M-BRe: Discovering Training Samples for Relation Extraction from Unlabeled Texts with Large Language Models

Zexuan Li, Hongliang Dai*, **Piji Li**
[1] College of Artificial Intelligence,
Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2] MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing, China
[3] The Key Laboratory of Brain-Machine Intelligence Technology,
Ministry of Education, Nanjing, China.
{zexuanli, hongldai, pjli}@nuaa.edu.cn

Abstract

For Relation Extraction (RE), the manual annotation of training data may be prohibitively expensive, since the sentences that contain the target relations in texts can be very scarce and difficult to find. It is therefore beneficial to develop an efficient method that can automatically extract training instances from unlabeled texts for training RE models. Recently, large language models (LLMs) have been adopted in various natural language processing tasks, with RE also benefiting from their advances. However, when leveraging LLMs for RE with predefined relation categories, two key challenges arise. First, in a multi-class classification setting, LLMs often struggle to comprehensively capture the semantics of every relation, leading to suboptimal results. Second, although employing binary classification for each relation individually can mitigate this issue, it introduces significant computational overhead, resulting in impractical time complexity for real-world applications. Therefore, this paper proposes a framework called M-BRe to extract training instances from unlabeled texts for RE. It utilizes three modules to combine the advantages of both of the above classification approaches: Relation Grouping, Relation Extraction, and Label Decision. Extensive experiments confirm its superior capability in discovering high-quality training samples from unlabeled texts for RE.

1 Introduction

Relation Extraction (RE) aims to identify specific relation categories between pairs of entities in texts. It is an essential part of information extraction and has been widely used in knowledge mining (Zhong et al., 2024; Wei et al., 2016), Q&A systems (Srihari and Li, 1999), etc. Although existing RE models (Paolini et al., 2021; Zhou and Chen, 2022; Chen et al., 2022b) have performed well on many benchmarks, the scarcity of high-quality training data remains a major problem due to the variety of relation categories in different application domains. Since the sentences that contain the target relation types can be scarce in unlabeled texts, the cost of manually annotating a large training set may become prohibitively expensive when many relation categories are concerned.

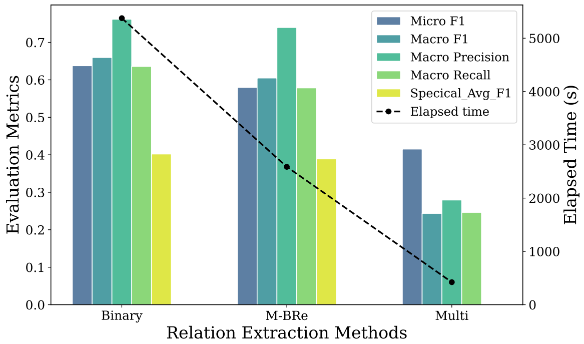


Figure 1: Evaluation metrics across three different Relation Extraction frameworks. The mathematical formulation of each metric is detailed in "§4.2 Evaluation".

To address this problem, some studies focus on zero- or few-shot scenarios, and exploit techniques such as meta-learning (Qu et al., 2020) and prototypical networks (Liu et al., 2022). Apart from these techniques, Large language models (LLMs) have also been employed for implementing RE with limited training data. Two main types of approaches have been investigated. One is to conduct RE directly with LLMs through prompt engineering (Xu et al., 2023). Another is to use LLMs to generate training examples, which can then be used to learn RE models (Liu et al., 2024; Li et al., 2025). In this paper, we also focus on applying LLMs to automatically produce training data for RE. However, instead of direct generation, we prompt LLMs to discover relation instances from unlabeled texts. This can typically be achieved by calling LLMs to perform a relation classification for each sentence in the texts. However, we observe that directly

[*] Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 5220–5238
November 4-9, 2025 ©2025 Association for Computational Linguistics

asking LLMs to conduct multi-class classification would make it difficult for them to understand the semantics of all the relation categories, therefore leading to low-quality prediction results. Another approach is to prompt the model with only one relation type at a time, asking it to perform a binary classification and determine whether this relation exists between the two queried entities in the sentence. However, this approach requires to run the LLM N times for every sentence when there are N target relation categories, which substantially increases the time cost. Figure 1 demonstrates the prediction quality and the time cost of these two approaches.

We therefore focus on how to reduce the time cost for LLMs to annotate the unlabeled sentences, while maintaining the correctness. To this end, we propose the M-BRe framework, which partitions all predefined relation types into multiple groups and ensures that the relations within each group are as distinguishable as possible. For an input example, it performs multi-class classification to differentiate the relations within each group and then further validates each predicted label using binary classification. This two-stage approach enables M-BRe to handle easily distinguishable relations via multi-class classification and more challenging cases via binary classification, thereby combining the strengths of both classification strategies. As shown in Figure 1, M-BRe achieves comparable performance to binary classification while requiring less than half the running time.

To comprehensively evaluate our framework, we conducted extensive experiments on standard relation extraction benchmarks, including SemEval and three variants of TACRED. Our results demonstrate that combining the original few-shot manually labeled samples with the framework-generated training samples significantly improves the performance of conventional RE models. In addition, we apply our approach to Fine-grained Entity Typing, showing its adaptability to other tasks.

Our main contributions are:

- We investigate a novel approach of using LLMs to discover RE training instances from unlabeled texts.

- We propose the M-BRe framework, which offers a novel strategy for LLMs to automatically annotate RE labels with both efficiency and accuracy.

- We conduct comprehensive experiments to demonstrate the effectiveness of the M-BRe framework and the benefit of the extracted training samples for RE.

Our code is available at https://github.com/Lzx-ZBC/M-BRe.

2 Related Work

2.1 Relation Extraction

Relation Extraction (RE) aims to identify relation categories between head and tail entity pairs in text. As a fundamental natural language processing task, it has been traditionally addressed through machine learning approaches such as Bootstrap and Snowball (Batista et al., 2015; Gao et al., 2020). Subsequent advancements in deep learning led to the adoption of pipeline architectures employing CNN, RNN and LSTM (Zeng et al., 2015; Miwa and Bansal, 2016; Zhou and Chen, 2022; Zhang et al., 2017), along with joint End2End framework and graph neural network model (Zhang et al., 2018; Guo et al., 2019, 2020) for RE. Since the emergence of Pre-trained Language Models (PLMs) like BERT, PLMs-based RE models (Devlin et al., 2019; Huang et al., 2019; Moreira et al., 2020; Chen et al., 2022b) have become the dominant paradigm due to their superior performance.

Recent research has demonstrated growing interest in employing Large Language Models (LLMs) for direct RE. Xu et al. (2023) introduced a prompting strategy that incorporates comprehensive relation category definitions and annotated examples, enabling LLMs to better comprehend RE task specifications. Their empirical results confirm that LLMs can generate highly accurate RE predictions. Zhang et al. (2023) developed QA4RE, a novel framework that formalizes RE as a Question Answering (QA) task. Additionally, Li et al. (2023) proposed SUMASK, an advanced prompting technique that reformulates RE through task decomposition into text summarization and QA components, thereby enhancing LLMs' compatibility.

2.2 Data Generation

The scarcity of high-quality data has long been a key factor constraining model performance, making data generation an emerging hotspot. There were already some studies on this topic before instruction tuned LLMs become popular (Ye et al., 2022; Meng et al., 2022; Gao et al., 2023). Meng

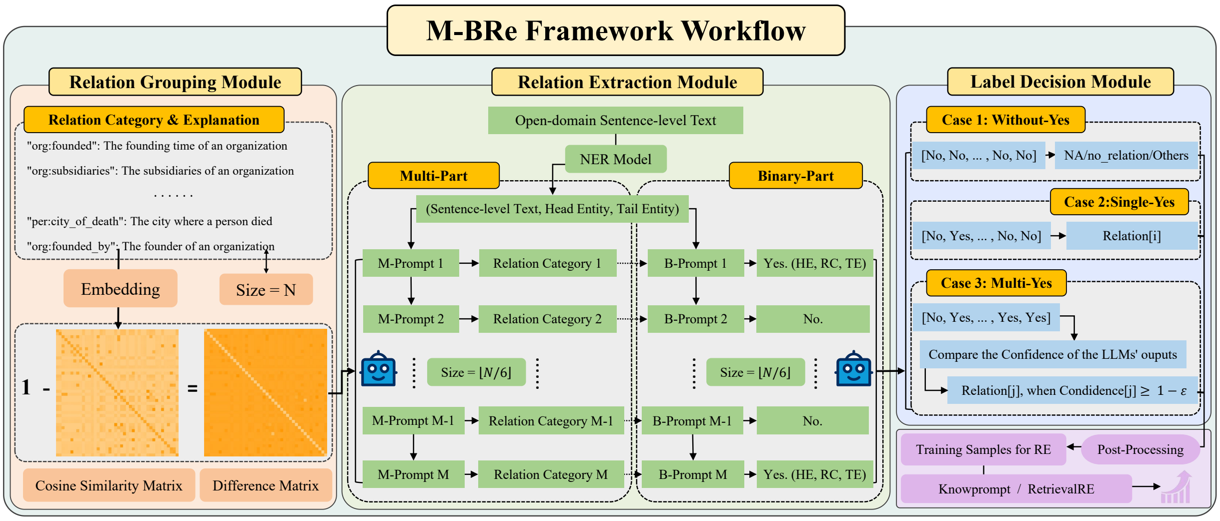


Figure 2: Workflow of the M-BRe Framework, which consists of three modules: Relation Grouping Module, Relation Extraction Module and Label Decision Module.

et al. (2022) prompts the PLMs to create training data for natural language understanding tasks. Chia et al. (2022) presents a framework that leverages language models to generate structured text for synthesizing unseen relation categories. Recently LLMs have demonstrated strong language generation capabilities, providing a viable alternative for synthetic data generation (Hartvigsen et al., 2022; Sahu et al., 2022). Xu et al. (2023) use the LLMs to generate data for assisting the models themselves on RE. Long et al. (2024) pointed out that effective prompts consist of three key elements: Task Specification, Generation Conditions, and In-Context Demonstrations. By clarifying these elements, the accuracy and diversity of generated data can be significantly improved. For complex data generation tasks, Multi-Step Generation has emerged as an important strategy. By breaking down the generation process into multiple simpler subtasks, complex data structures can be generated step by step.

These results show that using LLMs to generate data can effectively enhance the performance of RE models. However, they primarily focus on leveraging the intrinsic knowledge base of LLMs, without attempting to construct data from large-scale real-world unlabeled texts.

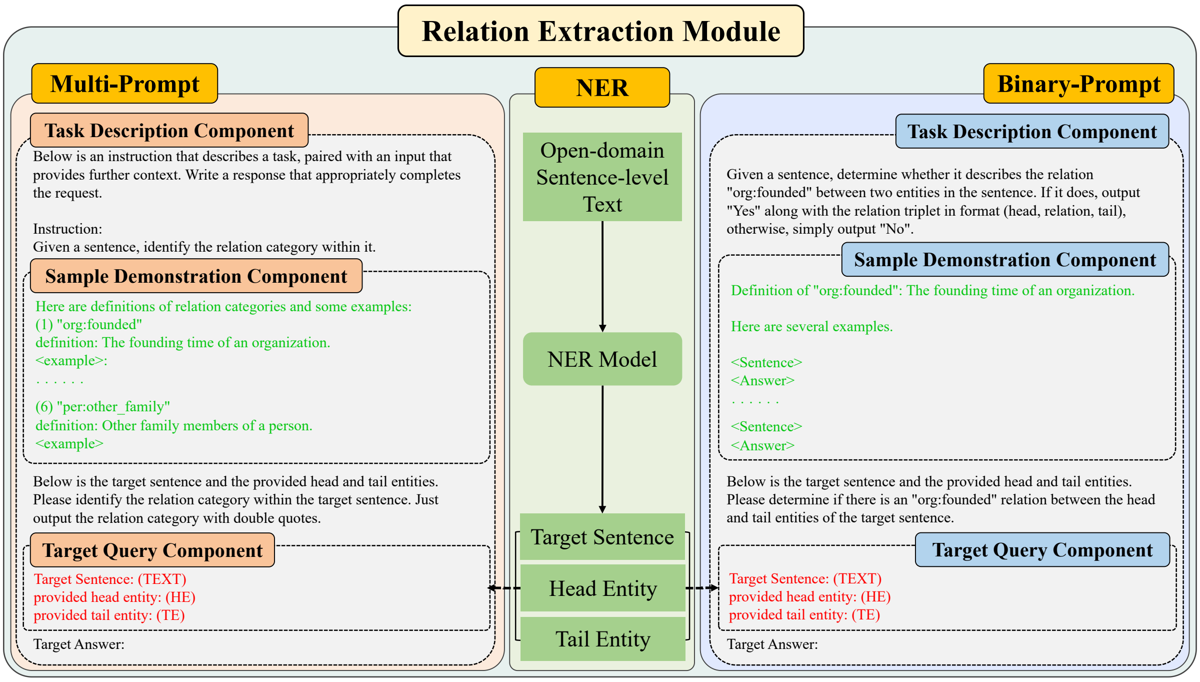
3 The M-BRe Framework

In this section, we present the architectural details of the M-BRe framework, which comprises three key modules as illustrated in Figure 2: Relation Grouping, Relation Extraction, and Label Decision.

The framework starts by partitioning all the predefined relation types into K groups with *Relation Grouping Module*, where the relations in each group should be as distinguishable as possible. Consequently, it decomposes the single large-scale multi-class classification task into K smaller-scale multi-class classification subtasks. For an input example, each subtask is completed with the multi-class classification prompt of *Relation Extraction Module*, thereby yielding K relation labels. To further select from these K labels, we consider each of them individually, and leverage the binary classification prompt of *Relation Extraction Module* K times to infer whether each of them is a valid label. Finally, *Label Decision Module* is used to determine the final labels for the input example.

3.1 Relation Grouping Module

We first construct an explanation for each predefined relation category in the RE dataset that follows the form “"org:founded": The founding time of an organization.” The details of the explanations for all relation categories are provided in Appendix A.4. We then vectorize them and compute their cosine similarities, deriving a matrix that quantifies the semantic disparities between different relation labels. To partition the relations based on this matrix, the two most dissimilar relation categories are used as initial seed groups. Then, the remaining relations are iteratively assigned to the group that minimizes the maximum similarity of the group using a greedy strategy, while keeping the size of each group balanced. The pseudo code for this process is shown in Appendix C.


Relation Extraction Module
Multi-Prompt
NER
Binary-Prompt
Task Description Component
Sample Demonstration Component
Target Query Component
Open-domain Sentence-level Text
NER Model
Target Sentence
Head Entity
Tail Entity
Task Description Component
Sample Demonstration Component
Target Query Component

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Given a sentence, identify the relation category within it.

Here are definitions of relation categories and some examples:
(1) "org:founded"
definition: The founding time of an organization.
<example>:
. . . . . .
(6) "per:other_family"
definition: Other family members of a person.
<example>

Below is the target sentence and the provided head and tail entities. Please identify the relation category within the target sentence. Just output the relation category with double quotes.

Target Sentence: (TEXT)
provided head entity: (HE)
provided tail entity: (TE)

Target Answer:

Given a sentence, determine whether it describes the relation "org:founded" between two entities in the sentence. If it does, output "Yes" along with the relation triplet in format (head, relation, tail), otherwise, simply output "No".

Definition of "org:founded": The founding time of an organization.

Here are several examples.

<Sentence>
<Answer>
. . . . . .
<Sentence>
<Answer>

Below is the target sentence and the provided head and tail entities. Please determine if there is an "org:founded" relation between the head and tail entities of the target sentence.

Target Sentence: (TEXT)
provided head entity: (HE)
provided tail entity: (TE)

Target Answer:

Figure 3: Construction of Multi-Prompt and Binary-Prompt, which consists of three component: Task Description Component, Sample Demonstration Component and Target Query Component.

Given N predefined relation categories, we set the number of groups to $\lfloor N/6 \rfloor$. This grouping strategy and its empirical validation are systematically analyzed in "§5.3 Number of Groups".

3.2 Relation Extraction Module

This module focuses on the prompt design for both multi-class classification and binary classification. Both prompts consist of three fundamental components: Task Description, Sample Demonstration, and Target Query. The details of the prompts are shown in Figure 3.

Task Description Component. This component consists of two basic parts. At the beginning of the prompt, we instruct the LLM that the objective is to identify the entity relation within a sentence. Then, following a few demonstrations, the model is told that the target sentence along with given head-tail entity pairs will be provided. The output format is also explicitly specified. In multi-class classification settings, the LLMs are required to predict the relation category directly. For binary classification, the LLMs are expected to output "Yes. (Head Entity, Relation Category, Tail Entity)" if they think that the current relation category does exist in the input unlabeled sentence, otherwise they should output "No.".

Sample Demonstration Component. We use each grouping result obtained in "§3.1 Relation Grouping Module" to construct the multi-class classification prompt, and employ In-Context Learning (ICL) (Brown et al., 2020; Wan et al., 2023) to provide definition and demonstrations for each relation category in the group, so that LLMs can distinguish the relation types within the group more easily. The binary classification prompt also applies ICL to provide 3 correct and 4 incorrect examples for the current relation category.

Target Query Component. We extract all entities from unlabeled sentence-level text and randomly sample two of them to form a head-tail entity pair. Finally, it can be injected into the prompt, which will enhance the LLMs' ability to discern potential relation categories.

3.3 Label Decision Module

With "§3.2 Relation Extraction Module", we will get $\lfloor N/6 \rfloor$ "Yes" or "No" results. In order to determine the final relation extraction results, a Confidence-based label decision strategy is introduced to handle the following three scenarios that may occur. The confidence for LLM-generated content is formulated as follows:

$$\mathrm{Confidence} = \frac{1}{\mathrm{M}} \sum_{\mathrm{i}=1}^{\mathrm{M}} \max(\mathrm{softmax}(\mathrm{logits_i})),$$

where M is the number of tokens generated.

Case 1: Without-Yes. The result of binary classification are all "No.", indicating that the LLMs consider that the unlabeled text does not have any relation categories, so we set the relation category between the head-tail entity pairs of the current unlabeled text to "NA" or "no_relation" or "Others".

Case 2: Single-Yes. The result of the binary classification is only one "Yes. (Head Entity, Relation Category, Tail Entity)", indicating that the LLMs consider that one of $\lfloor N/6 \rfloor$ relation categories, R_1, exists in the current unlabeled text, and we set the relation category between the head-tail entity pairs of the current unlabeled text to R_1.

Case 3: Multi-Yes. There are several "Yes. (Head Entity, Relation Category, Tail Entity)" results of the binary classification, indicating that the LLMs consider that there are more than one of $\lfloor N/6 \rfloor$ relation categories in the current unlabeled text, $[\mathrm{R_1}, \mathrm{R_2}, ..., \mathrm{R_i}]$. We calculate the confidence of LLMs' binary classification predictions for each relation category and set the relation categories between the head-tail entity pairs of the current unlabeled text to $[\mathrm{R_1}, \mathrm{R_2}, ..., \mathrm{R_j}]$ $(\mathrm{j} \leq \mathrm{i})$, where $\mathrm{Confidence}\ (\mathrm{R_j}) \geq 1 - \epsilon,\ \epsilon = 10^{-2}$.

In our current implementation, we set the confidence threshold to a fixed value $\theta = 10^{-2}$ based on preliminary experiments and empirical observations, which showed that this value provided a good balance between precision and recall in our label decision process.

- If we significantly reduce the $1 - \theta$ value dynamically, the subsequently constructed training samples would include relation categories less suitable for the current entity pairs. The more the $1 - \theta$ value decreases, the greater the noise in the training samples regarding relation categories.

- If we only make minimal dynamic adjustments to the $1 - \theta$ value, the results would be close to those obtained with the current $1 - \theta$ setting.

To give a better visualization of this phenomenon, we dynamically adjusted θ to take values in [0.001, 0.01, 0.02, 0.05, 0,1] and the experimental results are shown in Tabel 1. It can be seen that, in most settings, variations of θ in the range of [0.001, 0.01, 0.02, 0.05, 0,1] do not have a sensitive effect on the M-BRe framework, especially when 14B models are used. In general, our setting of $\theta = 10^{-2}$ allows the approach to achieve good performance results in most experimental setups.

Dataset	Tacred			SemEval		
θ	Qwen2.5-7B-Instruct-1M			Qwen2.5-7B-Instruct-1M		
	Mi-F1	Ma-F1	S_A_F1	Mi-F1	Ma-F1	S_A_F1
0.001	55.07	58.80	36.28	34.07	31.73	25.64
0.01	53.14	56.08	35.05	43.66	43.36	33.68
0.02	56.04	60.28	33.65	47.25	46.59	35.53
0.05	59.90	61.98	32.72	43.96	42.72	28.57
0.1	58.94	60.16	33.17	36.26	34.91	23.44
θ	Qwen2.5-14B-Instruct-1M			Qwen2.5-14B-Instruct-1M		
	Mi-F1	Ma-F1	S_A_F1	Mi-F1	Ma-F1	S_A_F1
0.001	59.90	63.71	34.37	52.75	51.77	37.55
0.01	58.94	63.95	33.86	52.75	51.14	38.64
0.02	58.12	62.48	33.93	54.95	52.95	36.63
0.05	59.08	62.55	30.76	56.04	56.47	34.80
0.1	61.50	65.08	31.39	54.95	54.79	36.26
θ	Qwen3-14B			Qwen3-14B		
	Mi-F1	Ma-F1	S_A_F1	Mi-F1	Ma-F1	S_A_F1
0.001	75.36	76.72	40.75	20.88	20.88	16.67
0.01	76.33	77.46	39.58	21.98	22.28	16.85
0.02	76.33	77.81	37.79	19.78	19.12	15.56
0.05	76.33	77.05	36.51	19.78	19.16	14.84
0.1	77.78	78.97	37.60	17.58	16.05	12.27

Table 1: Performance results of the M-BRe framework under different threshold θ. Mi-F1, Ma-F1 and S_A_F1 represent Micro F1, Macro F1 and Special_Avg_F1.

4 Experimental Settings

4.1 Datasets

We conducted experiments on SemEval and three versions of TACRED: SemEval 2010 Task 8 (SemEval) (Hendrickx et al., 2010), TACRED (Zhang et al., 2017), TACRED-Revisit (Alt et al., 2020), Re-TACRED (Stoica et al., 2021). Statistical details are given in Appendix A.1.

4.2 Evaluation

For M-BRe Framework. We randomly select 5 test samples for each relation category from the TACRED and SemEval test datasets. Since some relation categories have fewer than 5 test samples, the final number of samples used for testing is 207 for TACRED and 91 for SemEval. As long as the predicted relation list includes the ground truth, the prediction of the M-BRe framework is considered correct. In this way, Micro F1, Macro F1, Macro precision, Macro recall, and Elapsed time are used as preliminary evaluation metrics. Meanwhile, in order to reflect the absoluteness of the correct pre-

	Dataset / Method	TACRED			TACRED-Revisit			Re-TACRED			SemEval		
		K=2	K=4	K=8	K=2	K=4	K=8	K=2	K=4	K=8	K=2	K=4	K=8
Qwen2.5-7B-Instruct-1M	Knowprompt	5.91	11.24	22.07	12.44	14.04	26.31	12.35	14.48	44.77	15.71	38.18	61.39
	Mix 4	14.95	18.73	28.20	19.72	20.85	27.36	20.89	34.73	39.34	32.97	43.74	66.98
	Mix P	**20.32**	**25.36**	**30.13**	**24.52**	**25.59**	**28.96**	**27.48**	**35.07**	**53.25**	**37.81**	**44.71**	**68.13**
	RetrievalRE	15.34	13.34	28.14	17.35	19.53	28.95	9.76	17.54	33.72	38.95	52.92	72.53
	Mix 4	16.71	20.56	27.25	21.18	26.56	29.61	18.10	25.19	35.13	40.12	53.27	68.94
	Mix P	**20.44**	**25.37**	**30.30**	**22.25**	**26.84**	**31.27**	**21.72**	**26.44**	**36.30**	**42.71**	**56.89**	**73.33**
Qwen2.5-14B-Instruct-1M	Knowprompt	5.91	11.24	22.07	12.44	14.04	26.31	12.35	14.48	44.77	15.71	38.18	61.39
	Mix 4	12.26	19.30	25.65	14.55	19.18	26.06	19.03	26.55	36.28	34.28	38.33	45.18
	Mix P	**16.76**	**21.94**	**27.05**	**17.44**	**21.41**	**30.00**	**20.20**	**41.86**	**50.97**	**44.88**	**54.37**	**71.38**
	RetrievalRE	15.34	13.34	28.14	17.35	19.53	28.95	9.76	17.54	33.72	38.95	52.92	72.53
	Mix 4	17.19	21.94	26.71	19.40	23.34	30.02	20.78	25.16	39.00	50.77	59.01	64.60
	Mix P	**18.80**	**23.78**	**28.91**	**21.51**	**26.48**	**30.25**	**21.24**	**25.84**	**40.79**	**56.09**	**61.00**	**76.20**
Qwen3-14B	Knowprompt	5.91	11.24	22.07	12.44	14.04	26.31	12.35	14.48	44.77	15.71	38.18	61.39
	Mix 4	19.14	25.34	29.24	23.30	27.01	30.10	22.67	33.46	48.25	31.43	33.01	53.98
	Mix P	**21.14**	**25.96**	**30.12**	**24.77**	**29.22**	**31.23**	**27.66**	**38.84**	**51.15**	**35.44**	**45.89**	**62.67**
	RetrievalRE	15.34	13.34	28.14	17.35	19.53	28.95	9.76	17.54	33.72	38.95	52.92	72.53
	Mix 4	16.34	21.82	27.81	21.04	21.78	29.46	18.39	23.86	40.02	40.98	53.41	66.76
	Mix P	**19.73**	**22.14**	**28.64**	**21.54**	**25.69**	**29.72**	**21.19**	**31.52**	**47.46**	**41.18**	**58.54**	**72.54**

Table 2: Micro F1 (%) of few-shot performance. **Knowprompt** and **RetrievalRE** mean the performance of manually labeled training samples only. **Mix 4** and **Mix P** mean the performance of combining the use of manually labeled training samples and constructed training samples when relation categories are divided into 4 and P groups, where P = $\lfloor N/6 \rfloor$ and N is the total number of relation categories for each RE dataset.

LLMs	Model	Tacred	Tacrev	Retacred	SemEval
Qwen2.5-7B-Instruct-1M	Kp.	9.80	11.61	14.03	19.15
		17.52	**19.03**	**20.36**	**19.70**
	Rt.	13.23	15.10	13.82	18.97
		16.57	**20.60**	**17.53**	**19.46**
Qwen2.5-14B-Instruct-1M	Kp.	9.58	11.43	11.34	28.99
		9.64	**15.60**	**15.31**	**37.64**
	Rt.	12.92	13.86	12.04	31.98
		17.69	**16.22**	**14.13**	**32.77**
Qwen3-14B	Kp.	13.42	14.69	13.68	11.88
		17.99	**16.67**	**19.54**	**26.53**
	Rt.	12.50	11.77	9.77	11.56
		16.53	**16.38**	**15.13**	**29.99**

Table 3: Micro F1 (%) of few-shot performance using constructed training samples only. The first and second line of each model correspond to **Pure 4** and **Pure P**.

diction of the M-BRe framework, we introduce the Special_Avg_F1 metric, which takes into account the length of the predicted relation list. The formula is specified as follows:

$$\mathrm{Precison_i} = \frac{|\cap(\mathrm{PredictionSet_i}, \mathrm{ReferenceSet_i})|}{|\mathrm{PredictionSet_i}|},$$

$$\mathrm{Recall_i} = \frac{|\cap(\mathrm{PredictionSet_i}, \mathrm{ReferenceSet_i})|}{|\mathrm{ReferenceSet_i}|},$$

$$\mathrm{Special\_Avg\_F1} = \frac{1}{N}\sum_{i=0}^{N} \frac{2 \cdot \mathrm{Precision}_i \cdot \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i}$$

where for each test sample i, $\mathrm{Precision_i}$ and $\mathrm{Recall_i}$ take 10^{-10} only if the intersection of $\mathrm{PredictionSet_i}$ and $\mathrm{ReferenceSet_i}$ is $\varnothing$.

For Generated Training Samples. Considering that the quality of the generated training samples is difficult to directly assess, we put the generated training samples into KnowPrompt (Chen et al., 2022b) and RetrievalRE (Chen et al., 2022a) for training. KnowPrompt achieves satisfying performance through Knowledge Injection and Synergistic Optimization, while RetrievalRE improves the model's generalization ability by combining retrieval enhancement and prompt tuning. Their performance on the test dataset reflects the quality of the framework-generated training samples. The better the performance of them, the higher the quality of the framework-generated training samples. Finally, we follow the existing RE studies and adopt Micro F1 as the main evaluation metric.

Meanwhile, in order to further evaluate the quality of the relation extraction training samples generated by the M-BRe framework, we reformat them into Supervised Fine-Tuning (SFT) datasets for multi-class classification to fine-tune LLMs. The

Method \ Dataset		TACRED			TACRED-Revisit			Re-TACRED			SemEval		
		K=2	K=4	K=8	K=2	K=4	K=8	K=2	K=4	K=8	K=2	K=4	K=8
Mix 4	Random	14.65	18.49	27.25	16.49	16.83	25.81	18.01	30.77	37.28	29.12	37.99	36.58
	Ours	**14.95**	**18.73**	**30.46**	**19.72**	**20.85**	**27.36**	**20.89**	**34.73**	**39.34**	**32.97**	**43.74**	**66.98**
Mix P	Random	19.08	23.06	26.01	21.43	21.77	27.11	22.65	26.63	42.95	21.66	23.87	29.93
	Ours	**20.32**	**25.36**	**30.13**	**24.52**	**25.59**	**28.96**	**27.48**	**35.07**	**53.25**	**37.81**	**44.71**	**68.13**

Table 4: Micro F1 (%) of different algorithm of M-BRe framework on KnowPrompt. **Random** and **Ours** mean random and Algorithm 1-based Relation Grouping. **Mix** means hybrid training samples combining constructed and manually annotated samples, where P = $\lfloor N/6 \rfloor$ and N is the total number of relation categories for each RE dataset.

Method		Tacred	Tacrev	Retacred	SemEval
Pure 4	Random	7.71	9.25	8.38	17.05
	Ours	**9.80**	**11.61**	**14.03**	**19.15**
Pure P	Random	12.21	12.66	16.39	8.63
	Ours	**17.52**	**19.03**	**20.36**	**19.70**

Table 5: Micro F1 (%) of different algorithm of M-BRe framework on KnowPrompt. **Pure** means purely constructed training samples.

observed improvement in the LLMs' multi-class classification performance post-SFT confirms the effectiveness of the generated samples.

5 Results and Analysis

5.1 Main Results

For M-BRe framework, we adopt 4-grouping and $\lfloor N/6 \rfloor$-grouping strategies to extract the relation categories from unlabeled sentence-level texts, which are then used to construct RE training samples. The sample sizes of RE datasets constructed by each LLM on identical unlabeled sentence-level texts are summarized in Appendix A.3. The main results are detailed in Table 2 and Table 3.

Comparing to Manually Labeled Training Samples. In Table 3, Pure 4 and Pure P mean only using framework-constructed samples for training, and they differ in the number of groupings in "Multi-Part", while Knowprompt and RetrievalRE mean only using manually labeled training samples in Table 2. K denotes the number of manually labeled training samples for each relation category. The performance of Pure 4 and Pure P on all datasets can only match or exceed that of KnowPrompt and RetrievalRE when K = 2 or 4. We first analyze that the unlabeled sentence-level texts have not undergone a thorough data cleaning and screening process. The second point is that the distribution of relation categories extracted by the M-BRe framework is highly imbalanced, exhibiting a long-tail issue.

Comparing to Mixed Training Samples. We mix manually labeled training samples with constructed training samples and use them to train RE models. This corresponds to Mix 4 and Mix P in Table 2. We observe that RE models' performance is higher than when only pure manually labeled training samples or constructed training samples are used under all settings. The results indicate that incorporating the constructed training samples with existing manually labeled training samples can strongly enhance the performance of RE models. However, the efficacy of the constructed training samples diminishes as the volume of manually labeled training data increase. Our analysis suggests that synthetically constructed training samples can effectively improve RE models' understanding of relation categories when manually labeled training data are scarce. As the volume of manually labeled data increase, the quality variability of the generated training samples introduce noise, due to uncleaned text sources. This results in potential misinterpretation of relation categories by RE models, ultimately leading to a slow performance improvement when using mixed training data.

Comparing to Number of Groups. We establish relation category groupings with sizes 4 and P, corresponding to Mix 4 and Mix P in Table 2, along with Pure 4 and Pure P in Table 3. The experimental results demonstrate that Pure P consistently outperforms Pure 4 across all datasets. We attribute this superiority to Pure P's larger number of groupings, which provides greater discriminability among relation categories, thereby enabling LLMs to make more accurate judgments. Moreover, under all few-shot settings, Mix P outperforms Mix 4, although the performance gap diminishes with increasing K. We attribute this to the reduced con-

Method	Qwen2.5-7B-Instruct-1M				Qwen2.5-14B-Instruct-1M			
	Macro-P	Macro-R	Macro-F1	Micro-F1	Macro-P	Macro-R	Macro-F1	Micro-F1
Base	41.54	27.94	24.64	24.37	57.00	53.25	47.80	47.36
Only Generated	50.72	47.64	45.32	43.95	61.84	57.37	58.22	56.22
Only Manual	64.97	61.59	60.33	64.73	71.46	70.20	68.00	66.85
Manual→Generated	58.94	56.87	58.14	55.73	64.73	66.38	63.72	61.42
Generated→Manual	**69.21**	**66.36**	**64.82**	**69.08**	**73.43**	**76.47**	**73.81**	**72.32**

Table 6: Micro F1 (%) of each LLM with different SFT routes. **Base** means LLM without SFT. **Only Generated** means only using LLMs-generated training samples for SFT. **Only Manual** means only using manually labeled training samples for SFT. **Manual→Generated** means using LLMs-generated training samples first and then manually labeled training samples for SFT. **Generated→Manual** means using manually labeled training samples first and then LLMs-generated training samples for SFT.

tribution of LLMs-generated training samples in RE model training as more manually labeled data become available, where certain inevitable noise even exerts negative effects on model training. This phenomenon demonstrate that the number of groupings significantly impacts RE performance of M-BRe framework. Therefore, we conduct a comprehensive experimental analysis into the effect of grouping quantity in “§5.3 Number of Groupings”.

5.2 Ablation Analysis

In order to verify the effectiveness of Relation Groupings Algorithm, we conduct comprehensive ablation experiments: relation categories were divided into 4 and P groups using both the Relation Grouping Algorithm and other grouping ways. As demonstrated in Figure 5 and Appendix B.3, our method significantly enhances the relation extraction capability of the M-BRe framework across all evaluation metrics, while simultaneously achieving faster processing speed than the random grouping approach.

To further validate the effectiveness of our approach, we employed random-based M-BRe framework on the same unlabeled sentence-level texts to extract relation categories and construct RE training samples with Qwen2.5-7B-Instruct-1M. These samples were subsequently utilized to train KnowPrompt for comparative performance evaluation. As shown in Table 4 and 5, Ours (Pure 4, P) consistently outperform Random (Pure 4, P), demonstrating that the constructed RE training samples with Algorithm 1-based Relation Grouping achieve higher quality than those with random Relation Grouping, under both grouping sizes of 4 and P. Ours (Mix 4, P) also consistently exhibit superior performance to Random (Mix 4, P) in all settings, further validating the aforementioned conclusions.

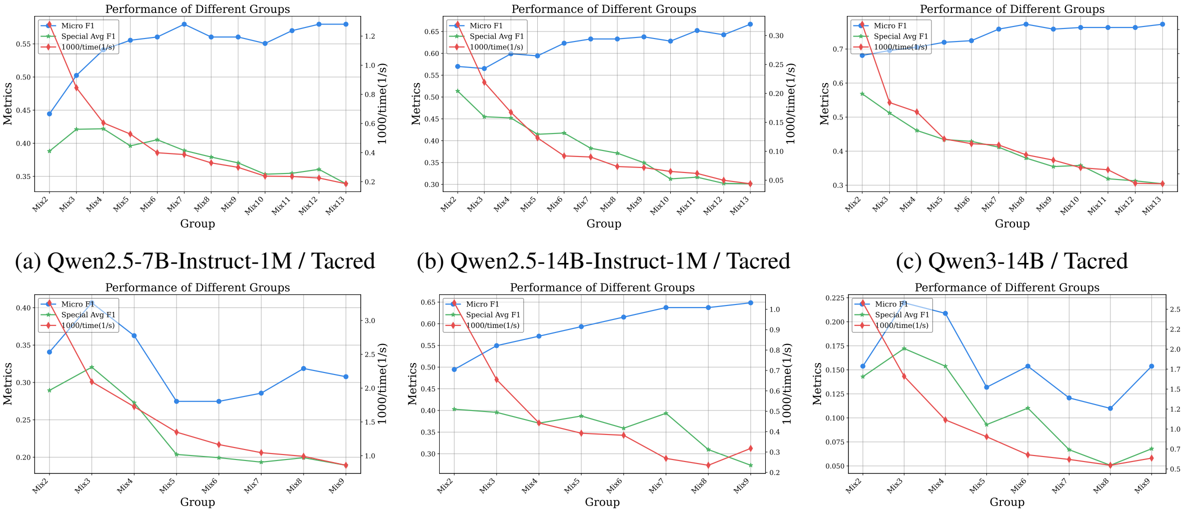
5.3 Number of Groups

Evidently, the number of relation category groups significantly impacts the overall performance of the M-BRe framework. Therefore, we conducted comprehensive experiments while controlling this variable as shown in Figure 4 and Appendix B.1. The relation extraction performance of the M-BRe framework gradually improves with increasing number of groups. At $\lfloor N/6 \rfloor$ groups, the performance essentially peaks, matching that of the binary classification approach while achieving more than twice the processing speed. Our analysis shows that at $\lfloor N/6 \rfloor$ groups, the M-BRe framework combines the advantages of both multi-class and binary classification, enabling rapid relation extraction from unlabeled sentence-level texts while maintaining accuracy. However, the framework's performance begins to gradually degrade as the number of groupings continues to increase. We attribute this to the lengthened output list in the Multi-prompt phase as group numbers increase, leading to accumulated error propagation. This phenomenon demonstrates the existence of an inflection point for group numbers, confirming that more groups do not necessarily yield better results.

5.4 Cold Start for SFT

We reconstruct the training samples generated by the M-BRe framework and the 8-shot manually labeled training samples into an SFT dataset, then fine-tune Qwen2.5-7B-Instruct-1M and Qwen2.5-14B-Instruct-1M through the following four routes:

- Using only LLMs-generated training samples.
- Using only manually labeled training samples.
- Using LLMs-generated training samples first and then manually labeled training samples.



(a) Qwen2.5-7B-Instruct-1M / Tacred (b) Qwen2.5-14B-Instruct-1M / Tacred (c) Qwen3-14B / Tacred

(d) Qwen2.5-7B-Instruct-1M / SemEval (e) Qwen2.5-14B-Instruct-1M / SemEval (f) Qwen3-14B / SemEval

Figure 4: Micro F1 (%) and Special_Avg_F1 (%) of different number of groupings on each LLM.

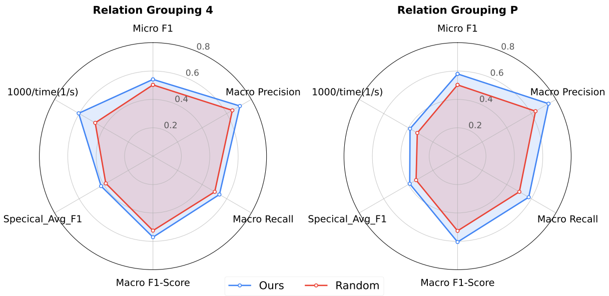


Figure 5: Comparative evaluation of the two methods' effects on M-BRe framework performance. **Ours** and **Random** represent Algorithm 1-based and random Relation Grouping in the M-BRe framework.

- Using manually labeled training samples first and then LLMs-generated training samples.

As shown in Table 6, all SFT routes significantly improve the performance of LLMs on direct multi-class relation extraction, further demonstrating the effectiveness of LLMs-generated training samples. Admittedly, the performance after SFT with manually labeled training samples surpasses that with LLMs-generated training samples, which we attribute to the two major characteristics of LLMs-generated training samples mentioned in "§5.1 Main Results". Meanwhile, it is encouraging to observe that the performance after two-stage fine-tuning (first with LLMs-generated then manually labeled samples) slightly outperforms using only manually labeled samples. We posit that the LLMs-generated training samples serve as a cold start for the second-stage fine-tuning, thereby effectively enhancing LLMs' performance.

6 Conclusion

In this paper, we propose a novel M-BRe framework for constructing RE training samples with LLMs. This framework effectively combines the advantages of multi-class and binary classification to efficiently utilize unlabeled texts to acquire sentence-level RE training samples, particularly through the Relation Grouping Module and the Label Decision Module. The former enables LLMs to rapidly comprehend and distinguish different relations while making an accurate classification judgment. The latter allows LLMs to verify their judgments while accommodating cases where single head-tail entity pair may correspond to multiple valid relations. Experimental results prove the effectiveness of RE training samples constructed by the M-BRe framework in few-shot scenarios.

In future work, we plan to explore the following directions: (1) more rational and effective approaches for Relation Grouping; (2) alternative methods for relation category judgment that outperform Confidence-based approaches.

Limitations

Despite our best efforts, the proposed M-BRe framework in this paper still has several limitations.

Long-Tail Issue: Constrained by the nature of unlabeled sentence-level texts, the distribution of RE training samples constructed by B-MRe framework exhibits a long-tail issue, where the scarcity of instances for certain relation categories may limit the performance of RE models.

LLMs: Although we have enabled the M-BRe framework to efficiently construct RE training samples, the quality of these constructed samples remains significantly influenced by the inherent capabilities of the open-source LLMs themselves.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 62306140, No. 62476127), the Natural Science Foundation of Jiangsu Province (No. BK20242039), the Basic Research Program of the Bureau of Science and Technology (ILF24001), the Fundamental Research Funds for the Central Universities (No. NJ2023032), the Scientific Research Starting Foundation of Nanjing University of Aeronautics and Astronautics (No. YQR21022), the Key Project of Jiangsu Collaborative Innovation Center of Chinese Medicinal Resources Industrialization (No. 000003401025-6), the Open Project of Chinese Materia Medica First-Class Discipline of Nanjing University of Chinese Medicine (No. ZYXJC2024-010) and the High Performance Computing Platform of Nanjing University of Aeronautics and Astronautics.

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1558–1569. Association for Computational Linguistics.

David S. Batista, Bruno Martins, and Mário J. Silva. 2015. Semi-supervised bootstrapping of relationship extractors with distributional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 499–504. The Association for Computational Linguistics.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xiang Chen, Lei Li, Ningyu Zhang, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022a. Relation extraction as open-book examination: Retrieval-enhanced prompt tuning. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2443–2448. ACM.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022b. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2778–2788. ACM.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022. Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 45–57. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jiahui Gao, Renjie Pi, Yong Lin, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023. Self-guided noise-free data generation for efficient zero-shot learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Tianyu Gao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2020. Neural snowball for few-shot relation learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7772–7779. AAAI Press.

Zhijiang Guo, Guoshun Nan, Wei Lu, and Shay B. Cohen. 2020. Learning latent forests for medical relation extraction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3651–3657. ijcai.org.

Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. In *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 241–251. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3309–3326. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010*, pages 33–38. The Association for Computer Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Weipeng Huang, Xingyi Cheng, Taifeng Wang, and Wei Chu. 2019. Bert-based multi-head selection for joint entity-relation extraction. In *Natural Language Processing and Chinese Computing - 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9-14, 2019, Proceedings, Part II*, volume 11839 of *Lecture Notes in Computer Science*, pages 713–723. Springer.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023. Revisiting large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 6877–6892. Association for Computational Linguistics.

Zexuan Li, Hongliang Dai, and Piji Li. 2025. Generating diverse training samples for relation extraction with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 713–726. Association for Computational Linguistics.

Yang Liu, Jinpeng Hu, Xiang Wan, and Tsung-Hui Chang. 2022. Learn from relation information: Towards prototype representation rectification for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1822–1831. Association for Computational Linguistics.

Ye Liu, Kai Zhang, Aoran Gan, Linan Yue, Feng Hu, Qi Liu, and Enhong Chen. 2024. Empowering few-shot relation extraction with the integration of traditional RE methods and large language models. In *Database Systems for Advanced Applications - 29th International Conference, DASFAA 2024, Gifu, Japan, July 2-5, 2024, Proceedings, Part V*, volume 14854 of *Lecture Notes in Computer Science*, pages 349–359. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 11065–11082. Association for Computational Linguistics.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *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*.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Johny Moreira, Chaina Oliveira, David L. Macêdo, Cleber Zanchettin, and Luciano Barbosa. 2020. Distantly-supervised neural relation extraction with side information using BERT. In *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pages 1–7. IEEE.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Meng Qu, Tianyu Gao, Louis-Pascal A. C. Xhonneux, and Jian Tang. 2020. Few-shot relation extraction via bayesian meta-learning on relation graphs. In

Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of *Proceedings of Machine Learning Research*, pages 7867–7876. PMLR.

Gaurav Sahu, Pau Rodríguez, Issam H. Laradji, Parmida Atighehchian, David Vázquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI, ConvAI@ACL 2022, Dublin, Ireland, May 27, 2022*, pages 47–57. Association for Computational Linguistics.

Rohini K. Srihari and Wei Li. 1999. Information extraction supported question answering. In *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999*, volume 500-246 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the TACRED dataset. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13843–13850. AAAI Press.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: in-context learning for relation extraction using large language models. In *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.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. *Database*, 2016:baw032.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. 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 2023, Toronto, Canada (Hybrid), July 13, 2023*, pages 190–200. Association for Computational Linguistics.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. 2025. Qwen2.5-1m technical report. *CoRR*, abs/2501.15383.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Zerogen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11653–11669. Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1753–1762. The Association for Computational Linguistics.

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 794–812. Association for Computational Linguistics.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2205–2215. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 35–45. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. 2024. A comprehensive survey on automatic knowledge graph construction. *ACM Comput. Surv.*, 56(4):94:1–94:62.

Wenxuan Zhou and Muhao Chen. 2022. 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, AACL/IJCNLP 2022 - Volume 2: Short Papers, Online only, November 20-23, 2022*, pages 161–168. Association for Computational Linguistics.

A Experimental Details

A.1 Datasets

For comprehensive experiments, we conducted experiments on four relation extraction datasets: TACRED, TACRED-Revisit, Re-TACRED and SemEval 2010 Task 8 (SemEval). The statistics of the RE datasets are shown in Table 7. A brief introduction to these data is given below:

TACRED: a large-scale sentence-level relation extraction dataset from the annual TACBP4 challenge, containing over 106,000 sentences. It involves 42 different relation categories, including 41 common relation categories and a special "no relation" relation category.

TACRED-Revisit: a dataset constructed on the basis of the TACRED dataset. The researchers found errors in the development and test sets of the original TACRED dataset and corrected them while keeping the training set intact.

Re-TACRED: another version of the TACRED dataset, which addresses some of the shortcomings of the original TACRED dataset by reconstructing the training, validation and test sets. Meanwhile, this dataset removes the original 6 relation categories and adds 4 new relation categories to the TACRED dataset, so that a dataset with 40 relation categories is finally obtained.

SemEval: a traditional relation extraction dataset, containing 10,717 annotated samples, covers 9 bi-directional relation categories and a special "no relation" relation category.

Dataset	Train	Val	Test	Relation
SemEval	6,507	1,493	2,717	19
TACRED	68,124	22,631	15,509	42
TACRED-Revisit	68,124	22,631	15,509	42
Re-TACRED	58,465	19,584	13,418	40

Table 7: Statistics of the RE datasets. Including the numbers of instances in different splits and the numbers of relation categories.

A.2 Implementation Details

For KnowPrompt and RetrievalRE. We follow (Chen et al., 2022b,a) and use RoBERTA_LARGE (Liu et al., 2019) in all experiments for comparison.

For Large Language Models. Considering the cost requirements and the strength of the different LLMs themselves, we used Qwen2.5-7B-Instruct-1M, Qwen2.5-14B-Instruct-1M (Yang et al., 2025), Qwen3-14B in our experiments, setting temperature = 0.6. For post-processing, since the unlabeled text has not been filtered, the following two data distributions appear after extracting relation categories with the B-MRe framework: (1) Most sentence-level texts are meaningless, so the number ratio of "NA" to "non-NA" is relatively large. (2) In sentence-level text, the frequency of different "non-NA" relation categories is different, so their number shows a long-tailed distribution. To combat this, we calculated the average number of "non-NA" relation categories. This number of samples are randomly selected from the "NA" relation category and combined with the "non-NA" relation category to form the final training data.

For Supervised Fine-Tuning. All models are fine-tuned using the LLaMA-Factory framework (Zheng et al., 2024) on 8 NVIDIA RTX 3090 GPUs. We employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) with rank $r = 16$ applied to all linear layers. Other hyperparameters include a batch size of 4, a learning rate of 10^{-4}, and 3 training epochs.

A.3 Statistics on the number of samples constructed by each LLM

The number of unlabeled sentences used in our experiments is 4401. Tabel 8 details the number of generated RE training samples for each LLM, it can be seen that the existence of relations within unlabeled sentences is indeed very sparse. Table 9 details the number of constructed RE training samples for each LLM. Our analysis reveals that individual LLMs with stronger performance exhibit enhanced capability in constructing RE training samples from unlabeled text, whereas the grouping factor demonstrates no statistically significant impact on the yield of RE training instances.

A.4 Relation Explanation

We give explanations for each relation in the four datasets. The detailed explanation for each relation is shown in Table 14.

B Experimental Results

B.1 Number of Groupings

Table 13 details the concrete relation extraction performance of the M-BRE framework, evaluating the comprehensive metrics under different number of groupings. Regarding the Tacred and SemEval

LLMs	Dataset	Number of non-NA	Relevant ratio	Relevant quantity
Qwen2.5-7B-Instruct-1M	Tacred 4/P	807/799	151/146	5/6
	Semeval 4/P	628/498	152/144	3/3
Qwen2.5-14B-Instruct-1M	Tacred 4/P	855/974	102/253	3/2
	Semeval 4/P	1561/1657	403/224	0/0
Qwen3-14B	Tacred 4/P	890/827	122/110	6/6
	Semeval 4/P	803/1041	330/197	5/4

Table 8: **Number of non-NA** represents the number of non-NA relation categories, **relevant ratio** represents the number ratio of relations with maximum generated quantities to relations with minimum generated quantities, **relevant quantity** represents the number of relations generated with a quantity of 0.

LLMs	Tacred	Tacrev	Retacred	SemEval
Qwen2.5-7B-Instruct-1M	411	411	376	670
	429	429	383	531
Qwen2.5-14B-Instruct-1M	476	476	448	1648
	510	510	445	1749
Qwen3-14B	915	915	833	865
	851	851	759	1115

Table 9: Number of samples constructed by each LLM. The first and second rows of each LLM correspond to group 4 and group p, where p = $\lfloor N/6 \rfloor$ and N is the total number of relation categories for each RE dataset.

datasets, our experimental configurations employ group sizes of [2, 13] and [2, 9] respectively.

B.2 Case Study

In response to the special phenomenon of "Multi-Yes" mentioned in §3.3 Confidence Judgement Module, we conduct a case study on RE training samples constructed by the M-BRe framework.

As shown in Figure 6, some head-tail entity pairs from unlabeled texts can be plausibly explained by multiple relation categories. Our M-BRe framework does not restrict the LLMs to output single relation category or "NA/no_relation/Others". By granting the LLMs greater creative diversity, more and more valid high-quality RE training samples can be constructed from single sentence-level unlabeled text.

B.3 Other Grouping Methods

We have supplemented experiments on the relation grouping algorithms, including K-Means and hierarchical clustering. The experiments were conducted with the number of clusters set to 4 and $\lfloor N/6 \rfloor$, using Qwen2.5-7B-Instruct-1M, Qwen2.5-14B-Instruct-1M, and Qwen3-14B. The primary evaluation metrics were Micro-F1, Macro-F1, and Special_Avg_F1. As shown in Table 10. Our relation grouping algorithm outperforms both K-Means and hierarchical clustering across most experimental conditions. This further validates the superiority of our proposed approach.

C Relation Grouping Algorithm

The detailed Relation Grouping Algorithm is given in Algorithm 1.

Algorithm 1: Relation Grouping

```
Input: Relation list $R$, $k \leftarrow \lfloor |R|/6 \rfloor$
Output: Groups $G = \{g_1, ..., g_k\}$
$V \leftarrow$ TF-IDF$(R)$ {Vectorize relations}
$S \leftarrow$ cosine$(V)$,
$D \leftarrow 1 - S$ {Similarity matrices}
$G \leftarrow \{\varnothing\}^k$ {Initialize groups}
$(i, j) \leftarrow \arg\max(D)$ {Select seeds}
$g_1 \leftarrow \{R_i\}, g_2 \leftarrow \{R_j\}$
while $\exists$ unassigned relations do
    for $r_u \in$ unassigned do
        $g^* \leftarrow \arg\min_{g_i} \max_{r_g \in g_i} S[r_u][r_g]$
    end for
    $g^* \leftarrow g^* \cup \{r^*\}$ {Assign best candidate}
end while
$G \leftarrow$ sort$(R$[indices]$)$
{Recover original relations}
return $G$
```

D Resource Consumption

The detailed Resource Consumption is given in Table 12.

Method	Qwen2.5-7B-Instruct-1M for Tacred			Qwen2.5-7B-Instruct-1M for SemEval		
	Mi-F1	Ma-F1	S_A_F1	Mi-F1	Ma-F1	S_A_F1
K-Means 4/P	48.79/51.21	50.65/53.02	40.71/37.80	20.88/37.36	18.38/31.64	19.78/31.07
Hierarchical 4/P	48.31/50.72	50.72/52.62	40.39/38.58	28.57/26.37	23.97/23.21	26.37/25.64
Ours 4/P	**57.10/60.50**	**54.05/57.86**	**42.19/38.91**	**30.98/39.52**	**34.74/38.95**	**27.29/32.05**
Method	Qwen2.5-14B-Instruct-1M for Tacred			Qwen2.5-14B-Instruct-1M for SemEval		
	Mi-F1	Ma-F1	S_A_F1	Mi-F1	Ma-F1	S_A_F1
K-Means 4/P	54.11/57.97	58.30/62.08	42.09/35.36	47.25/51.65	44.05/47.72	30.29/32.86
Hierarchical 4/P	53.62/58.94	57.79/62.71	42.24/38.23	50.04/46.15	51.61/41.98	33.92/35.77
Ours 4/P	**61.35/65.55**	**59.76/63.10**	**45.22/38.27**	**51.12/52.75**	**54.74/56.84**	**37.07/39.56**
Method	Qwen3-14B for Tacred			Qwen3-14B for SemEval		
	Mi-F1	Ma-F1	S_A_F1	Mi-F1	Ma-F1	S_A_F1
K-Means 4/P	67.63/71.01	70.29/73.68	42.46/40.90	19.57/20.88	17.79/18.74	14.65/16.21
Hierarchical 4/P	64.73/70.53	68.08/74.66	42.82/40.92	18.27/20.47	21.58/22.49	**19.01/20.88**
Ours 4/P	**70.68/76.89**	**70.95/76.19**	**46.07/41.15**	**19.76/21.16**	**24.21/25.26**	15.38/17.22

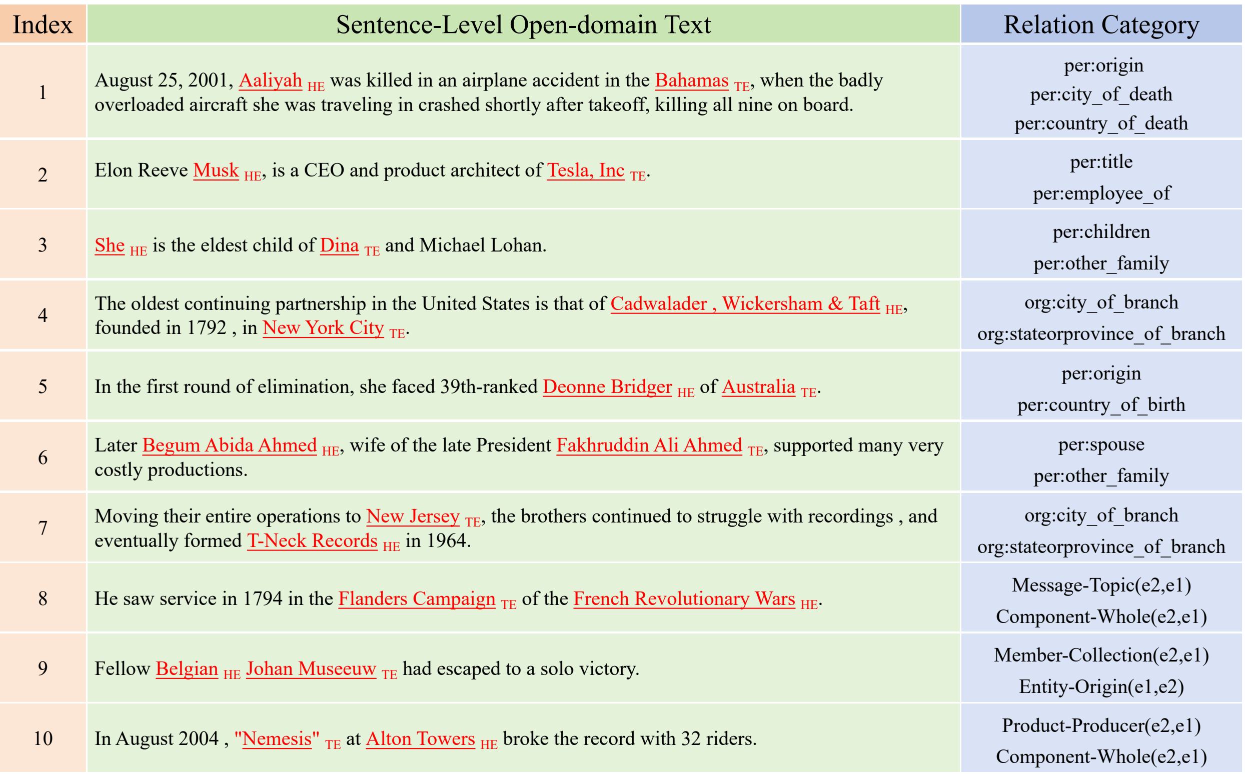
Table 10: Performance Results of Different Relation Grouping Methods for group 4 and group p, where p = $\lfloor N/6 \rfloor$ and N is the total number of relation categories for each RE dataset. Mi-F1, Ma-F1 and S_A_F1 represent Micro F1, Macro F1 and Special_Avg_F1.

E M-BRe for Fine-Grained Named Entity Recognition

We applied the M-BRe framework (Qwen2.5-14B-Instruct-1M) to Fine-Grained Named Entity Recognition. Experimental results on two datasets are shown in Table 11. It can be observed that the M-BRe framework also demonstrates high performance with low computational and time cost.

Datasets	Methods	Precison	Recall	S_A_F1	Time(h)
BNN	Multi	28.10	1.27	2.39	0.14
	M-BRe	30.20	40.00	32.04	2.52
	Binary	26.09	94.05	37.93	5.18
OntoNotes	Multi	31.48	0.70	1.36	0.22
	M-BRe	30.91	23.81	24.03	4.12
	Binary	27.97	79.80	37.73	7.55

Table 11: Performance comparison of M-BRe Framework on Fine-Grained Named Entity Recognition.



Index	Sentence-Level Open-domain Text	Relation Category
1	August 25, 2001, Aaliyah $_{\text{HE}}$ was killed in an airplane accident in the Bahamas $_{\text{TE}}$, when the badly overloaded aircraft she was traveling in crashed shortly after takeoff, killing all nine on board.	per:origin per:city_of_death per:country_of_death
2	Elon Reeve Musk $_{\text{HE}}$, is a CEO and product architect of Tesla, Inc $_{\text{TE}}$.	per:title per:employee_of
3	She $_{\text{HE}}$ is the eldest child of Dina $_{\text{TE}}$ and Michael Lohan.	per:children per:other_family
4	The oldest continuing partnership in the United States is that of Cadwalader , Wickersham & Taft $_{\text{HE}}$, founded in 1792 , in New York City $_{\text{TE}}$.	org:city_of_branch org:stateorprovince_of_branch
5	In the first round of elimination, she faced 39th-ranked Deonne Bridger $_{\text{HE}}$ of Australia $_{\text{TE}}$.	per:origin per:country_of_birth
6	Later Begum Abida Ahmed $_{\text{HE}}$, wife of the late President Fakhruddin Ali Ahmed $_{\text{TE}}$, supported many very costly productions.	per:spouse per:other_family
7	Moving their entire operations to New Jersey $_{\text{TE}}$, the brothers continued to struggle with recordings , and eventually formed T-Neck Records $_{\text{HE}}$ in 1964.	org:city_of_branch org:stateorprovince_of_branch

8	He saw service in 1794 in the Flanders Campaign $_{\text{TE}}$ of the French Revolutionary Wars $_{\text{HE}}$.	Message-Topic(e2,e1)
		Component-Whole(e2,e1)

9	Fellow Belgian $_{\text{HE}}$ Johan Museeuw $_{\text{TE}}$ had escaped to a solo victory.	Member-Collection(e2,e1)
		Entity-Origin(e1,e2)

10	In August 2004 , "Nemesis" $_{\text{TE}}$ at Alton Towers $_{\text{HE}}$ broke the record with 32 riders.	Product-Producer(e2,e1)
		Component-Whole(e2,e1)

Figure 6: A case study of relation extraction on sentence-level unlabeled text with the M-BRe framework, where HE and TE represents the head entity and the tail entity.

LLMs	Method	Average Tokens Processed (Tacred/Semeval)	GPU Type
Qwen2.5-7B-Instruct-1M	Multi-Class	1920/860	NVIDIA RTX 3090*1
	M-BRe 4	3560/2400	NVIDIA RTX 3090*1
	M-BRe $\lfloor N/6 \rfloor$	4655/2100	NVIDIA RTX 3090*1
	Binary-Class	14070/5700	NVIDIA RTX 3090*1
Qwen2.5-14B-Instruct-1M	Multi-Class	1920/860	NVIDIA RTX 3090*2
	M-BRe 4	3560/2400	NVIDIA RTX 3090*2
	M-BRe $\lfloor N/6 \rfloor$	4655/2100	NVIDIA RTX 3090*2
	Binary-Class	14070/5700	NVIDIA RTX 3090*2
Qwen3-14B	Multi-Class	1920/860	NVIDIA RTX 3090*2
	M-BRe 4	3560/2400	NVIDIA RTX 3090*2
	M-BRe $\lfloor N/6 \rfloor$	4655/2100	NVIDIA RTX 3090*2
	Binary-Class	14070/5700	NVIDIA RTX 3090*2

Table 12: Resource consumption of various LLMs under different frameworks.

Group	Qwen2.5-7B-Instruct-1M for Tacred						Qwen2.5-7B-Instruct-1M for SemEval					
	Ma-P	Ma-R	Ma-F1	Mi-F1	S_A_F1	Time(h)	Ma-P	Ma-R	Ma-F1	Mi-F1	S_A_F1	Time(h)
Mix2	44.44	54.49	44.52	45.51	38.81	0.22	34.07	37.59	32.63	32.08	28.94	0.09
Mix3	50.24	68.79	50.24	52.20	42.11	0.33	**40.66**	**47.06**	**38.95**	**39.52**	**32.05**	**0.13**
Mix4	54.11	70.70	54.05	57.10	42.19	0.46	36.26	35.25	34.74	30.98	27.29	0.16
Mix5	55.55	72.11	55.48	58.48	39.61	0.53	27.47	26.13	26.32	23.47	20.37	0.21
Mix6	56.04	70.33	55.95	57.86	40.53	0.69	27.47	24.04	26.32	21.52	19.93	0.24
Mix7	**57.97**	**73.97**	**57.86**	**60.50**	**38.91**	**0.72**	28.57	31.43	27.37	24.85	19.34	0.27
Mix8	56.04	71.68	55.95	58.41	37.92	0.84	31.87	23.72	30.53	25.28	19.93	0.28
Mix9	56.04	74.35	55.95	59.37	37.04	0.93	30.77	35.98	29.47	26.35	18.92	0.32
Mix10	55.07	67.38	55.00	56.15	35.33	1.17	-	-	-	-	-	-
Mix11	57.00	74.66	56.90	59.74	35.47	1.18	-	-	-	-	-	-
Mix12	57.97	73.85	57.86	61.27	36.06	1.23	-	-	-	-	-	-
Mix13	57.97	73.60	57.86	61.51	33.89	1.49	-	-	-	-	-	-

Group	Qwen2.5-14B-Instruct-1M for Tacred						Qwen2.5-14B-Instruct-1M for SemEval					
	Ma-P	Ma-R	Ma-F1	Mi-F1	S_A_F1	Time(h)	Ma-P	Ma-R	Ma-F1	Mi-F1	S_A_F1	Time(h)
Mix2	57.00	68.32	56.90	57.75	51.37	0.87	49.45	53.44	47.37	46.19	40.29	0.27
Mix3	56.52	68.03	56.43	58.19	45.49	1.27	**54.95**	**55.75**	**56.84**	**52.75**	**39.56**	**0.42**
Mix4	59.90	73.86	59.76	61.35	45.22	1.66	57.14	58.48	54.74	51.12	37.07	0.63
Mix5	59.42	69.12	59.29	59.76	41.47	2.26	59.34	65.61	61.05	59.25	38.72	0.71
Mix6	62.32	81.09	62.14	65.37	41.76	3.02	61.54	54.50	58.95	54.93	35.90	0.73
Mix7	**63.29**	**77.79**	**63.10**	**65.55**	**38.27**	**3.08**	63.74	61.05	61.05	58.41	39.35	0.87
Mix8	63.29	80.42	63.10	65.09	37.15	3.77	63.74	72.58	61.05	61.22	30.97	1.03
Mix9	63.77	84.45	63.57	67.65	34.96	3.87	64.84	77.98	66.32	65.91	27.35	1.18
Mix10	62.80	80.64	62.62	65.88	31.26	4.27	-	-	-	-	-	-
Mix11	65.22	85.64	65.00	68.87	31.65	4.50	-	-	-	-	-	-
Mix12	64.25	84.97	64.05	67.80	30.23	5.57	-	-	-	-	-	-
Mix13	66.67	85.19	66.43	70.14	30.15	6.33	-	-	-	-	-	-

Group	Qwen3-14B for Tacred						Qwen3-14B for SemEval					
	Ma-P	Ma-R	Ma-F1	Mi-F1	S_A_F1	Time(h)	Ma-P	Ma-R	Ma-F1	Mi-F1	S_A_F1	Time(h)
Mix2	68.12	77.20	68.57	69.43	56.84	0.34	15.38	24.85	14.74	14.23	14.29	0.11
Mix3	69.57	77.15	69.29	70.03	51.21	0.56	**21.98**	**30.50**	**25.26**	**21.16**	**17.22**	**0.17**
Mix4	70.53	76.91	70.95	70.68	46.07	0.61	20.88	28.06	24.21	19.76	15.38	0.25
Mix5	71.98	81.21	72.38	72.83	43.41	0.80	13.19	11.22	12.63	7.84	9.30	0.31
Mix6	72.46	87.53	72.86	74.73	42.87	0.85	15.38	21.12	14.74	11.69	11.02	0.41
Mix7	**75.85**	**84.19**	**76.19**	**76.89**	**41.15**	**0.87**	12.08	24.01	11.58	10.38	6.69	0.45
Mix8	77.29	89.87	77.62	79.32	37.96	0.99	10.99	17.00	10.53	8.27	5.07	0.51
Mix9	75.85	86.52	76.19	77.74	35.46	1.08	15.38	15.65	18.95	12.87	6.79	0.44
Mix10	76.33	89.20	76.67	78.45	35.82	1.23	-	-	-	-	-	-
Mix11	76.33	85.86	75.95	77.47	31.88	1.28	-	-	-	-	-	-
Mix12	76.33	86.08	76.67	78.49	31.27	1.73	-	-	-	-	-	-
Mix13	77.29	88.28	77.62	79.25	30.42	1.75	-	-	-	-	-	-

Table 13: Comprehensive assessment of different Number of Groupings. **Ma-P**, **Ma-R**, **Ma-F1**, **Mi-F1** and **S_A_F1** respectively represent Macro Precision, Macro Recall, Macro F1, Micro F1 and Special_Avg_F1. **Bold** denotes the optimal trade-off point.

Relation	**Explanation**
Component-Whole (e2,e1)	Tail entity e2 is the component of head entity e1, and head entity e1 is the whole of tail entity e2.
Instrument-Agency (e2,e1)	Tail entity e2 is the instrument of head entity e1, and head entity e1 is the agency of tail entity e2.
Member-Collection (e1,e2)	Head entity e1 is the member of tail entity e2, and tail entity e2 is the collection of head entity e1.
Cause-Effect (e2,e1)	Tail entity e2 is the cause of head entity e1, and head entity e1 is the effect of tail entity e2.
Entity-Destination (e1,e2)	Head entity e1 is the entity of tail entity e2, and tail entity e2 is the destination of head entity e1.
Content-Container (e1,e2)	Head entity e1 is the content of tail entity e2, and tail entity e2 is the container of head entity e1.
Message-Topic (e1,e2)	Head entity e1 is the message of tail entity e2, and tail entity e2 is the topic of head entity e1.
Product-Producer (e2,e1)	Tail entity e2 is the product of head entity e1, and head entity e1 is the producer of tail entity e2.
Member-Collection (e2,e1)	Tail entity e2 is the member of head entity e1, and head entity e1 is the collection of tail entity e2.
Entity-Origin (e1,e2)	Head entity e1 is the entity of tail entity e2, and tail entity e2 is the origin of head entity e1.
Cause-Effect (e1,e2)	Head entity e1 is the cause of tail entity e2, and tail entity e2 is the effect of head entity e1.
Component-Whole (e1,e2)	Head entity e1 is the component of tail entity e2, and tail entity e2 is the whole of head entity e1.
Message-Topic (e2,e1)	Tail entity e2 is the message of head entity e1, and head entity e1 is the topic of tail entity e2.
Product-Producer (e1,e2)	Head entity e1 is the product of tail entity e2, and tail entity e2 is the producer of head entity e1.
Entity-Origin (e2,e1)	Tail entity e2 is the entity of head entity e1, and head entity e1 is the origin of tail entity e2.
Content-Container (e2,e1)	Tail entity e2 is the content of head entity e1, and head entity e1 is the container of tail entity e2.
Instrument-Agency (e1,e2)	Head entity e1 is the instrument of tail entity e2, and tail entity e2 is the agency of head entity e1.
Entity-Destination (e2,e1)	Tail entity e2 is the entity of head entity e1, and head entity e1 is the destination of tail entity e2.
Other	There is no relationship or unrecognized relationship between the head and tail entities.
org:founded	The founding time of an organization.
org:subsidiaries	The subsidiaries of an organization.
per:date_of_birth	The date of birth of a person.
per:cause_of_death	The cause of death of a person.
per:age	The age of a person.
per:stateorprovince_of_birth	The state or province of birth of a person.

Relation	**Explanation**
per:countries_of_residence	The countries where a person resides.
per:country_of_birth	The country of birth of a person.
per:stateorprovinces_of_residence	The states or provinces where a person resides.
org:website	The website of an organization.
per:cities_of_residence	The cities where a person resides.
per:parents	The parents of a person.
per:employee_of	The organization where a person is employed.
NA/no_relation	Unknown or non-existent relation.
per:city_of_birth	The city of birth of a person.
org:parents	The parent company of an organization.
org:political/religious_affiliation	The political or religious affiliation of an organization.
per:schools_attended	The schools attended by a person.
per:country_of_death	The country where a person died.
per:children	The children of a person.
org:top_members/employees	The top members/employees of an organization.
per:date_of_death	The date of death of a person.
org:members	The members of an organization.
org:alternate_names	The alternate names of an organization.
per:religion	The religion of a person.
org:member_of	The organization to which a member belongs.
org:city_of_headquarters	The city where the headquarters of an organization is located.
per:origin	The origin of a person.
org:shareholders	The shareholders of an organization.
per:charges	The charges against a person.
per:title	The occupation of a person.
org:number_of_employees/members	The number of employees/members in an organization.
org:dissolved	The date of dissolution of the organization.
org:country_of_headquarters	The country where headquarters of an organization is located.
per:alternate_names	The alternate names of a person.
per:siblings	The siblings of a person.
org:stateorprovince_of_headquarters	The state or province where headquarters of an organization is located.
per:spouse	The spouse of a person.
per:other_family	Other family members of a person.
per:city_of_death	The city where a person died.
per:stateorprovince_of_death	The state or province where a person died.
org:founded_by	The founder of an organization.
org:country_of_branch	The country where a branch of an organization is located.
org:city_of_branch	The city where a branch of an organization is located.
org:stateorprovince_of_branch	The state or province where branch of an organization is located.
per:identity	The identity information or characteristics of a person.

Table 14: Explanation of each relation in the four datasets.