CRE-LLM: A Domain-Specific Chinese Relation Extraction Framework with Fine-tuned Large Language Model

Zhengpeng Shi[1] , **Haoran Luo**[2*]

[1]College of Statistics and Mathematics, Zhejiang Gongshang University, China
[2]School of Computer Science, Beijing University of Posts and Telecommunications, China
shizhengpeng@outlook.com, luohaoran@bupt.edu.cn

Abstract

Domain-Specific Chinese Relation Extraction (DS-CRE) aims to extract relations between entities from domain-specific Chinese text. Despite the rapid development of PLMs in recent years, especially LLMs, DSCRE still faces three core challenges: complex network structure design, poor awareness, and high consumption of fine-tuning. Given the impressive performance of large language models (LLMs) in natural language processing, we propose a new framework called CRE-LLM. This framework is based on fine-tuning open-source LLMs, such as Llama-2, ChatGLM2, and Baichuan2. CRE-LLM enhances the logic-awareness and generative capabilities of the model by constructing an appropriate prompt and utilizing open-source LLMs for instruction-supervised fine-tuning. And then it directly extracts the relations of the given entities in the input textual data, which improving the CRE approach. To demonstrate the effectiveness of the proposed framework, we conducted extensive experiments on two domain-specific CRE datasets, FinRE and SanWen. The experimental results show that CRE-LLM is significantly superior and robust, achieving state-of-the-art (SOTA) performance on the FinRE dataset. This paper introduces a novel approach to domain-specific relation extraction (DSCRE) tasks that are semantically more complex by combining LLMs with triples. Our code is publicly available[1].

1 Introduction

Domain-Specific Chinese Relation Extraction (DSCRE) is a crucial task in the field of Natural Language Processing (NLP). Its objective is to extract relations between given entities from domain-specific unstructured Chinese text. Examples of such relations include financial and biomedical relations. The main difficulties of this task are due to the fact that Chinese datasets specific to certain domains are mostly

---

* Corresponding author.
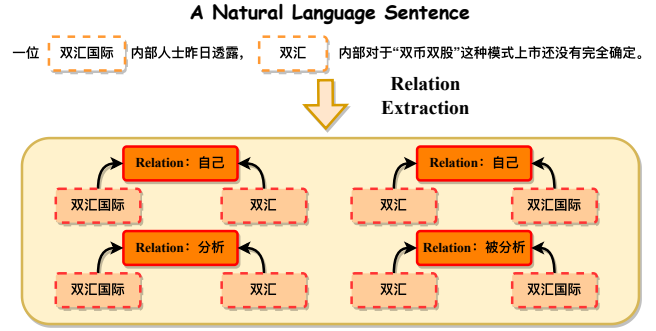[1]https://github.com/SkyuForever/CRE-LLM

Figure 1: An example of Domain-specific CRE Task.

private, with limited informative data and resources. Additionally, dealing with the diversity of linguistic expressions and potential ambiguities in the text is a challenge. Furthermore, the limitations of the Chinese corpus and the low use of dummy words and lexemes in Chinese, which makes it challenging to extract relation between entities from domain-specific Chinese texts. A specific example as shown in Figure 1.

The field of deep learning has rapidly developed and achieved significant success in relation extraction tasks. Pre-trained models (shown in Figure 2a), such as BERT [Devlin *et al.*, 2019] and T5 [Raffel *et al.*, 2019], have been widely used in this area. For instance, when dealing with domain-specific Chinese relation extraction, BERT-PAGG [Xu *et al.*, 2023] combines the positional information of entities with local features extracted by the PAGG module and entity vector representations outputted by BERT to achieve relation extraction. Similarly, MoVE [Yang *et al.*, 2023b] achieves this by dynamically learning multi-view features. LLMs can also be used for domain-specific relation extraction (shown in Figure 2b), such as ERNIE 3.0 [Sun *et al.*, 2021] and ERNIE 3.0 TITAN [Wang *et al.*, 2021], the continuous learning semantic understanding frameworks were developed based on knowledge augmentation. It utilized self-supervised contrastive learning pre-training technique of word-mixing and self-adversarial fine-tuning with word-mixing data augmentation to improve RE tasks. Additionally, GPT-FinRE [Rajpoot and Parikh, 2023] can also be used for this purpose (shown in Figure 2c). It uses OpenAI models under the ICL frame-
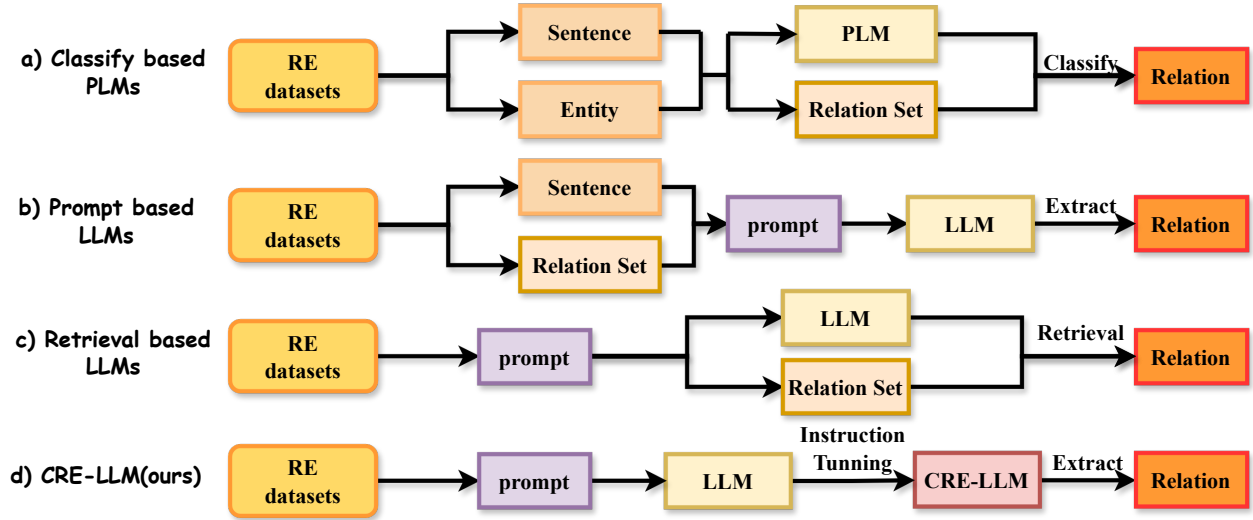
Figure 2: Illustration of 4 different paradigms for solving CRE task. As shown in Figure 2a, entities and texts from the RE datasets are inputted separately into the PLM. And the PLM is combined with the Relation Set and output the relation with the highest probability as result. As shown in Figure 2b, prompts are constructed based on the texts and Relation Set from the RE dataset and input them into the LLM to generate relation. As shown in Figure 2c, the RE dataset is employed to construct the prompts and input them into the LLM to generate preliminary results, which are subsequently retrieved with the Relation Set to obtain relation extraction results. As shown in Figure 2d, our method directly utilizes a fine-tuning dataset constructed from the RE dataset to fine-tune the LLM and then generate accurate relation extraction results.

work and incorporates retrieval mechanisms to achieve financial relation extraction, has strong potential and performance in general.

Several approaches have been proposed for this task, resulting in significant performance improvements. However, three main challenges and difficulties remain. **(1) Complex network design is required.** Methods such as BERT-PAGG [Xu *et al.*, 2023], MoVE [Yang *et al.*, 2023b], and other BERT-type pre-trained models have been proposed, which require the construction of complex networks and the combination of external and internal information to improve model performance. **(2) Poor perception is a challenge.** Direct use of large models for event extraction may not perceive internal Relations, especially in the domain-specific of Chinese text. A sentence contains semantic information from different perspectives, including words, structure and contextual semantic information, as well as domain-specific proper nouns. These difficulties can impede the understanding of the model, so direct use of the model for relation extraction often yields unsatisfactory results. **(3) Fine-tuning the model to achieve the domain-specific CRE task results in significant memory consumption.** GPT-3 [Brown *et al.*, 2020] demonstrates that expanding the pre-trained models can further exploit its potential. However, fine-tuning large-scale pre-trained models such as ERNIE 3.0 [Sun *et al.*, 2021] and ERNIE 3.0 TITAN [Wang *et al.*, 2021] to achieve domain-specific Chinese relation extraction requires a large amount of memory consumption, which may be difficult for a typical team to achieve.

To address the aforementioned challenges, we introduce CRE-LLM (shown in Figure 2d), a framework for extracting domain-specific Chinese relations using fine-tuned open-

source large models such as Llama-2-7B [Touvron *et al.*, 2023], ChatGLM2-6B [Du *et al.*, 2022], and Baichuan2-7B [Yang *et al.*, 2023a]. CRE-LLM proposes a direct and concise method for relation extraction by calling a fine-tuned open-source large model and constructing an appropriate prompt. (1) The method selects a pre-trained open-source large model and constructs an appropriate prompt to directly complete relation extraction. This eliminates the need to design and build complex network structures. Instead, the PEFT [Mangrulkar *et al.*, 2022] framework can be used to achieve end-to-end question and answer quickly and efficiently [Luo *et al.*, 2023]. Additionally, larger models typically have more parameters and often yield better relation extraction results. (2) In order to enable the model to perceive and comprehend the internal relations between the given sentences and entities, we simultaneously utilize instruction-supervised fine-tuning methods. This enhances its logical perception and generation capabilities, resulting in appropriate extraction outcomes. (3) To leverage the enhanced performance of large-scale language models, we apply the PEFT [Mangrulkar *et al.*, 2022] framework to fine-tune LLMs. This substantially decreases memory consumption and enhances training efficiency, enabling typical projects and teams to harness LLMs for domain-specific CRE tasks.

To evaluate the performance of our proposed framework, we applied it to datasets of CRE tasks from two different domains: FinRE [Li *et al.*, 2019] and SanWen [Li *et al.*, 2019]. The experimental results demonstrate the excellent performance of CRE-LLM on the domain-specific CRE tasks, while achieving new state-of-the-art (SOTA) performance on FinRE. We also conducted additional experiments

to verify whether our relation extraction framework improves the relation extraction accuracy and efficiency. Finally, we also discuss how the insights from this framework allow us to envision future combinations of LLMs and domain-specific CRE.

2 Related Work

2.1 Domain-Specific RE with PLMs

Categorical relation extraction based on pre-trained language models (PLMs) such as BERT [Devlin *et al.*, 2019] and T5 [Raffel *et al.*, 2019] is currently one of the mainstream methods for extracting relations in specific domains. The two main pre-trained language models used in the biomedical domain are BioBERT [Peng *et al.*, 2019] and PubMedBERT [Gu *et al.*, 2022]. These models are suitable for real-world tasks such as entity and relation extraction. In the financial domain, T5-base [Zhao *et al.*, 2019] is a commonly used pre-trained language model. It is important to note that technical term abbreviations should always be explained when first used. In order to improve the overall level of Chinese financial natural language processing (NLP) and promote the development of information extraction, several models have been proposed, including FinBERT [Araci, 2019] and Mengzi-BERT-base-fin [Zhang *et al.*, 2021]. Furthermore, in the domain-specific CRE task, BERT-PAGG [Xu *et al.*, 2023] has developed the PAGG module to enable the model to synthesize multiple pieces of information about entities. Additionally, MoVE [Yang *et al.*, 2023b] has integrated different view representations through hybrid viewpoints and expert mechanisms, which effectively filter out noisy information and improve the efficiency of model training and inference.

This paper introduces CRE-LLM, a framework for DSCRE that deviates from the traditional classification-based approach for relation extraction and instead uses a generative method. Fine-tuned open-source LLMs are utilized to directly discern relations between given entities through the generation process. The framework aims to solve the complex network structure design problem of previous pre-trained models like BERT [Devlin *et al.*, 2019] and T5 [Raffel *et al.*, 2019].

2.2 Domain-Specific RE with LLMs

The emergence phenomenon has been demonstrated by the introduction of ChatGPT [Ouyang *et al.*, 2022] and GPT-4 [OpenAI, 2023], which are decoder-only Large Language Models (LLMs) with a large number of parameters. These models have shown strong performance on natural language problems, making many traditional NLP tasks easier [Zhao *et al.*, 2023]. The emergence of open-source large models such as Llama2-7B [Touvron *et al.*, 2023], ChatGLM2-6B [Du *et al.*, 2022], and Baichuan2-7B [Yang *et al.*, 2023a] have made the use of LLMs more convenient. For example, GCRE-GPT [Wang *et al.*, 2023] uses a generative approach to extract complex multiple comparison relations from input text. On the other hand, GPT-RE [Wan *et al.*, 2023] focuses on entity and relation information and use gold label induced reasoning to bridge the performance gap of RE. GPT-FinRE [Rajpoot and

Parikh, 2023], which has strong potential to implement financial relation extraction using OpenAI models and combining retrieval mechanisms under the ICL framework. Therefore, leveraging powerful semantic parsing capabilities of the LLMs to design and implement relation extraction through multiple rounds of dialogue or retrieval is a promising approach.

In this paper, our proposed CRE-LLM is a novel approach to relation extraction. It employs a fine-tuned open-source LLMs and leverages the powerful capabilities of LLMs in text understanding, generation and generalization. This enables the extraction of relations between specified entities in a simple and direct end-to-end method from unstructured Chinese text.

2.3 Fine-Tuning for Large Pre-Trained Models

With the open-source LLMs like Llama-2-7B [Touvron *et al.*, 2023], ChatGLM2-6B [Du *et al.*, 2022] and Baichuan2-7B [Yang *et al.*, 2023a] emerging. It can be supervised fine-tuning (SFT) by using Parameter-Efficient Fine-Tuning (PEFT) technologies [Mangrulkar *et al.*, 2022] such as LoRA [Hu *et al.*, 2021], QLoRA [Dettmers *et al.*, 2023], P-Tuning v2 [Liu *et al.*, 2021], and Freeze [Geva *et al.*, 2021], enhancing the capabilities of LLMs for specific tasks. For instance, ERNIE 3.0 [Sun *et al.*, 2021] and ERNIE 3.0 TI-TAN [Wang *et al.*, 2021] achieved significant performance improvements on domain-specific CRE datasets through full fine-tuning. Additionally, the fine-tuning approach was used to construct the BBT-FinT5 [Lu *et al.*, 2023] model based on the T5 model, which promotes the development of natural language processing (NLP) in Chinese finance. In conclusion, the implementation of domain-specific CRE tasks through fine-tuning LLMs is highly effective and significant. This enables more general projects and teams to deploy implementations.

The CRE-LLM integrates the semantic parsing capability of LLMs with the advantages of instruction-supervised fine-tuning. This provides convenience and possibility for LLMs to be deployed in DSCRE tasks.

3 Preliminaries

3.1 Problem Definition

Given an input sentence of n words $s = \{x_1, x_2, ..., x_n\}$, an entity e is a consecutive span of words where $e = \{xi, xi + 1, ..., x_j\}$, $i, j \in (1, ..., n)$. For each sentence s, the output of the CRE-LLM is a set of facts where each fact consists of a relation triplet. A relation triplet consists of the relation $r \in R$ between head entity e_{head} and tail entity e_{tail} where R is the predefined set of relation labels. Hence, a relation triplet has three components: (e_{head}, r, e_{tail}). A specific example is shown in Figure 3.

3.2 Parameter-Efficient Fine-Tuning Methods

The current trend in pre-trained language models (PLMs), represented by models like ChatGPT [Ouyang *et al.*, 2022], is towards increasing scale. This leads to a growing cost for
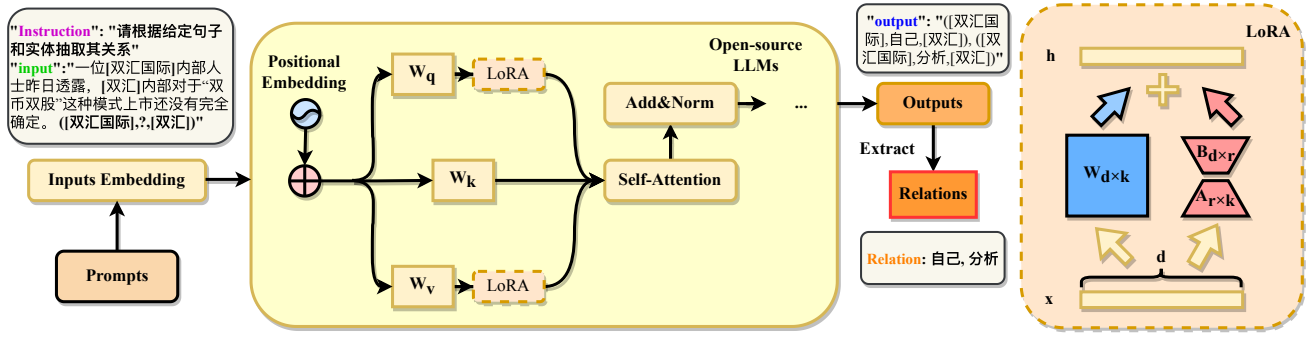
Figure 3: The overview of CRE-LLM for domain-specific Chinese relation extraction method with supervised fine-tuned LLMs by using Parameter-Efficient Fine-Tuning (PEFT) technologies (e.g. LoRA).

full fine-tuning. PEFT [Mangrulkar *et al.*, 2022] methods address this challenge by freezing the majority of model parameters and fine-tuning only a small or additional set of parameters. This approach achieves comparable performance to full fine-tuning while significantly reducing the overall fine-tuning costs. LoRA [Hu *et al.*, 2021] , as a form of PEFT [Mangrulkar *et al.*, 2022], which employs low-rank approximation, introducing low-rank matrix modules labeled as A and B. This approach helps to reduce memory consumption during fine-tuning of LLMs by minimizing changes in weights associated with the parameters of the model. For instance, considering the weight matrix W_0 of a pre-trained model, its update is constrained through low-rank decomposition, involving training only B and A, where $B \in R_{d \times r}, A \in Rr \times k, r \ll min(d, k)$. The specific update is expressed as follows:

$$h = W_0 x + \Delta W x = W_0 x + BAx. \qquad (1)$$

Therefore, LoRA [Hu *et al.*, 2021] framework offers the following advantages: generalization comparable to full fine-tuning, no additional inference latency, and reduced consumption of memory and storage resources.

4 Methodology

In this section, we first give a brief overview of CRE-LLM and introduced by the following three main sections: Instruction design, Efficient Fine-Tuning on LLMs and Discussion of Methods. We explain how the fine-tuning dataset is constructed, why we use the PEFT [Mangrulkar *et al.*, 2022] framework to fine-tune LLMs and comparison of CRE-LLM to other methods.

4.1 Overview of CRE-LLM

CRE-LLM is a framework for domain-specific CRE based on generative question-answering, utilizing open-source LLMs fine-tuned with instruction supervision. Firstly, CRE-LLM involves constructing effective instructions based on natural language and specified entities in CRE datasets. Simultaneously, it sets up reasonable input and output configurations to enable the model to better understand and accomplish the task. Subsequently, leverage PEFT [Mangrulkar *et al.*, 2022] framework to achieve efficient fine-tuning of LLMs. And then, from the given Chinese text data, the fine-tuned LLMs

are employed to generate inference results in the form of triplets by means of prompts. Finally, the framework employs a direct extraction process to derive the relations from the generated triplets, thereby elucidating the relations between the specified entities in the Chinese text. The detailed framework description is illustrated in Figure 3.

4.2 Instruction Design

To construct an instruction fine-tuning dataset, we need to design instruction, input and output. In this context, we set the "instruction" for inputting LLMs as "Please extract the relation based on the given sentence and entities." And the entities in each instance of the relation extraction dataset (e.g., "Shuanghui") are labeled with the given entities using "[]" (e.g., "[Shuanghui]") to help LLMs understand the meaning of the concept referred to by the term "given entities" and better focus on and comprehend the relation between the given entities.

Subsequently, to facilitate LLMs understanding of the task requirements and ensure the accuracy of inferring relation extraction, we append a triplet and the format is that $([e_{head}], ?, [e_{tail}])$ (e.g., "([Shuanghui International], ? ,[Shuanghui])"), where "?" represents the relation between the given entities. The specific format of the instance data is detailed in Figure 3. Combining the relations between given entities, we construct triplets (e.g., "([Shuanghui International], analysis ,[Shuanghui])") as both "input" and "output," and finally, the previously constructed instruction is added to complete an instance in the fine-tuning dataset.

Additionally, if multiple relation exist between entities in a sentence, we list all relation in the inference results, separated by "," (e.g."([Shuanghui International], self, [Shuanghui]), ([Shuanghui], analysis, [Shuanghui International])"). Following this structured process, the instruction fine-tuning training dataset for open-source LLMs is constructed.

4.3 Efficient Fine-Tuning on LLMs

To reduce the substantial cost of fine-tuning LLMs, overcome the limitations of the available information resources for domain-specific CRE tasks and ensure that LLMs can generate standardized relation extraction results, it is necessary to implement a solution that addresses these issues. The CRE-LLM employs the PEFT [Mangrulkar *et al.*, 2022] framework

to address these challenges and minimize the number of fine-tuning parameters.

The CRE-LLM allows switching between all the aforementioned fine-tuning methods and open-source LLMs, such as Llama-2-7B [Touvron *et al.*, 2023], ChatGLM2-6B [Du *et al.*, 2022], and Baichuan2-7B [Yang *et al.*, 2023a]. As shown in Figure 3, for these large-parameter-only decoder LLMs, CRE-LLM adopts PEFT [Mangrulkar *et al.*, 2022] technology. It fine-tunes the Q and V parts of the input in the GQA section using LoRA [Hu *et al.*, 2021], adds them to the K part, and then calculates the attention using the following formula:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \qquad (2)$$

Then, through a series of network architecture layers, the relations of the given entities in the input text are generated. The underlying mechanism can be outlined by the following formula:

$$p_\theta(\mathcal{Y}|\mathcal{X}, \mathcal{P}) = \prod_{i=1}^{m} p_\theta(y_i|\mathcal{X}, \mathcal{P}, y < i), \qquad (3)$$

where $\mathcal{X} = [x_1, x_2, ..., x_n]$ is the text sequence to be extracted, $\mathcal{Y} = [y_1, y_2, ..., y_m]$ is the target sequence, and $\mathcal{P}$ is the prompt. Through PEFT [Mangrulkar *et al.*, 2022] framework, the problem of poor internal perceptual capabilities in general open-source LLMs is addressed. It simultaneously enhances the generation understanding and generalization abilities of LLMs with respect to texts in this domain specific. This can be widely applied to domain-specific CRE tasks, generating more accurate results.

4.4 Discussion of Methods

Based on the explanation and process description of the CRE-LLM method above, let's discuss and compare it with existing CRE methods:

(1) Comparison with Classify based PLMs. The PLMs, such as BERT [Devlin *et al.*, 2019], T5 [Raffel *et al.*, 2019], ERNIE 3.0 [Sun *et al.*, 2021] and ERNIE 3.0 TITAN [Wang *et al.*, 2021] primarily implement CRE using classification-based methods. In contrast, the CRE-LLM utilizes PEFT framework to fine-tune LLMs and employs a generative approach for CRE.

(2) Comparison with Classify-then-Extract based LLMs. The direct invocation of LLMs, such as ChatGPT [Ouyang *et al.*, 2022] and GPT4 [OpenAI, 2023] involves constructing a prompt based on the Classify-then-Extract approach. This prompt must include the Relation Set as options for the relation between the given entities. In contrast, CRE-LLM employs a fine-tuned LLM that has incorporated the knowledge of the Relation Set, eliminating the need for an excessively lengthy prompt while achieving the domain-specific CRE task.

(3) Comparison with Generate-then-Retrieval based LLMs. Regarding the Generate-then-Retrieval method, LLMs generate the relation between given entities directly and then align it with the Relation Set through retrieval. In contrast, CRE-LLM does not require additional retrieval alignment with the Relation Set. It can directly infer and generate more accurate relation extraction results.

5 Experiments

This section covers experimental configurations, results, and analysis. We address the following research questions (RQs): **RQ1**: Is CRE-LLM superior to other CRE methods? **RQ2**: Is fine-tuning of LLMs effective? **RQ3**: Is the design of the prompts for LLMs reasonable? **RQ4**: Why use fine-tuned open-source LLMs instead of directly invoking ChatGPT for CRE? **RQ5**: Does LoRA fine-tuning reduce GPU memory consumption and environmental configuration requirements, and does it improve training efficiency? **RQ6**: How about error analysis?

5.1 Experimental Setup

Datasets. All experiments were conducted on two Chinese datasets from different domains: FinRE [Li *et al.*, 2019], based on 2647 financial news articles from Sina Finance, containing 44 distinguished relations, including the special relation "NA" indicating no relation between the marked entity pairs. The dataset was split into 26,971, 2,977, and 7,453 relation extraction instances for training, validation and testing respectively. SanWen [Li *et al.*, 2019] is based on 837 Chinese literary works, comprising 10 distinguishable relations, also including the special relation "NA." It was respectively divided into 17,227, 1,793 and 2,220 relation extraction instances for training, validation, and testing.

Baselines. CRE-LLM was compared with several CRE baselines, including ERNIE 3.0 TITAN [Wang *et al.*, 2021], Bert-PAGG [Xu *et al.*, 2023], MoVE [Yang *et al.*, 2023b], and other CRE methods mentioned in Section 2.

Evaluation Metrics. Following previous work [Sun *et al.*, 2019; Sun *et al.*, 2021; Wang *et al.*, 2021; Lu *et al.*, 2023; Xu *et al.*, 2023; Yang *et al.*, 2023b], we used Precision, Recall, and F1-Score to evaluate the performance of methods.

Hyperparameters and Environment. For the FinRE dataset [Li *et al.*, 2019], we fine-tuned the LLM for 5 epochs with a learning rate of 5e-5. For SanWen [Li *et al.*, 2019], fine-tuning was performed for 10 epochs with a learning rate of 5e-4. The batch size was set to 4, and the gradient accumulation steps were 5e-5. All experiments were conducted on a single NVIDIA A40 GPU (48 GB), and the results are averages from five experiments with different random seeds.

5.2 Main Result (RQ1)

For the DSCRE task, to validate that CRE-LLM outperforms other CRE methods as proposed in this paper, we conducted inference tests through designed experiments. The experimental results, listed in Table 1, include fine-tuning Baichuan2-13B [Yang *et al.*, 2023a] on the FinRE [Li *et al.*, 2019] and SanWen [Li *et al.*, 2019] using LoRA [Hu *et al.*, 2021]. Additionally, we compared CRE-LLM with other baseline models. From Table 1, it can be observed that CRE-LLM outperforms the majority of baseline models on the SanWen [Li *et al.*, 2019] and demonstrates significant improvements over all existing CRE methods on the FinRE [Li *et al.*, 2019]. Comparing with the previous best scores, CRE-LLM increased F1 score on the FinRE [Li *et al.*, 2019] validation set and test set by approximately 5.98% and 6.68%, respectively. This reflects the state-of-the-art performance of CRE-LLM in domain-specific CRE capabilities.

Method	FinRE		SanWen	
	Dev	Test	Dev	Test
T5-base	-	54.93	-	-
ERNIE 2.0	63.33	60.60	79.92	77.97
FinBERT-base	-	55.33	-	-
ERNIE 3.0	64.87	62.88	81.32	82.59
Mengzi-BERT-base-fin	-	58.25	-	-
ERNIE 3.0 TITAN	65.27	63.15	**83.07**	82.70
BBT-FinT5-base	-	60.62	-	-
BBT-FinT5-large	-	61.88	-	-
Bert-PAGG	-	53.01	-	73.83
BERT+MultiView	-	53.89	-	72.98
CRE-LLM(ours)	**69.17**	**67.37**	81.30	**82.74**

Table 1: The F1 score comparison of CRE-LLM with other baselines on FinRE [Li *et al.*, 2019] and SanWen [Li *et al.*, 2019] datasets. The results are mainly taken from their original paper. For our proposed CRE-LLM, we display the results of the best setup on FinRE and SanWen. The results of the FinRE and SanWen datasets in RQ1 are all attained utilising Baichuan2-13B, which is fine-tuned with LoRA. The best results in each metric are in **bold**.

5.3 Effectiveness of LLM's Fine-Tuning (RQ2)

To validate the effectiveness of fine-tuning LLMs, we chose to use the FinRE dataset [Li *et al.*, 2019]. We randomly selected 20%, 40%, 60%, and 80% of the training data and fine-tuned Baichuan2-13B [Yang *et al.*, 2023a] using LoRA [Hu *et al.*, 2021] on each subset. We then compared their inference test results with the results of the model fine-tuned on the complete dataset.

Fine-Tuning Setting	Precision	Recall	F1 score
Baichuan2-13B *+20%Training Data*	63.25	60.04	61.10
Baichuan2-13B *+40%Training Data*	66.97	63.74	64.73
Baichuan2-13B *+60%Training Data*	69.44	66.25	67.20
Baichuan2-13B ***+80%Training Data***	**69.71**	**66.89**	**67.65**
Baichuan2-13B	69.43	66.65	67.37

Table 2: Ablation study for LLM's Fine-Tuning on FinRE.

As shown in the Table 2, the performance of the model on domain-specific CRE tasks gradually improves with an increase in the training data, demonstrating the effectiveness of fine-tuning. Additionally, we observed that when using only 40% of the training data for fine-tuning, the F1 score already surpassed the original best performance. This indicates that fine-tuning allows LLMs to learn effectively from a limited dataset, achieving commendable performance. Moreover, when fine-tuning with 60% or more of the training data, there is minimal improvement in the F1 score. This suggests that instruction fine-tuning exhibits good generalization capabilities, emphasizing the importance of the quality of training samples over quantity.

5.4 The Impact of Instruction Design (RQ3)

To validate the effectiveness and rationality of the dataset constructed for the DSCRE task using our proposed fine-tuning-based generative approach with open-source LLMs, we conducted ablation experiments on various components. Table 3 shows the effectiveness of the different components of Instruction Design on FinRE [Li *et al.*, 2019].

Fine-Tuning Setting	Precision	Recall	F1 score
Baichuan2-13B w/o EM	61.09	61.54	61.24
Baichuan2-13B w/o AT	68.67	65.13	66.23
Baichuan2-13B w/o TR	69.28	**67.07**	**67.50**
Baichuan2-13B w/o AT+TR	67.44	64.42	65.22
Baichuan2-13B	**69.43**	66.65	67.37

Table 3: Ablation study for Instruction Design on FinRE.

Effectiveness of Using "[]" Entity Markers (EM). As shown in Tabel 3, the model when fine-tuned without utilizing "[]" entity markers in the dataset resulted in a decrease in Precision, Recall and F1 score by 13.65%, 8.30% and 10.11% respectively. Using "[]" entity markers proves to be effective. It can facilitate LLMs in locating and understanding the given entities in the sentence, as well as aligning entities during the inference process. This helps improve the performance of CRE .

Effectiveness of Adding Triplets (AT) After the Input Text Tabel 3 demonstrates that removing the triplet part leads to a decrease in Precision, Recall, F1 score by 1.11%, 2.33% and 1.81% respectively. The absence of the triplet in the input text during the fine-tuning phase makes the model only observe the positional relation between entities from the output part. It hinders the learning process and reducing the performance of CRE.

Effectiveness of the Triplet Results(TR) As shown in Table 3, from the results of the ablation experiments, it can be observed that TR does not significantly improve Precision, Recall and F1 Score. However, in the absence of AT, TR can respectively increase Precision, Recall and F1 Score by 2.95%, 3.46%, and 3.39%. This is because in such a scenario, demanding direct output of relation extraction results from LLM may lead to insufficient understanding of entities and their relation reasoning. The setting of TR provides partial informative cues and aligns with entities indicated in the input, leveraging the well-established learning patterns of LLM. This allows for better understanding of task requirements, resulting in superior performance on domain-specific CRE tasks. Therefore, we consider TR to be effective, contributing to enhanced robustness of LLM in domain-specific CRE performance.

Comparison among all components. The results of ablation experiments for all components are shown in Table 3, where the settings of EM and AT significantly improve Precision, Recall and F1 score. Particularly, EM has the most significant impact on the performance of the model in CRE.

For the TR part, its influence on the performance of the model is substantial without AT settings. However, with AT settings, its impact becomes very limited. Consequently, AT has a stronger effect on LLM's performance in CRE than TR, with TR playing a role similar to AT but to a lesser extent. In summary, EM has the most significant impact on LLM's performance in DSCRE, followed by AT, and TR has the smallest impact.

5.5 Comparison with ChatGPT (RQ4)

To illustrate why CRE-LLM chooses to fine-tune open-source LLMs for relation extraction, instead of directly using Chat-GPT to accomplish this task, experiments were designed to compare these two approaches. Therefore, we conducted experiments on FinRE [Li *et al.*, 2019] for relation extraction. Table 4 shows the Comparison with ChatGPT[Ouyang *et al.*, 2022].

Method	Precision	Recall	F1 score
ChatGPT *+Classify-then-Extract*	27.59	25.86	26.44
ChatGPT *+Generate-then-Retrieval*	25.86	24.14	24.71
Baichuan2-13B+LoRA	**69.43**	**66.65**	**67.37**

Table 4: CRE result comparison of CRE-LLM with ChatGPT on FinRE. For our proposed CRE-LLM, we display the results of the basic setup on FinRE. The best results in each metric are in **bold**.

Comparison with Classify-then-Extract. This method involves constructing a Relation Set containing all distinguished relations from the RE dataset. The prompt is modified to introduce the Relation Set, which allows ChatGPT [Ouyang *et al.*, 2022] to select the appropriate relation between the given entities from the Relation Set. As shown in Table 4, despite ChatGPT [Ouyang *et al.*, 2022] having a larger number of model parameters, it is not open-source and cannot be fine-tuned, posing a challenge for generating standard relation extraction results directly. The challenge presented by the dataset is the necessity to provide 44 distinguished relations for ChatGPT [Ouyang *et al.*, 2022] to differentiate. This necessitates the input prompt being longer and more challenging for the model to understand, consequently reducing the performance of CRE.

Comparison with Generate-then-Retrieval. This method involves using ChatGPT [Ouyang *et al.*, 2022] to directly infer and generate relations. In contrast to the Classify-then-Extract approach, this method does not necessitate the input of all special relations, simplifying the prompt significantly. However, the generated results are more diverse. SimCSE [Gao *et al.*, 2021] is employed to align the results with the Relation Set for relation extraction. As shown in Table 4, due to the lack of fine-tuning, ChatGPT [Ouyang *et al.*, 2022] exhibits weaker internal understanding and logical reasoning capabilities for domain-specific Chinese text. It fails to generate results with precise meanings. Consequently, the performance of relation extraction is naturally not ideal after retrieval alignment through SimCSE [Gao *et al.*, 2021].

5.6 Analysis of Efficiency (RQ5)

For domain-specific CRE tasks, training efficiency and costs also require significant attention. To illustrate that CRE-LLM achieves higher efficiency and lower cost, our experimental results as shown in Table 5.

Method	Trainable params↓	Training Time↓
ERNIE 3.0 *+Progressive Learing*	100M	11h30m 4h
Baichuan2-13B+LoRA *+40%Training Data*	**0.0655M**	**1h08m**
Baichuan2-13B+LoRA		2h32m

Table 5: Comparison of CRE-LLM with ERNIE 3.0 on FinRE.

The specific experimental results are presented in Table 5. Although ERNIE 3.0 [Sun *et al.*, 2021] exhibits superior performance after fine-tuning on the SanWen [Li *et al.*, 2019], Table 5 reveals that the number of parameters and GPU memory consumption required for general NLP tasks after fine-tuning ERNIE 3.0 is significantly more than for CRE-LLM. For instance, during the fine-tuning phase, ERNIE 3.0 would need at least eight 32GB V100 GPUs, indicating higher requirements for environmental configuration. Additionally, the fine-tuning training duration is also substantially longer. In contrast, CRE-LLM, which employs PEFT [Mangrulkar *et al.*, 2022] framework, achieves efficient parameter fine-tuning for LLMs. This results in a significant reduction in GPU memory usage and a lowering of environmental configuration demands, thereby enabling more general teams and projects to adopt it.

5.7 Error Analysis (RQ6)

We analyzed instances in the FinRE [Li *et al.*, 2019] test set where CRE-LLM failed to achieve correct CRE, and conducted a statistical analysis of the errors. The specific findings are summarized as follows:

Entity Relation Understanding Errors (52.21%). The primary source of errors in relation extraction is a misunderstanding of relations between entities. This is mainly due to the inherent difficulty of relation extraction in the dataset and the model's limited ability to precisely infer relations between entities in domain-specific natural language texts.

Errors with Multiple Relations between Entities (26.69%). Another category of errors arises when there are multiple relations between entities. For instance, in sentences like "With the establishment of [Ant Financial], [Alibaba]'s layout in the financial business has been officially clarified," , and the relations of the given entities are "ownership" and "establishment". The model may overlook one of the relations while generating the relation extraction result, leading to lower the performance of CRE.

"NA" Relation Errors (20.47%). The FinRE [Li *et al.*, 2019] dataset contains a special "NA" relation that is challenging to express in the text. Even for general readers, understanding the relation between the given entities may be

difficult. This complexity poses a challenge for the model to achieve accurate CRE.

Nonexistent Relation Errors (0.622%). Since CRE-LLM does not provide any relation options and relies on learning from training data to obtain the Relation Set of the dataset. Consequently, there may be instances where the model, despite fine-tuning, generates relation extraction results that do not exist in the Relation Set. This is due to the fine-tuned LLMs still retain certain independent capabilities in text understanding, generation, and generalization.

6 Conclusion

In this work, we propose CRE-LLM, a large language model framework for domain-specific Chinese relation extraction (DSCRE) based on fine-tunned open-source large models. This method represents a significant shift from traditional approaches. It employs the PEFT [Mangrulkar *et al.*, 2022] framework and open-source LLMs with numerous parameters to achieve a simple and efficient end-to-end generative relation extraction. It addresses inherent challenges such as complex network structure design, poor perception and high consumption of fine-tuning.

Our experimental results, based on two standard domain-specific CRE benchmarks, namely FinRE [Li *et al.*, 2019] and SanWen [Li *et al.*, 2019], demonstrate that CRE-LLM achieves state-of-the-art(SOTA) performance on the DSCRE tasks. Moreover, the simplicity, flexibility, and especially the efficiency of our framework make it a promising direction for applying LLMs to DSCRE tasks that involve stronger domain specificity and more challenging semantic understanding.

Acknowledgment

References

[Araci, 2019] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *CoRR*, abs/1908.10063, 2019.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

[Dettmers *et al.*, 2023] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient fine-tuning of quantized llms. *CoRR*, abs/2305.14314, 2023.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio,

editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[Du *et al.*, 2022] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022.

[Gao *et al.*, 2021] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Conference on Empirical Methods in Natural Language Processing*, 2021.

[Geva *et al.*, 2021] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5484–5495. Association for Computational Linguistics, 2021.

[Gu *et al.*, 2022] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Heal.*, 3(1):2:1–2:23, 2022.

[Hu *et al.*, 2021] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021.

[Li *et al.*, 2019] Ziran Li, Ning Ding, Zhiyuan Liu, Haitao Zheng, and Ying Shen. Chinese relation extraction with multi-grained information and external linguistic knowledge. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4377–4386. Association for Computational Linguistics, 2019.

[Liu *et al.*, 2021] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602, 2021.

[Lu *et al.*, 2023] Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *CoRR*, abs/2302.09432, 2023.

[Luo *et al.*, 2023] Haoran Luo, E. Haihong, Zichen Tang, Shiyao Peng, Yikai Guo, Wentai Zhang, Chenghao Ma, Guanting Dong, Meina Song, and Wei Lin. Chatkbqa: A generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models. *ArXiv*, abs/2310.08975, 2023.

[Mangrulkar *et al.*, 2022] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

[OpenAI, 2023] OpenAI. Gpt-4 technical report, 2023.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[Peng *et al.*, 2019] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. In Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors, *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 58–65. Association for Computational Linguistics, 2019.

[Raffel *et al.*, 2019] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

[Rajpoot and Parikh, 2023] Pawan Kumar Rajpoot and Ankur Parikh. Gpt-finre: In-context learning for financial relation extraction using large language models. *CoRR*, abs/2306.17519, 2023.

[Sun *et al.*, 2019] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE 2.0: A continual pre-training framework for language understanding. *CoRR*, abs/1907.12412, 2019.

[Sun *et al.*, 2021] Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *CoRR*, abs/2107.02137, 2021.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023.

[Wan *et al.*, 2023] Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. GPT-RE: in-context learning for relation extraction using large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3534–3547. Association for Computational Linguistics, 2023.

[Wang *et al.*, 2021] Shuohuan Wang, Yu Sun, Yang Xiang, Zhihua Wu, Siyu Ding, Weibao Gong, Shikun Feng, Junyuan Shang, Yanbin Zhao, Chao Pang, Jiaxiang Liu, Xuyi Chen, Yuxiang Lu, Weixin Liu, Xi Wang, Yangfan Bai, Qiuliang Chen, Li Zhao, Shiyong Li, Peng Sun, Dianhai Yu, Yanjun Ma, Hao Tian, Hua Wu, Tian Wu, Wei Zeng, Ge Li, Wen Gao, and Haifeng Wang. ERNIE 3.0 titan: Exploring larger-scale knowledge enhanced pre-training for language understanding and generation. *CoRR*, abs/2112.12731, 2021.

[Wang *et al.*, 2023] Yequan Wang, Hengran Zhang, Aixin Sun, and Xuying Meng. GCRE-GPT: A generative model for comparative relation extraction. *CoRR*, abs/2303.08601, 2023.

[Xu *et al.*, 2023] Bin Xu, Shuai Li, Zhaowu Zhang, and Tongxin Liao. BERT-PAGG: a chinese relationship extraction model fusing PAGG and entity location information. *PeerJ Comput. Sci.*, 9:e1470, 2023.

[Yang *et al.*, 2023a] Ai Ming Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Hai Zhao, Hang Xu, Hao-Lun Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kuncheng Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Pei

Guo, Ruiyang Sun, Zhang Tao, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yan-Bin Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. Baichuan 2: Open large-scale language models. *ArXiv*, abs/2309.10305, 2023.

[Yang *et al.*, 2023b] Jing Yang, Bin Ji, Shasha Li, Jun Ma, Long Peng, and Jie Yu. Dynamic multi-view fusion mechanism for chinese relation extraction. In Hisashi Kashima, Tsuyoshi Idé, and Wen-Chih Peng, editors, *Advances in Knowledge Discovery and Data Mining - 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2023, Osaka, Japan, May 25-28, 2023, Proceedings, Part I*, volume 13935 of *Lecture Notes in Computer Science*, pages 405–417. Springer, 2023.

[Zhang *et al.*, 2021] Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. *CoRR*, abs/2110.06696, 2021.

[Zhao *et al.*, 2019] Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. UER: an open-source toolkit for pre-training models. In Sebastian Padó and Ruihong Huang, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019 - System Demonstrations*, pages 241–246. Association for Computational Linguistics, 2019.

[Zhao *et al.*, 2023] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *CoRR*, abs/2303.18223, 2023.