive Learning (SCL) to enhance relational reasoning and generalization. Together, these components enable TKRE to effectively handle FSRE tasks. Comprehensive experiments on benchmark datasets demonstrate the efficacy of TKRE, achieving new state-of-the-art performance in FSRE and underscoring its potential for broader application in low-resource scenarios. The code and data are released on https://github.com/UESTC-GQJ/TKRE.

1 Introduction

Relation Extraction (RE) focuses on identifying semantic relationships between entities in unstructured text, serving as a cornerstone for knowledge graph construction and event analysis. Although pre-trained language models (PLMs) [Wei *et al.*, 2020; Li *et al.*, 2022; Wang *et al.*, 2023a; Zhu *et al.*, 2024] have significantly advanced RE performance, their success
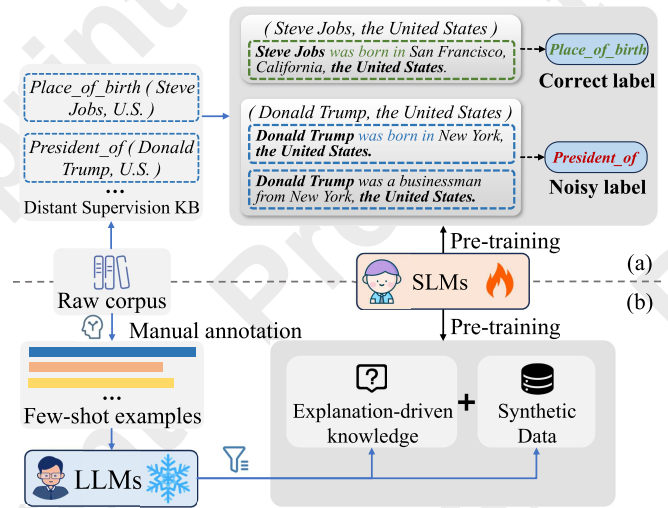
---

*Corresponding Author

Figure 1: (a) Distant supervision is an automated and efficient approach for generating labeled data, but its effectiveness depends on the quality of the knowledge base. Challenges such as noisy labels and context loss are common issues with distant supervision. (b) In contrast, our approach leverages explanation-driven knowledge and synthetic data automatically generated by LLMs, which significantly enhance the performance of small language models (SLMs).

depends on the availability of large-scale annotated datasets, a requirement that is often unattainable in real-world applications such as rare disease relationship mining and financial event detection. These scenarios are particularly challenging due to the severe scarcity of annotated examples. As a result, Few-Shot Relation Extraction (FSRE) has emerged as a critical research area, necessitating the development of models capable of generalizing from only a handful of labeled examples per relation type.

Existing FSRE methodologies face three critical challenges. **Data Scarcity**: Traditional approaches rely on extensive annotations to distinguish between relations. While distant supervision [Vrandečić and Krötzsch, 2014] generates labeled data aligned with knowledge bases, it introduces noisy labels and domain mismatches [Soares *et al.*, 2019;

Peng *et al.*, 2020; Qin *et al.*, 2021], as shown in Figure 1(a). **Limited Generalization**: Models trained on in-distribution data struggle to adapt to cross-domain or fine-grained relations. For example, discriminative models frequently misclassify semantically similar relations, such as confusing "Place_of_birth" with "Place_of_death", with error rates exceeding 30% [Alt *et al.*, 2020]. **Generative-Discriminative Disconnect**: Large language models (LLMs) like GPT-4 [Achiam *et al.*, 2023] show potential to generate synthetic data, but treat RE as a black-box text generation task [Li *et al.*, 2023; Xu *et al.*, 2023b]. This results in a misalignment between the outputs of LLMs and the structured learning objectives required by FSRE models.

To address these challenges, we propose Two-stage Knowledge-guided Relational Extraction (TKRE), a novel framework that integrates generative and discriminative paradigms through structured knowledge infusion. We first leverage large language models (LLMs) to generate relational logic explanations (e.g., "Subject X exhibits relation R with Object Y because ..."), explicitly capturing fine-grained semantic dependencies between entities. These explanations create interpretable reasoning pathways that guide downstream model training. Then, we employ constrained decoding strategies to generate schema-compliant training instances by enforcing entity type compatibility and relation-specific lexico-syntactic patterns. This dual approach ensures that the synthetic data preserve domain-specific relational semantics while mitigating distributional changes, as illustrated in Figure 1(b).

Finally, we introduce a two-stage knowledge-guided pre-training strategy that combines Masked Span Language Modeling (MSLM) and Span-level Contrastive Learning (SCL). MSLM focuses on selectively masking relation-indicative spans and optimizing their reconstruction, enhancing the model's ability to capture contextual dependencies and infer relational semantics. Complementing this, SCL operates within a contrastive learning framework, where positive spans (semantically aligned with target relations) are contrasted with negative spans (relationally discordant but contextually plausible). This approach sharpens the model's discriminative boundaries for entity interaction modeling, fostering a more precise understanding of relational structures.

As illustrated in Figure 2, TKRE establishes a self-reinforcing knowledge integration cycle: (a) relational explanations guide the generation of synthetic data, ensuring semantic consistency; (b) the synthetic data is leveraged during pre-training to incorporate generative knowledge; and (c) pre-training objectives explicitly align the representations of synthetic and golden data. This closed-loop architecture facilitates the iterative refinement of relational understanding through knowledge distillation and representation alignment, effectively bridging the gap between generative and discriminative paradigms.

In summary, our contributions are as follows:

- We leverage LLMs to automatically generate explanation-driven knowledge and structured synthetic data, addressing data scarcity while enhancing the traditional RE model's comprehension of underlying relational logic.

- We introduce a novel two-stage knowledge-guided pre-training framework that seamlessly integrates generative knowledge with discriminative objectives, enabling models to effectively learn from sparse annotations and noisy synthetic data simultaneously.

- Extensive experiments on four benchmark datasets demonstrate that TKRE significantly improves the performance of traditional FSRE models, achieving F1 gains of 7.8% and 5.0% over strong baselines (TYP Marker, GenPT) and outperforming both pure LLM-based and hybrid methods in few-shot settings.

2 Related Work

Few-Shot Relation Extraction with Traditional Models.
Traditional FSRE methods predominantly rely on discriminative learning to optimize $P(y|x)$, but they face significant challenges under low-resource conditions. Approaches such as [Peng *et al.*, 2020], [Zhou and Chen, 2022] and [Zhang and Kang, 2024] enhance intra-class representation consistency but struggle to generalize to fine-grained relations, due to their reliance on rigid in-distribution assumptions. Prompt-based methods such as [Han *et al.*, 2022] and PTRE [Zhang *et al.*, 2024b]) reformulate classification as an infilling task, yet require precise semantic alignment and fail to scale with extreme data scarcity. Distant supervision [Mintz *et al.*, 2009] alleviates annotation costs via knowledge base alignment, and many recent works investigate how to improve learning with such distant data [Soares *et al.*, 2019; Peng *et al.*, 2020; Qin *et al.*, 2021]. Despite its benefits, distant supervision introduces noisy labels and domain mismatches, which significantly undermine model robustness.

These limitations arise from a common bottleneck: *discriminative models excel at distinguishing predefined classes but lack the ability to reason over unseen relational logic or contextual dependencies*. This deficiency is particularly critical in FSRE, where success depends on capturing subtle semantic distinctions and adapting to cross-domain variations. To address these challenges, our proposed TKRE framework integrates generative knowledge to bridge this gap, enabling more robust relational reasoning and improved adaptability.

Large Language Models for FSRE. The emergence of LLMs like GPT-4 [Achiam *et al.*, 2023] and LLaMA-2 [Touvron *et al.*, 2023] has generated significant interest in leveraging their **in-context learning** capabilities for FSRE. Early efforts [Xu *et al.*, 2023a; Li *et al.*, 2023; Xu *et al.*, 2023b] query LLMs using task descriptions and demonstrations, achieving only marginal improvements over traditional baselines. These approaches, however, treat LLMs as black-box generators and fail to align their **generative knowledge** (e.g., contextual reasoning and dependency modeling) with the structured discriminative objectives of RE models. As a result, the generated outputs often lack task-specific structural awareness, leading to suboptimal performance on fine-grained relations [Li *et al.*, 2023]. Furthermore, directly fine-tuning LLMs is computationally prohibitive [Yang *et al.*, 2024], and their zero-shot performance is constrained by the complexity of prompt design [Whitehouse *et al.*, 2023; Guo *et al.*, 2025]. These limitations underscore the urgent
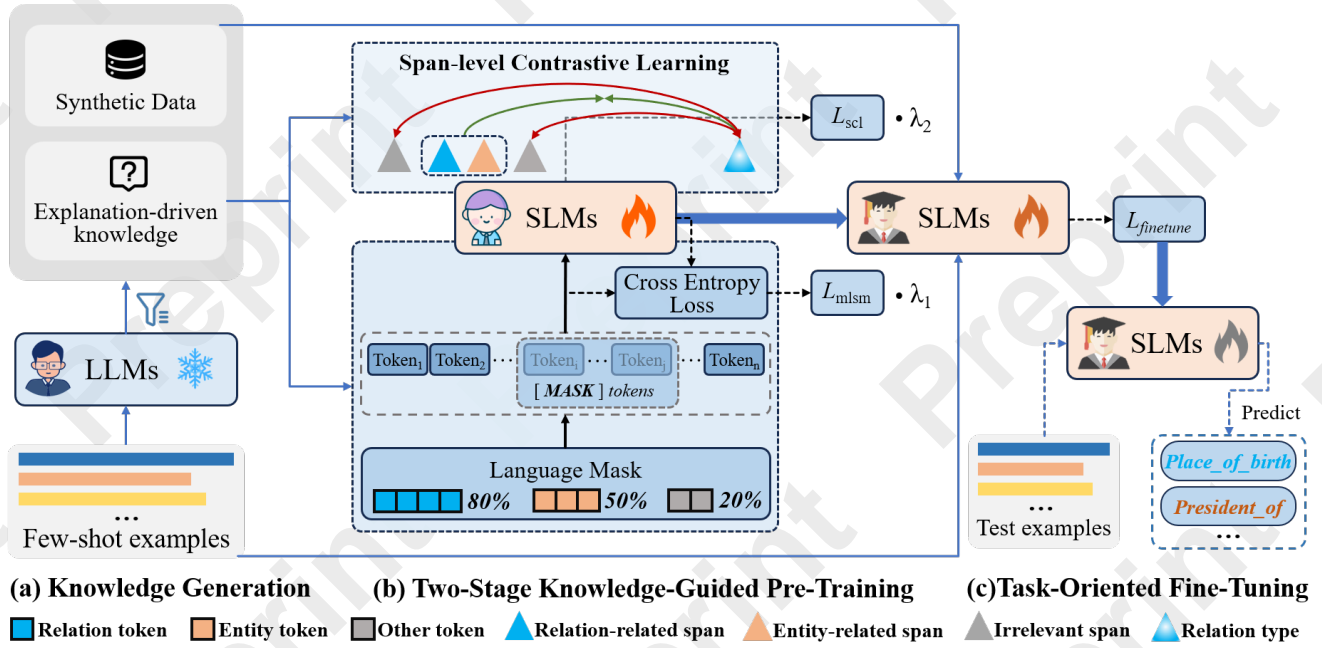
Figure 2: Overall framework of the proposed TKRE. (a) Leveraging LLMs to generate explanation-driven knowledge and schema-constrained synthetic data. (b) Implementing a two-stage knowledge-guided pre-training framework to enhance the model's ability to understand relational structures and contextual dependencies. (c) Performing task-oriented fine-tuning on few-shot examples and synthetic data.

need for frameworks that effectively bridge generative and discriminative learning paradigms to unlock the full potential of LLMs for FSRE.

Hybrid Methods: Combining LLMs with Traditional Models. Recent hybrid approaches attempt to address data scarcity by leveraging LLMs to augment training data for smaller RE models. Editing-based methods [Wei and Zou, 2019; Jiao *et al.*, 2020] replace tokens with synonyms or back-translated sentences but often fail to maintain relational consistency. Generative methods [Anaby-Tavor *et al.*, 2020; Schick and Schütze, 2021] use LLM prompting to synthesize labeled instances, yet their outputs frequently suffer from noise and lack diversity in structured RE tasks. Semi-supervised techniques [Vu *et al.*, 2021] combine synthetic and golden data using confidence filtering or auxiliary tasks but treat synthetic data as static inputs rather than dynamic knowledge carriers. For instance, [Xu *et al.*, 2023a] introduces a two-stage self-training framework but does not explicitly model entity-span interactions. Similarly, [Wang *et al.*, 2023b] generates label-conditioned data but neglects alignment between synthetic examples and relational reasoning. These methods fail to fully exploit the potential synergy between generative and discriminative paradigms, limiting their ability to transfer LLM-derived knowledge into task-specific representations effectively.

3 Methodology

In this section, we present the detailed modules of our TKRE: explanation-driven knowledge generation, two-stage knowledge-guided pre-training, and task-oriented fine-tuning. Figure 2 depicts the overall framework.

3.1 Task Formulation

Let $\mathcal{R} = \{r_1, r_2, \ldots, r_{|\mathcal{R}|}\}$ represent the set of $|\mathcal{R}|$ relation types in the dataset (e.g., TACRED contains $|\mathcal{R}| = 42$). Given a sentence x with two marked entities e_{sub} and e_{obj}, the goal is to predict the relation $r \in \mathcal{R}$ between these entities. In the few-shot setting, only K labeled examples $\mathcal{D}_{\text{train}} = \{(x, e_{\text{sub}}, e_{\text{obj}}, r)_i\}_{i=1}^{K \times |\mathcal{R}|}$ per relation type are provided for training, where K is a small constant (e.g., $K = 8$).

3.2 Data Preparation

To address the challenges of data scarcity and limited generalization, we employ LLMs to generate **knowledge explanations** and **synthetic examples** for each relation type, followed by multi-task pre-training and fine-tuning.

Explanation-Driven Knowledge Generation

To further improve domain adaptation and enhance the task relevance of the pre-training strategy, we utilize LLMs to construct task-oriented, generated knowledge. LLMs, trained on diverse corpora that likely encompass domains relevant to NER tasks, provide a rich source of contextual and domain-specific information. However, directly fine-tuning LLMs is computationally expensive and resource-intensive, making it impractical for many downstream tasks.

To address this, we design an intuitive instruction to guide the LLM in explaining why a given text span should be recognized as an entity, thereby generating a task-oriented corpus. For a sentence $x \in \mathcal{D}_{\text{train}}$, the golden relation type r, and an entity pair $(e_{\text{sub}}, e_{\text{obj}}) \in x$, the instruction X is constructed as described in Figure 3.

> ### Instruction:
> Take the text $< x >$ and give an explanation of why the subject $< e_{\text{sub}} >$ has the relation $< r >$ with the object $< e_{\text{obj}} >$.

Figure 3: Instruction of explanation-driven knowledge generation.

Given this instruction X, we denote LLM as f_{LLM}. The generated sequence regarding the explanation of relation $< y >$ is denoted as $Y = f_{\text{LLM}}(X) = \{y_i\}_{i=1}^{L}$, where L is defined as the length of the generated sequence, i.e., the number of tokens in Y. For the classic auto-regressive generative model, the generated sequence Y is predicted by the following conditional probability:

$$p(Y \mid X) = \prod_{t=1}^{L} p(y_t \mid X, y_{<t}), \tag{1}$$

where $y_{<t}$ represents the tokens before y_t.

Consequently, we can obtain several sentences of a relation extraction flow by reasoning in the raw textual context $< x >$, such as the grey part in Figure 1. Then, with respect to all relation types in $\mathcal{D}_{\text{train}}$, we employ the frozen LLM $\mathcal{M}$ to obtain K explanations cluster of each relation type. Formally,

$$\{\mathbf{Y}_i\} = \mathcal{M}_{\text{Frozen}}(X_{r_i}), r_i \in \mathcal{R} \tag{2}$$

where X_{r_i} denotes the instruction X with the corresponding slots of relation r_i. Following [Qin *et al.*, 2021], we build the explanation-driven knowledge corpus $\mathcal{C}$ from the labeled raw texts in $\mathcal{D}_{\text{train}}$.

Schema-Constrained Data Generation
In this section, we aim to perform data augmentation using LLMs to enrich the training data for RE. Inspired by [Xu *et al.*, 2023b], we design prompts that inform the LLM about the essential components of an RE training sample: context text, subject entity, object entity, subject entity type, object entity type, and the relation. The LLM is then guided to generate additional pseudo RE samples based on this structure. To ensure the augmented data aligns with the desired format, we apply predefined rules, such as regular expressions, to process and transform the LLM-generated outputs. The specific instruction provided to the LLM is illustrated in Figure 4.

> ### Instruction:
> One sample consists of a relation, a context, a pair of head and tail entities… Generate more samples like above for the relation:

Figure 4: Instruction of schema-constrained synthetic data generation.

3.3 Two-Stage Knowledge-guided Pre-training
In this section, we present a *two-stage knowledge-guided pre-training framework* that integrates **Masked Span Language Modeling (MSLM)** and **Span-level Contrastive Learning** tasks. This framework is designed to equip traditional FSRE models with the ability to effectively capture contextual relationships between entities while simultaneously learning robust span-level representations. By combining these complementary tasks, the pre-training strategy enhances the model's capability to reason over entity relationships and adapt to fine-grained relational distinctions.

Masked Span Language Modeling
MSLM is an extension of traditional Masked Language Modeling (MLM) [Devlin *et al.*, 2019], where instead of masking individual tokens, we mask entire spans (sequences of tokens). Inspired by the better learning ability of span masking [Liu *et al.*, 2021] and recent advances in span-based masking methods [Zhang *et al.*, 2024a], we adopt this strategy to improve the model's ability to capture long-range dependencies and complex relations between entities.

Let $x = \langle x_1, x_2, \ldots, x_n \rangle$ represent a given sentence, where $x_i \in \mathcal{C}$ is a token from the explanation-driven knowledge corpus $\mathcal{C}$, and n is the sentence length. In MSLM, we mask a continuous span $\mathcal{S} = \langle x_i, x_{i+1}, \ldots, x_j \rangle$ by replacing it with a special sentinel token, such as $[MASK]$. The task is to predict the masked span $\mathcal{S}$ based on the surrounding context.

Given a masking strategy with different probabilities for different types of spans:

- 80% probability to mask relation spans,
- 50% probability to mask subject/object entity spans,
- 20% probability to mask other words.

We define a custom probability distribution for masking tokens based on their linguistic role in the sentence. The mask distribution $M = \langle m_1, m_2, \ldots, m_n \rangle$ is determined by the specific type of each token x_i. For each token x_i, the masking probability p_i is defined according to the following rules:

$$p_i = \begin{cases} 0.8 & \text{if } x_i \text{ belongs to relation span,} \\ 0.5 & \text{if } x_i \text{ belongs to subject/object entity span,} \\ 0.2 & \text{if } x_i \text{ belongs to other type of words.} \end{cases} \tag{3}$$

This strategy allows for a more targeted masking process, where different parts of the sentence are masked with different probabilities based on their syntactic roles. We model this process using a Bernoulli distribution for each token, where the probability of masking a token x_i is determined by its type:

$$P(M = m_i \mid p_i) = p_i^{m_i}(1 - p_i)^{1-m_i} \tag{4}$$

where $m_i \in \{0, 1\}$ indicates whether the token x_i is masked ($m_i = 1$) or not ($m_i = 0$).

The model is trained to predict the masked span by minimizing the *cross-entropy loss*:

$$L_{\text{mslm}} = -\frac{1}{\gamma} \sum_{i=1}^{\gamma} \log P(y_i \mid x_i) \tag{5}$$

where γ represents the maximum length of input sequence, and y_i denotes the predicted output for each masked token x_i. The loss function trains the model to recover the masked spans, encouraging it to learn rich contextual information for span prediction.

Span-level Contrastive Learning

In addition to MSLM, we apply **Span-level Contrastive Learning** to further refine the traditional model's ability to differentiate between different spans in the context of specific relations. Contrastive learning encourages the model to learn useful embeddings that can distinguish positive spans (relevant to the relation) from negative spans (irrelevant to the relation).

For each sentence x, the model learns to predict both the *positive span* s_p (e.g., "was born in") and the *negative span* s_n (e.g., "New York"). Let h_a, h_p, h_n represent the embeddings of the anchor (relation type), positive, and negative spans, respectively. We optimize the following contrastive loss:

$$L_{\text{scl}} = -\log \frac{\exp(\text{sim}(h_a, h_p)/\tau)}{\exp(\text{sim}(h_a, h_p)/\tau) + \exp(\text{sim}(h_a, h_n)/\tau)}$$
(6)

where $\text{sim}(h_a, h_x)$ is the cosine similarity between the anchor h_a and the span h_x, and τ is the temperature parameter controlling the sharpness of the similarity function.

This loss function encourages the model to bring positive spans closer to their respective anchor contexts and push negative spans further away, improving the model's span-level understanding.

Multi-task Training Objective

The total loss for pre-training is a weighted combination of MSLM and SCL losses:

$$L_{\text{pre-train}} = \lambda_1 L_{\text{mslm}} + \lambda_2 L_{\text{scl}},$$
(7)

where λ_1 and λ_2 control the relative importance of the two tasks.

3.4 Task-Oriented Fine-tuning

After pre-training, the model is fine-tuned on the specific few-shot relation extraction task. Fine-tuning is performed using both the original small-shot examples and the synthetic samples generated in the previous stages. The goal of fine-tuning is to adapt the pre-trained model to the downstream task of relation classification, which involves predicting the correct relation between two entities in a given sentence.

For each input sentence x with entities e_{sub} and e_{obj}, the model is fine-tuned to predict the correct relation $r \in \mathcal{R}$ from the set of relations. The input to the model consists of the sentence x_i, along with the two entities e_{sub_i} and e_{obj_i}. The output is the predicted relation label r_i. The final relation prediction is made by passing the output of the encoder through a fully connected classification layer:

$$P(r_i|x_i, e_{\text{sub}_i}, e_{\text{obj}_i}; \theta) = \text{softmax}(W \cdot h(x_i, e_{\text{sub}_i}, e_{\text{obj}_i}) + b)$$
(8)

where $h(x_i, e_{\text{sub}_i}, e_{\text{obj}_i})$ is the representation of the sentence and entities from the pre-trained model, and W and b are the learned weights and bias of the classifier. The classification loss is computed as:

$$L_{\text{finetune}} = -\sum_{i=1}^{N} \log P(r_i|x_i, e_{\text{sub}_i}, e_{\text{obj}_i}; \theta)$$
(9)

where $N \approx 2 \times K \times |\mathcal{R}|$ is the total number of samples in the training set, because the number of samples for some relation types may be less than K. And $P(r_i|x_i, e_{\text{sub}_i}, e_{\text{obj}_i}; \theta)$ is the predicted probability for the correct relation r_i.

4 Experiments

4.1 Experimental Design

Dataset	#Train				#Val	#Test	#Rel
	8-Shot	16-Shot	32-Shot	All			
SemEval	148	295	589	6,507	1,439	2,717	19
TACRED	334	662	1,305	68,124	22,631	15,509	42
TACREV	334	662	1,305	68,124	22,631	15,509	42
Re-TACRED	318	630	1,238	58,465	19,584	13,418	40

Table 1: Statistics of our experimental datasets. K-shot means sampling K instances from each relation type. For relation types with fewer than K instances, we sample all available data. All refers to the complete training dataset.

Datasets. **SemEval 2010 Task 8** [Hendrickx *et al.*, 2010] is a widely used testbed for relation extraction. The schema is targeted at semantic relations between pairs of nominals, which requires a certain level of abstractive capabilities. **TACRED** [Zhang *et al.*, 2017] is a large-scale dataset annotated using Amazon Mechanical Turk crowdsourcing. It was initially created for the TAC knowledge base population and mainly covers common relations between people, organizations, and locations based on the TAC KBP scheme. **TACREV** [Alt *et al.*, 2020] is a label-corrected version of the TACRED dataset, which motivates from the unresolved challenging cases in the original TACRED dataset. **Re-TACRED** [Stoica *et al.*, 2021] further conducted a more comprehensive analysis and re-annotated the entire dataset. Besides, it made alternations to the schema to make it more clear and intuitive, which greatly improved the dataset quality. Following the strategy adopted by [Gao *et al.*, 2021; Xu *et al.*, 2023b], for each relation type, we randomly sample K instances per relation (K-shot) for the training phase. The whole validation set and test set are preserved to ensure the effectiveness of the evaluation. The statistics of datasets are presented in Table 1.

Baselines. We compare our proposed TKRE with *Traditional RE*, *LLM-based* and *Hybrid* Methods. The *Traditional RE* methods include **CP** [Peng *et al.*, 2020], **TYP Marker** [Zhou and Chen, 2022], **Knowprompt** [Chen *et al.*, 2022], **GenPT** [Han *et al.*, 2022], and **PTRE** [Zhang *et al.*, 2024b]. The *LLM-based* methods include **LLaMA-2** [Touvron *et al.*, 2023], **GPT-3.5** [Ouyang *et al.*, 2022], and **GPT-4** [Achiam *et al.*, 2023]. The *Hybrid* methods include **Unleash** [Xu *et al.*, 2023b], and **S²ynRE** [Xu *et al.*, 2023a].

Implementation Details. In the LLM data generation part, we leverage the 13B version for LLama-2 (llama-2-13b-chat-hf) to conduct experiments, and we double the K-shot training set through LLMs. For example, for an 8-shot training set, we construct 8 pieces of pseudo data per relation, thereby creating the final augmented training set.

In the Traditional RE Model Pre-Training part, we adopt TYP Marker and GenPT as the base architecture. Regarding

Paradigms	Models	SemEval			TACRED			TACREV			Re-TACRED			Avg.
		k=8	k=16	k=32	k=8	k=16	k=32	k=8	k=16	k=32	k=8	k=16	k=32	
Traditional Methods	CP	75.6	81.8	83.8	33.7	34.9	36.4	33.2	33.9	34.8	55.8	58.9	65.4	52.3
	TYP Marker	70.9	77.1	80.3	28.9	32.0	32.4	27.6	31.2	32.0	44.8	54.1	60.0	47.6
	KnowPrompt	74.3	82.9	84.8	32.0	35.4	36.5	32.1	33.1	34.7	55.3	63.3	65.0	52.5
	GenPT	77.1	81.5	83.9	35.7	36.6	37.4	34.4	34.6	36.2	57.2	60.4	65.2	53.3
	PTRE	79.0	83.0	84.6	32.7	36.5	<u>39.2</u>	33.9	36.3	38.1	58.8	<u>64.4</u>	66.3	54.4
LLM-based Methods	LLama-2*		56.9			26.5			27.3			38.1		37.2
	GPT-3.5*		60.5			29.7			30.0			39.6		39.9
	GPT-4*		65.3			32.3			32.1			45.9		43.9
Hybrid Methods	Unleash (TYP Marker)	78.3	81.7	84.8	35.8	36.6	37.1	36.7	36.5	37.0	58.4	60.6	64.8	54.0
	S^{2}ynRE (CP)	<u>79.2</u>	<u>83.6</u>	<u>85.3</u>	36.3	37.2	38.7	35.2	36.1	36.9	<u>59.4</u>	63.9	<u>67.5</u>	54.9
	TKRE (TYP Marker)	77.6	81.7	84.1	<u>36.7</u>	<u>38.3</u>	39.1	<u>37.2</u>	<u>38.2</u>	<u>39.0</u>	58.7	63.8	65.5	<u>55.0</u>
	TKRE (GenPT)	**80.7**	**84.9**	**86.6**	**39.6**	**42.1**	**43.2**	**40.2**	**41.3**	**42.9**	**61.3**	**67.6**	**68.9**	**58.3**

Table 2: Micro F1 (%) of few-shot performance. The best results are in **bold** and the second best ones are <u>underlined</u>. * refers to the performance with one-shot demonstrations. (·) refers to the backbone model used.

the evaluation metrics, we adopt the micro F1 scores of RE as the primary metric to evaluate models, considering that F1 scores can assess the overall performance of precision and recall. More details are shown in Appendix.

4.2 Main Results

The main results are illustrated in Table 2. Our proposed **TKRE** model outperforms all baselines across all metrics. Particularly compared to the base architecture TYP Marker and GenPT, our method manifests a significant advantage. This demonstrates the effectiveness of our designs and the benefits of integrating traditional RE models and LLMs. Furthermore, there are also some interesting phenomena:

1) The vast majority of methods exhibit superior performance on the ReTACRED dataset compared to the TACRED and TACREV datasets. This is reasonable as Re-TACRED is an improved version among these three datasets, which addresses some shortcomings of the original TACRED dataset and refactors its training/development/test sets. The more precise labels contribute to the learning process of these models, thereby yielding superior performance.

2) Among these LLM-based methods, **GPT-4** demonstrates competitive performance and significantly outperforms **GPT-3.5** and **LLama-2** on all datasets. This proves its strong ability, as claimed in [Achiam *et al.*, 2023]. However, it exhibits the generation performance in FSRE. This may be because the training objective of **GPT-4** is on generative tasks, which predict the next word based on context, rather than optimizing specifically for RE tasks even though it utilized various very largescale corpora for training.

3) **Unleash** introduces a schema-constrained data augmentation method using LLMs to enhance the **TYP Marker** baseline by 6.4%. Similarly, **S^{2}ynRE** presents a two-stage self-training framework leveraging LLM-generated synthetic data, improving the **CP** baseline by 2.6%. Both approaches demonstrate a measurable improvement over their respective baselines, validating the feasibility of this line of research.

In comparison, our **TKRE** model significantly outperforms **TYP Marker** by 7.8% and **GenPT** by 5.0%, further confirming the effectiveness of our approach from a different perspective.

4.3 Ablation Study

To validate the effectiveness of main components in TKRE, we introduce the following variants for the ablation study:

TKRE *w/o* EDKG: This variant removes explanation-driven knowledge generation at the data preparation stage.

TKRE *w/o* SCDG: This variant removes schema-constrained data generation at the data preparation stage.

TKRE *w/o* MSLM: This variant abides by the mask setting of BERT[Devlin *et al.*, 2019] instead of masked span language modeling at the two-stage knowledge guidance pre-Training stage.

TKRE *w/o* SCL: This variant skips the span-level contrastive learning at the two-stage knowledge guidance pre-Training stage.

TKRE *w/o* ALL: This variant addresses the FSRE task by fine-tuning the baseline model (e.g., TYP Marker) solely on the golden training dataset, without incorporating any of the enhancements introduced by our method.

Models	SemEval			TACRED			Avg.
	k=8	k=16	k=32	k=8	k=16	k=32	
TKRE	**77.6**	**81.7**	**84.1**	**36.7**	**38.3**	**39.1**	**59.6**
w/o EDKG	73.1	75.8	80.2	30.9	33.7	36.1	55.0
w/o SCDG	75.2	79.3	82.1	34.1	35.3	37.3	57.2
w/o MSLM	75.3	78.5	81.1	35.4	36.3	36.9	57.2
w/o SCL	76.2	79.8	82.9	35.2	35.9	37.4	57.9
w/o ALL	70.9	77.1	80.3	28.9	32.0	32.4	53.6

Table 3: Ablation study results for TACRED.

From Table 3, we observe the following:

1) Removing EDKG significantly reduces performance, especially in smaller settings, highlighting its importance in generating relation-relevant explanations for FSRE.

2) The absence of SCDG leads to a performance decline, especially on TACRED with fewer training examples. This demonstrates the importance of schema constraints help LLMs toward more accurate and relevant data generation.

3) Replacing MSLM with traditional BERT-style masking results in a modest performance reduction, emphasizing the significance of span-based masking in capturing relational dependencies.

4) Omitting SCL leads to a slight drop in performance, suggesting that it plays a key role in differentiating relevant spans, which is crucial for fine-grained RE.

5) The variant with all enhancements removed exhibits the lowest performance, particularly for smaller K values. This reinforces the necessity of these enhancements in boosting the performance of FSRE.

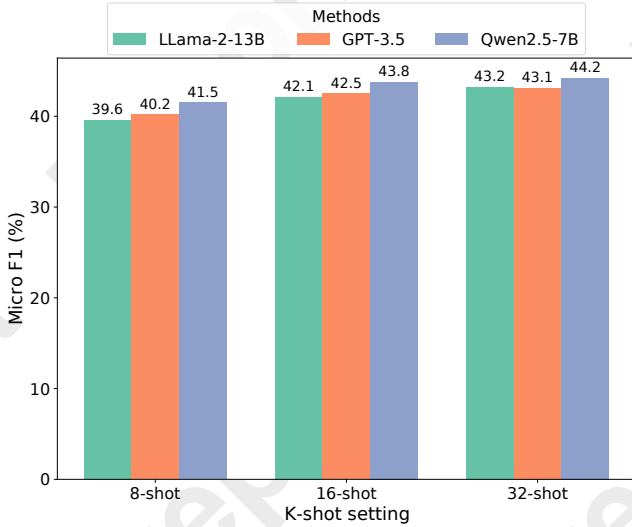### 4.4 Analysis and Discussions

Figure 5: Micro F1 (%) of **TKRE** (GenPT) with generated explanation-driven knowledge by various LLMs.

Explanation-driven Knowledge Generated from Different LLMs. We evaluate the impact of different LLMs applied to generate explanation-driven knowledge. We extend **GPT-3.5** and **Qwen2.5-7B** [Team, 2024] as other knowledge generators under the TACRED dataset, and continue model pre-training as well as fine-tuning under the same setting of Llama-2. As shown in Figure 5, the models pre-trained on different LLMs have similar performance. This indicates that our framework is not sensitive to different LLMs for FSRE.

Scale of Synthetic data. As shown in Figure 6, performance initially improves with increasing volume of schema-constrained data, but eventually declines, while still consistently surpassing the performance of the scenario without generated data. This observation suggests that incorporating schema-constrained data generated by LLMs can enhance
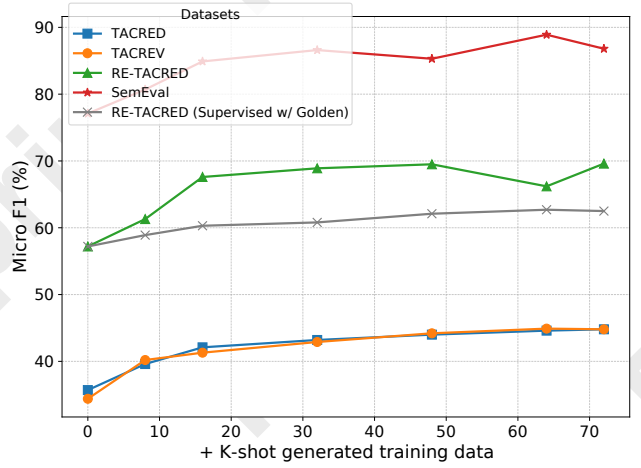

Figure 6: Micro F1 (%) of **TKRE** (GenPT) with generated training data and original 8-shot data.

FSRE performance up to a certain threshold. However, beyond this point, the inclusion of low-quality or noisy data begins to negatively affect the training process. This behavior is consistent with previous studies indicating that excessive noise in training data can undermine model performance, even considering the noise-resilience capabilities of LLMs [Song *et al.*, 2022].

We also report supervised results using additional golden training data to measure the utility of synthetic data. We can achieve two conclusions: 1) the advantage of golden training data are more significant when it is scaled up. However, this also takes more expensive costs. 2) TKRE demonstrates the ability to achieve performance on par with models trained using significantly larger volumes of manually annotated golden data, despite being trained with only a small amount of synthetic data. Furthermore, the generation of the requisite synthetic data requires only a few minutes of GPU computation.

5 Conclusion

In this work, we propose TKRE, a Two-Stage Knowledge Guided framework for FSRE, which effectively bridges the gap between generative and discriminative learning paradigms. By leveraging LLMs for explanation-driven knowledge generation and schema-constrained synthetic data, TKRE addresses key challenges in FSRE, including data scarcity and limited generalization ability. Our two-stage knowledge-guided pre-training approach integrates Masked Span Language Modeling and Span-Level Contrastive Learning, enhances traditional FSRE models' ability to capture relational reasoning and fine-grained semantic distinctions. Extensive experiments on multiple benchmark datasets demonstrate that TKRE significantly improves the performance of traditional FSRE models. Ablation studies further validate the effectiveness of each proposed component. Future work could explore extending TKRE to broader relation extraction tasks, such as zero-shot scenarios, or further optimizing the generation process to mitigate noise in synthetic data.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Alt *et al.*, 2020] Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. Tacred revisited: A thorough evaluation of the tacred relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

[Anaby-Tavor *et al.*, 2020] Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7383–7390, 2020.

[Chen *et al.*, 2022] Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web conference 2022*, pages 2778–2788, 2022.

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

[Gao *et al.*, 2021] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, 2021.

[Guo *et al.*, 2025] Quanjiang Guo, Yihong Dong, Ling Tian, Zhao Kang, Yu Zhang, and Sijie Wang. Baner: Boundary-aware llms for few-shot named entity recognition. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10375–10389, 2025.

[Han *et al.*, 2022] Jiale Han, Shuai Zhao, Bo Cheng, Shengkun Ma, and Wei Lu. Generative prompt tuning for relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3170–3185, 2022.

[Hendrickx *et al.*, 2010] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *ACL 2010*, page 33, 2010.

[Jiao *et al.*, 2020] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, 2020.

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

[Li *et al.*, 2023] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, 2023.

[Liu *et al.*, 2021] Zihan Liu, Yan Xu, Tiezheng Xu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13452–13460, 2021.

[Mintz *et al.*, 2009] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, 2009.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[Peng *et al.*, 2020] Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Learning from context or names? an empirical study on neural relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672, 2020.

[Qin *et al.*, 2021] Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. Erica: Improving entity and relation understanding for pre-trained language models via contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3350–3363, 2021.

[Schick and Schütze, 2021] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, 2021.

[Soares *et al.*, 2019] Livio Baldini Soares, Nicholas Fitzgerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, 2019.

[Song *et al.*, 2022] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from

noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153, 2022.

[Stoica *et al.*, 2021] George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. Re-tacred: Addressing shortcomings of the tacred dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13843–13850, 2021.

[Team, 2024] Qwen Team. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Vrandečić and Krötzsch, 2014] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[Vu *et al.*, 2021] Tu Vu, Minh-Thang Luong, Quoc Le, Grady Simon, and Mohit Iyyer. Strata: Self-training with task augmentation for better few-shot learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5715–5731, 2021.

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

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

[Wei and Zou, 2019] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, 2019.

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

[Whitehouse *et al.*, 2023] Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. Llm-powered data augmentation for enhanced cross-lingual performance. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[Xu *et al.*, 2023a] Benfeng Xu, Quan Wang, Yajuan Lyu, Dai Dai, Yongdong Zhang, and Zhendong Mao. S2ynre: Two-stage self-training with synthetic data for low-resource relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8186–8207, 2023.

[Xu *et al.*, 2023b] Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. How to unleash the power of large language models for few-shot relation extraction? In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 190–200, 2023.

[Yang *et al.*, 2024] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32, 2024.

[Zhang and Kang, 2024] Yu Zhang and Zhao Kang. Tpn: Transferable proto-learning network towards few-shot document-level relation extraction. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2024.

[Zhang *et al.*, 2017] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. Position-aware attention and supervised data improve slot filling. In *Conference on empirical methods in natural language processing*, 2017.

[Zhang *et al.*, 2024a] Zhihao Zhang, Sophia Lee, Junshuang Wu, Dong Zhang, Shoushan Li, Erik Cambria, and Guodong Zhou. Cross-domain ner with generated task-oriented knowledge: An empirical study from information density perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1595–1609, 2024.

[Zhang *et al.*, 2024b] Zirui Zhang, Yiyu Yang, and Benhui Chen. Prompt tuning for few-shot relation extraction via modeling global and local graphs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13233–13242, 2024.

[Zhou and Chen, 2022] Wenxuan Zhou and Muhao Chen. An improved baseline for sentence-level relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168, 2022.

[Zhu *et al.*, 2024] Xudong Zhu, Zhao Kang, and Bei Hui. Fcds: Fusing constituency and dependency syntax into document-level relation extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7141–7152, 2024.