Rethinking the Role of LLMs for Document-level Relation Extraction: a Refiner with Task Distribution and Probability Fusion

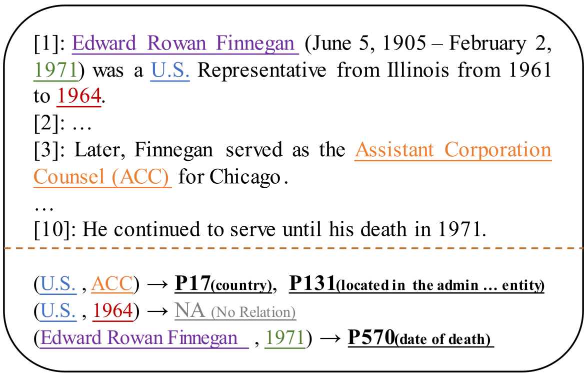
Fu Zhang[†], Xinlong Jin[†], Jingwei Cheng[*], Hongsen Yu, Huangming Xu
School of Computer Science and Engineering, Northeastern University, China
{zhangfu,chengjingwei}@mail.neu.edu.cn; drasick59596@163.com

Abstract

Document-level relation extraction (DocRE) provides a broad context for extracting one or more relations for each entity pair. Large language models (LLMs) have made great progress in relation extraction tasks. However, one of the main challenges we face is that LLMs have difficulty in multi-label relation prediction tasks. Additionally, another noteworthy challenge and discovery we reveal: the small language models (SLMs) for DocRE tend to classify existing relations as "no relation" (NA), while LLMs tend to predict existing relations for all entity pairs. To address these challenges, we propose a novel method that utilizes LLMs as a refiner, employing task distribution and probability fusion. The task distribution we carefully designed aims to distinguish hard and easy tasks, and feed hard tasks to our LLMs-based framework to reevaluate and refine. Further, in order to effectively solve the multi-label relation prediction problem in the refinement process, we propose a probability fusion method, ensuring and enhancing fusion predictions by maintaining a balance between SLMs and LLMs. Extensive experiments on widely-used datasets demonstrate that our method outperforms existing LLM-based methods without fine-tuning by an average of 25.2% F1. Refining SLMs using our method consistently boosts the performance of the SLMs, achieving new state-of-the-art results compared to existing SLMs and LLMs[1].

1 Introduction

Relation extraction (RE) is the task of extracting semantic relations among entities within a given text, which has abundant applications such as knowledge graph construction, question answering, and text analysis (Vaswani et al., 2017; Distiawan et al., 2019; Shi et al., 2019). Prior studies mostly focus on predicting a relation between two entities mentioned in a single sentence which is called sentence-level relation extraction. By contrast, document-level relation extraction (DocRE) (Yao et al., 2019) offers a broader context for analysis and poses greater challenges, as it involves identifying one or more relations for entities that span multiple sentences or paragraphs as illustrated in Figure 1.


[1]: Edward Rowan Finnegan (June 5, 1905 – February 2, 1971) was a U.S. Representative from Illinois from 1961 to 1964.
[2]: …
[3]: Later, Finnegan served as the Assistant Corporation Counsel (ACC) for Chicago.
…
[10]: He continued to serve until his death in 1971.

(U.S. , ACC) → **P17(country)**, **P131(located in the admin … entity)**
(U.S. , 1964) → NA (No Relation)
(Edward Rowan Finnegan , 1971) → **P570(date of death)**

Figure 1: A DocRE example: predicting the existence of one or more relations, or no relations for entity pairs.

Recent advancements in DocRE focus on the use of neural models for sequence-based, graph-based, and transformer-based approaches (Delaunay et al., 2023). Meanwhile, large language models (LLMs) have achieved significant success in a wide range of natural language processing tasks, leveraging emergent capabilities of reasoning and in-context learning across diverse domains such as commonsense reasoning and open-domain question answering (Yu et al., 2023; Sun et al., 2023). Recent studies have utilized LLMs for relation extraction tasks, including few-shot RE (Wei et al., 2023; Xu et al., 2023b; Ma et al., 2023b) and sentence-level RE (Wadhwa et al., 2023; Zhang et al., 2023). Methods that involve designing prompts and fine-tuning LLMs for specific tasks have been shown to outperform fine-tuned small language models (SLMs) in several relation extraction domains (Gutierrez

[1] Our code: https://github.com/Drasick/Drell.
[†] Equal contribution. [*] Corresponding author.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6293–6312
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

et al., 2022; Xu et al., 2023a).

However, there are few studies on applying LLMs to DocRE, including PromptRE (Gao et al., 2023) and DocGNRE (Li et al., 2023) without fine-tuning, and AutoRE (Lilong et al., 2024) with fine-tuning. Building on our in-depth research, we identify several key focal points that warrant attention and resolution: **(i)** LLMs **struggle to handle datasets with a large number of negative samples effectively**. Our research reveals that the limited predictive performance of SLMs owing to its tendency to classify existing relations as "no relation" (NA), while LLMs tend to predict existing relations for all entity pairs. **(ii)** LLMs, which are essentially generative models, **struggle with multi-label relation prediction tasks**. When multiple relations need to be predicted for an entity pair, the answer generated by LLMs may not exactly match the relation labels. Even when LLMs are employed to choose from multiple labels, they prefer to choose a single label.

To address these points, aimed at exploratory of the ability differed between SLMs and LLMs for DocRE, we propose a method that utilizes LLMs as a *refiner*, employing task distribution and probability fusion to complete the DocRE task. The *task distribution* initially uses SLMs to address the issue of a large number of negative samples and subsequently uses LLMs as a refiner to resolve tasks that are difficult for SLMs. The *probability fusion* provides a stable solution to the multi-label classification problem, ensuring that the predictions are not overly dependent on either the SLMs or the LLMs. Our contributions are as follows:

- We explore the performance of SLMs and LLMs in DocRE and reveal several noteworthy findings.
- We propose a task distribution method that allows LLMs, acting as a refiner, to effectively assist in DocRE tasks that are difficult for SLMs.
- We innovatively propose a probability fusion method for multi-label classification, balancing and enhancing the predictions made by both SLMs and LLMs.
- Experiments on widely-used DocRE datasets demonstrate that using LLMs as a refiner can consistently enhance the performance of SLMs. Moreover, our method is cost-effective, saving both time and cost without requiring additional fine-tuning[2].

2 Related Work

DocRE involves extracting relations between entities within a document. Let the document be defined as D, composed of N sentences $\{s_n\}_{n=1}^{N}$. We need to combine all mentioned entities $\{e_q\}_{q=1}^{E}$ pairwise to form entity pairs $(e_h, e_t)_{h,t \in \{1,2,\ldots,E\};h \neq t}$, where h represents the head entity and t represents the tail entity. The task is to predict the relation r for this entity pair, where r belongs to the pre-defined set $\{r_i\}_{i=1}^{R} \cup \{\mathrm{NA}\}$. For the entity pair (e_h, e_t), we define its probability relative to relation r as $P(r)$. Therefore, we define the *probability distribution* $F^{(h,t)}$ of the entity pair (e_h, e_t) as:

$$F^{(h,t)} = \{P(r) | r \in \{r_i\}_{i=1}^{R} \cup \{\mathrm{NA}\}\} \quad (1)$$

2.1 DocRE based on SLMs

The utilization of SLMs for DocRE can be roughly categorized into sequence-based (e.g. CNN (Yao et al., 2019), BiLSTM (Yao et al., 2019)), graph-based (e.g. GAIN (Zeng et al., 2020), SIRE (Zeng et al., 2021)), and transformer-based. *We prefer transformer-based models with the same underlying architecture as LLMs*, so we select several recently representative and competitive transformer-based models: ATLOP (Zhou et al., 2021), Eider (Xie et al., 2022), DREEAM (Ma et al., 2023a), and AA (Lu et al., 2023) as SLMs baselines for subsequent experiments. Descriptions of these SLMs can be found in the review work (Delaunay et al., 2023) in details.

Transformer-based models utilize pre-trained models (such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019)) to produce probability predictions for relations in $R \cup \{\mathrm{NA}\}$. By using the predicted value for the NA relation as a threshold, and considering all relations with values greater than this threshold as the predicted relation labels for (e_h, e_t), we define this process as:

$$P_{slm}(r|e_h, e_t; D) = \sigma(\sum_{i=1}^{k} z_h^{(h,t)} W_r^i z_t^{(h,t)} + b_r) \quad (2)$$

where $z_h^{(h,t)}$ and $z_t^{(h,t)}$ are the embeddings with the in-context information for the head and tail

[2] All LLMs discussed in this paper *are not fine-tuned*, and the results for LLMs are based on default weights or APIs.

entities, W_r^i and b_r is the weight matrix and bias for relation r, σ is the activation. For simplicity, $P_{slm}(r|e_h, e_t; D)$ is abbreviate as $P_{slm}(r)$. Hence, the probability distribution $F_{slm}^{(h,t)}$ is denoted as :

$$F_{slm}^{(h,t)} = \{P_{slm}(r)|r \in R \cup \{\text{NA}\}\} \qquad (3)$$

The final probability set $P^{(h,t)}$ of relation labels for the entity pair (e_h, e_t) is defined as:

$$P^{(h,t)} = \{P_{slm}|P_{slm}(r) > P_{slm}(\text{NA})\} \qquad (4)$$

This approach is widely used in SLMs and completes the multi-label classification for DocRE.

2.2 DocRE based on LLMs

Currently, there are few LLMs-based models evaluated on document-level datasets. Two notable studies *without fine-tuning* conducted on DocRE are discussed: PromptRE (Gao et al., 2023) and DocGNRE (Li et al., 2023). **PromptRE** combines diverse prompting, integrating label distribution and entity types to enhance DocRE. **DocGNRE** autonomously generates relation triples to apply LLMs for dataset augmentation, instead of using LLMs for DocRE.

The smallest unit of LLM generation is defined as a *token* w_i, its probability among all tokens in the vocabulary V can be positioned according to the previous token sequence $[w_1, \ldots, w_{i-1}]$ (commonly called as *prompt*) as follows:

$$P_{llm}(w_i|w_1, w_2, \ldots, w_{i-1}) \qquad (5)$$

$$F_{llm}(w_i) = \{P_{llm}(w_j)|w_j \in V\} \qquad (6)$$

where $F_{llm}(w_i)$ is denoted as the probability distribution of the output token. The LLM-based methods typically output tokens until the end. Subsequently, the output $[w_i, w_{i+1} \ldots, w_{end}]$ is regularized to predict multi-relation labels.

3 Exploratory Analysis of SLMs and LLMs for DocRE

To investigate the efficacy of LLMs in DocRE and explore their performance distinctions with SLMs, we devise two experiments: ($\boldsymbol{E}_1$) Similar to previous approaches based on LLMs, we mainly utilize prompts to predict relation labels, aiming to initially *assess the performance of LLMs* in DocRE. ($\boldsymbol{E}_2$) We conduct an in-depth *comparative analysis* of relation labels predicted by LLMs and SLMs to *insight the potential capabilities of LLMs* in DocRE and *uncover any limitations in SLMs*.

	DocRED	Re-DocRED
# Train	3053	3053
# Dev	1000	500
# Test	1000	500
Triples	50,503	120,664
Relation types	96	96

Table 1: Dataset statistics. # indicates document count.

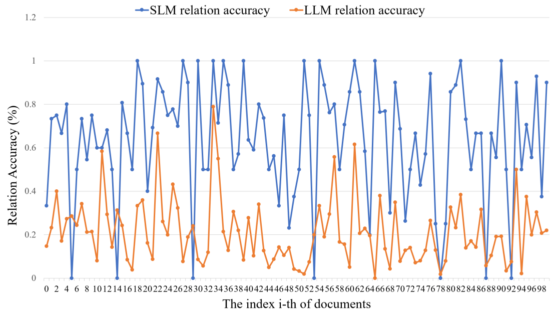


Figure 2: Comparison of relation prediction accuracy.

3.1 Experiment Setup

Datasets We evaluate on widely-adopted datasets for DocRE, including DocRED (Yao et al., 2019) and Re-DocRED (Tan et al., 2022) in Table 1. Re-DocRED improves annotation labels upon the popular DocRED dataset.

Implementation Settings Regarding the LLMs in $\boldsymbol{E}_1$, we apply GPT-3.5-turbo-0613 following PromptRE and DocGNRE, and comparing on Re-DocRED. In $\boldsymbol{E}_2$, we employ the SLM ATLOP (Zhou et al., 2021) and compare it with our LLM-based method on DocRED.

Evaluation Metric We adopt *F1* and *Ign F1* as the evaluation metrics for DocRE, as established in previous research. *Ign F1* disregards triples that are present in training set.

For each document, we will consider *relation accuracy* (defined as the proportion of correctly predicted labels among all predicted labels), and the prediction of *NA_but_relation* (defined as the number of entity pairs with relation labels predicted as NA).

3.2 $\boldsymbol{E}_1$: Observation of LLMs on DocRE

To validate the inherent capability of LLMs in DocRE, we design a document-fitting prompt template (as will be introduced in Section 4.1 and Figure 4)[3] to test each entity pair. This prompt is then

[3] *All prompt templates* mentioned in this paper can be found in Appendix A.

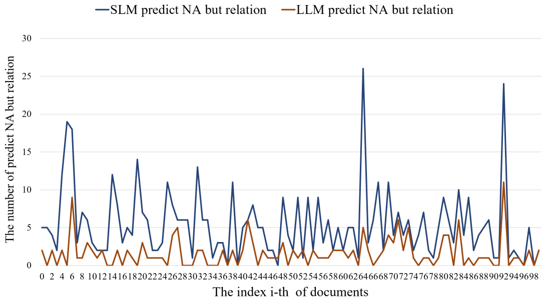


Figure 3: Comparison of predicting NA_but_relation.

fed to GPT-3.5 for querying and reasoning to derive the predicted relation labels.

Table 2 shows that our method achieves improvements over the baselines, but the performance of directly using LLM prompts in DocRE is still unsatisfactory, as well as other LLMs-based methods. This suggests that the direct utilization of prompts currently may not leverage the reasoning capabilities inherent in LLMs like other domains.

3.3 E_2: Tendency Analysis of SLMs & LLMs

To investigate why direct use of prompts does not achieve the same or better effectiveness as SLMs, we conduct experiments on each document, analyze the experimental results in depth, and randomly select 100 documents for visualization. The experiments' outcomes are illustrated in Figures 2 and 3: **(1)** The performance of the **SLM is constrained** because, **while it ensures high prediction relation accuracy, it discards many relations as NA**. Our in-depth analysis reveals that the high performance of SLMs is largely attributed to their higher prediction accuracy. Additionally, an interesting finding emerges from Figure 3: among the results of SLM prediction errors, there is a tendency to predict entity pairs that have relation labels as NA, thereby limiting their performance. **(2)** The **diminished efficacy of LLMs** stems from **their tendency to predict existence labels for many NA relations**. In Figure 2, compared to the SLM, the proportion of correct predictions by LLMs, excluding NA, is relatively low. Further, Figure 3 illustrates that LLMs tend to predict existence labels for many NA relations. This over-prediction phenomenon has also been observed in the few-shot relation extraction GPT-RE (Wan et al., 2023). We suspect that LLMs may rely on their own prior knowledge to make predictions.

Model	**F1**	**Ign F1**
PromptRE (Gao et al., 2023)*	10.56	9.04
DocGNRE (Li et al., 2023)*	11.73	-
Only-LLM (Ours)	22.95	22.36

Table 2: Performance of our LLM prompting-based method. Results of $*$ are from their original papers.

3.4 Discussion: Why not leverage LLMs to refine predictions by SLMs?

Based on Sections 3.2 and 3.3, the following *issues arise*: The potential of LLMs in DocRE still has significant room for exploration. And, leveraging characteristics of LLMs and SLMs to jointly enhance DocRE may be valuable.

Hence, **our main ideas** are as follows: *First*, we propose a task distribution method that allows LLMs as a refiner, to effectively assist SLMs. This involves identifying tasks where the SLMs discard many relations as NA, which should be refined by the LLMs. *Second*, even with LLMs assisting SLMs, LLMs themselves still struggle with multi-label relation prediction tasks. We need to design a method tailored for multi-label classification. *Moreover*, to further enhance fusion predictions by maintaining a balance between SLMs and LLMs, we propose a probability balance fusion method.

4 Methodology

Our refining method consists of Task Distribution and Self-supervised Probability Fusion as shown in Figure 4.

4.1 Task Distribution

As mentioned in Section 2, when given an entity pair (e_h, e_t) to be predicted in a document D, SLMs can derive the probability $P_{slm}(r)$ and distribution F_{slm} using Eq. (2) and (3).

For the tendency as discussed in Section 3.3, we posit that when $P_{slm}(\mathrm{NA})$ is greater than $P_{slm}(r)_{r\in R}$, but there is at least one $P_{slm}(r)$ in *close proximity* to $P_{slm}(\mathrm{NA})$, the task is defined as **Hard**. We define the close proximity as $\gamma_{(h,t)}$. In light of this, we introduce a partition method to distinguish between easy and hard tasks based on $P_{slm}(\mathrm{NA})$ as the threshold. The formulation of this method is defined:

$$\gamma_{(h,t)} = \frac{P_{slm}(\mathrm{NA}) - \max(P_{slm}(r))}{P_{slm}(\mathrm{NA})}, \gamma_{(h,t)} \leq \delta \tag{7}$$

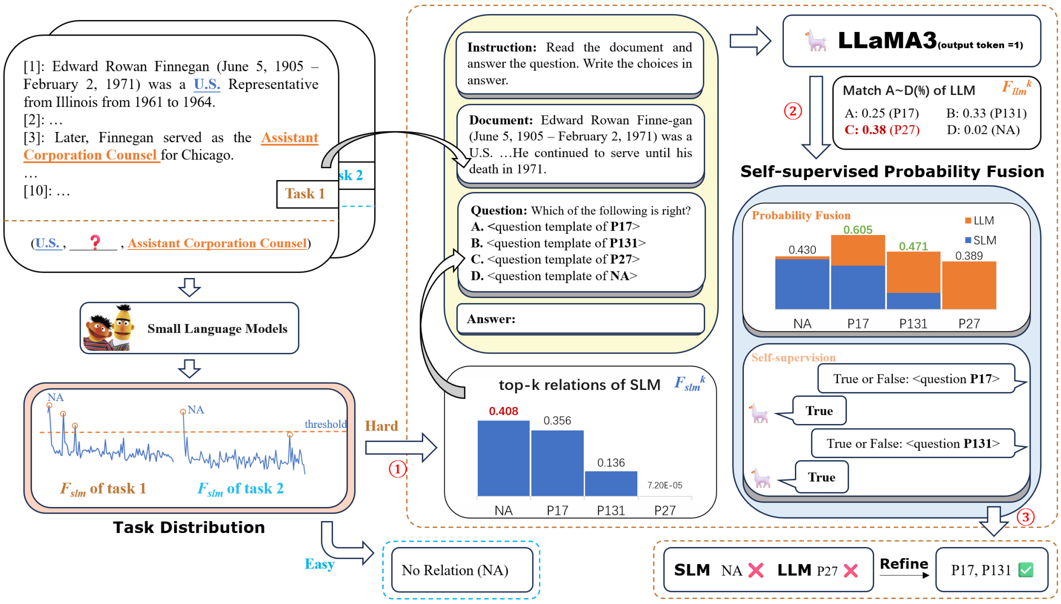

[1]: Edward Rowan Finnegan (June 5, 1905 – February 2, 1971) was a U.S. Representative from Illinois from 1961 to 1964.
[2]: …
[3]: Later, Finnegan served as the Assistant Corporation Counsel for Chicago.
…
[10]: …
Task 1
sk 2
(U.S. , ? , Assistant Corporation Counsel)
Small Language Models
NA
NA
threshold
F_{slm} of task 1
F_{slm} of task 2
Task Distribution
Hard
①
Easy
No Relation (NA)
Instruction: Read the document and answer the question. Write the choices in answer.
Document: Edward Rowan Finne-gan (June 5, 1905 – February 2, 1971) was a U.S. …He continued to serve until his death in 1971.
Question: Which of the following is right?
A. <question template of P17>
B. <question template of P131>
C. <question template of P27>
D. <question template of NA>
Answer:
top-k relations of SLM $F_{slm}{}^k$
0.408
0.356
0.136
7.20E-05
NA
P17
P131
P27
LLaMA3(output token =1)
②
Match A~D(%) of LLM $F_{llm}{}^k$
A: 0.25 (P17)
B: 0.33 (P131)
C: 0.38 (P27)
D: 0.02 (NA)
Self-supervised Probability Fusion
Probability Fusion
LLM
SLM
0.430
0.605
0.471
0.389
NA
P17
P131
P27
Self-supervision
True or False: <question P17>
True
True or False: <question P131>
True
③
SLM NA ✗ LLM P27 ✗
Refine
P17, P131 ✓

Figure 4: Illustration for our LLMs-based refiner framework. The task distribution is to distinguish hard and easy tasks, and feed hard tasks to refine (Step ①). The question templates converted from the top-k relations of SLMs and the document are together fed into the LLM, to get the distribution F_{llm^k} of LLM's relation predictions (Step ②). With the probability balance fusion of two distributions F_{slm^k} and F_{llm^k} and self-supervision, we obtain the final relation predictions (Step ③).

where $\gamma_{(h,t)}$ is denoted as the threshold, and δ is a parameter, controlling the range scaling of the task difficulty. If the task is regarded as hard, it will be distributed to the LLM for refinement on the basis of the distribution F_{slm}.

The idea of distinguishing hard tasks is similarly proposed by Ma et al. (2023b). Inspired by their work but differing from theirs, *first*, they focus on few-shot RE task (where a relation between two entities is predicted from a single sentence), and use a *fixed threshold* to give all relation prediction scores. We propose the idea of *dynamic threshold* above, which allows us to adjust the threshold adaptively according to the predicted score of each entity pair. *Second*, *our work after task distribution is completely different from theirs*, we for the first time propose the idea and method of balancing and fusing the probability distributions of LLMs and SLMs to achieve multi-label classification tasks.

After distributing the hard tasks to the LLM, considering that the LLM still struggles with multi-label relation prediction tasks, we carefully design a *document-fitting prompt template* consisting of *Instruction* + *Document* + *Question* + *Answer*: *Instruction* informs the LLM that it needs to undertake a multiple-choice task. *Document* contains the document information where (e_h, e_t) belongs to. The most important part, *Question*, is what the LLM needs to reason about and answer. The question includes the top-k relation probability labels (obtained by F_{slm^k}, where k indicates only the top-k $P_{slm}(r)$ are retained), filled with head and tail entities, and converted into a question template designed by ourselves. After our prompt is input into the LLM, unlike the regularization-based multi-label classification methods mentioned at the end of Section 2.2, **we focus only on the probability distribution F_{llm} of the first output token** (i.e., Eq. (6)). We then identify all tokens in this distribution F_{llm} that match our multiple-choice options. The probability values of these tokens are taken as the final distribution F_{llm^k} of the LLM's relation label predictions.

Now, we have two relation probability distributions for the entity pair (e_h, e_t): F_{slm^k} and F_{llm^k}.

4.2 Self-supervised Probability Fusion

Probability Fusion A challenge arises because F_{slm^k} is obtained through the prediction results of all relations, while F_{llm^k} is based on the token probability distribution of the entire vocabulary. Although we can add the probabilities from the two distributions one by one, intuitively, the two distributions are unbalanced[4].

[4] We also demonstrate in subsequent experiments of Section 5.3 that direct addition is unreasonable.

Therefore, we propose a *probability balance estimation*. Through this estimation, we aim to make the distributions of the two models balance, which helps the final prediction with the judgment of both F_{slm^k} and F_{llm^k}.

Considering that the probability distribution F_{llm^k} is calculated based on the softmax function including the temperature parameter τ, which transforms the logits (log-odds) z_i for each token into a distribution over the vocabulary V:

$$P_{llm}(r_i|z_i, \tau) = \frac{e^{\frac{z_i}{\tau}}}{\sum_{j=1}^{V} e^{\frac{z_j}{\tau}}} \tag{8}$$

thus, we adjust the distribution of next token output in LLMs by τ, which helps control the diversity of generated text.

As the hard task we defined in Eq. (7), we classify tasks where the probabilities $P_{slm}(\text{NA})$ and $P_{slm}(r)$ are close as hard tasks. This closeness may affect the variance of the distribution F_{slm^k}. At the same time, as shown in Eq. (8), adjusting τ can also change the variance of F_{llm^k} (we denote the adjusted F_{llm^k} as $F^{\tau}_{llm^k}$), where we observe that as τ increases, the variance will decrease. The variances σ_{slm} and σ_{llm} are defined based on their respective distributions F_{slm^k} and $F^{\tau}_{llm^k}$. Therefore, we propose to further determine whether F_{slm^k} and $F^{\tau}_{llm^k}$ are balanced by calculating the difference between their variances with a threshold ξ:

$$|\sigma_{slm}(F_{slm^k}) - \sigma_{llm}(F^{\tau}_{llm^k})| \leq \xi \tag{9}$$

After estimating probability balance, we combine $P_{slm}(r_i)$ and $P_{llm}(r_i)$ in F_{slm^k} and $F^{\tau}_{llm^k}$:

$$F_{ref} = \{P_{slm}(r_i) + P_{llm}(r_i) | i = 1, \ldots, k\} \tag{10}$$

According to the $P_{ref}(r)$ in F_{ref}, we finally obtain the refined relation prediction probability set $P^{(h,t)}_{ref}$ of the entity pair (e_h, e_t):

$$P^{(h,t)}_{ref} = \{P_{ref} | P_{ref}(r) > P_{ref}(\text{NA})\} \tag{11}$$

Self-supervision To further enhance the accuracy of the probability fusion results, we utilize LLM's self-supervision to enable it to make a second judgment on the document and each relation in $P^{(h,t)}_{ref}$. The method involves posing the question again, asking whether there is a relation. We judge based on the probability distribution of the first token output by the LLM, focusing on the token predictions of "T" (true) and "F" (false) as showed in Figure 4.

5 Experiments

5.1 Experimental Setup and Baselines

Datasets As detailed in Section 3.1, we evaluate our methods on two widely-adopted DocRE datasets DocRED and Re-DocRED, and report *Precision*, *Recall*, *F1* and *Ign F1*.

Baselines As detailed in Section 2, we choose two earlier SLMs CNN and BiLSTM that are often compared as baselines.

As our main baselines, transformer-based models, which have the same underlying architecture as LLMs, we select four representative state-of-the-art SLMs as the refining models, including ATLOP, Eider, DREEAM, and AA. We also compare with the existing LLMs-based DocRE methods without fine-tuning, PromptRE and DocGNRE.

Implementation Details We set parameters, including AdamW (Loshchilov and Hutter, 2019), warmup (Goyal et al., 2017), and learning rates as origin for SLMs to train and obtain F_{slm^k}. In all subsequent experiments, we test on an A100 with 40GB GPU, use LLaMA3-8B (Touvron et al., 2023) as the LLM, set τ to 1.8, top-k to 4, ξ to 0.03, and δ to 0.6/0.5 for two datasets.

Moreover, we conduct experiments without SLMs (i.e., Only-LLM as also mentioned in Section 3.2), and without the task distribution and probability fusion (denote as $\text{Refiner}_{\neg TD\&PF}$, which integrates the top-k relations predicted by SLM ATLOP for each entity pair into the LLM's prompt to obtain the final prediction result).

5.2 Main Results

Table 3 and 4 report the performance of our method and existing methods on two DocRE datasets.

(1) Compared with LLMs and earlier SLMs: In Table 3, our Only-LLM method demonstrates more competitive than previous LLM-based methods, and obtains improvements of **10.95** and **21.36** F1 on two datasets. Further, $\text{Refiner}_{\neg TD\&PF}$ achieves an average F1 improvement of **17.99** on DocRED and **32.49** on Re-DocRED. However, the performance of LLM-based methods is still unsatisfactory even compared to the earlier SLMs, as there remains a gap in F1 scores, as shown in Table 3 and Table 4, which suggests that the refinement may be necessary and valuable.

(2) Compared SLMs with our Refiner: The F1 and Ign F1 scores of SLMs with our Refiner show *consistent improvement* across the DocRED

Model	DocRED				Re-DocRED			
	P	R	F1	Ign F1	P	R	F1	Ign F1
PromptRE (Gao et al., 2023)	-	-	-	-	6.56	27.00	10.56	9.04
DocGNRE (Li et al., 2023)	14.61	9.8	11.73	-	24.45	5.77	9.33	-
Only-LLM (Ours)	23.63	21.81	22.68	21.37	43.83	25.11	31.92	30.48
Refiner$_{\neg TD\&PF}$ (Ours)	**37.14**	**24.77**	**29.72**	**28.55**	**63.33**	**31.89**	**42.43**	**41.67**

Table 3: Evaluation results of our method based mainly on LLM, with best scores **bold**. The results of LLMs-based PromptRE and DocGNRE are from their original papers.

Model	DocRED				Re-DocRED			
	Dev		Test		Dev		Test	
	Ign F1	F1	Ign F1	F1	Ign F1	F1	Ign F1	F1
(a) Earlier SLMs								
CNN (Yao et al., 2019)	41.58	43.45	40.33	42.26	-	-	-	-
BiLSTM (Yao et al., 2019)	48.87	50.94	48.78	51.06	-	-	-	-
(b) SLMs with Refiner								
ATLOP (Zhou et al., 2021)	59.11	61.01	59.31	61.30	76.79	77.46	76.82	77.56
ATLOP $_{+\text{Refiner}}$	↑59.42	↑61.31	↑59.55	↑61.62	↑77.48	↑78.05	↑77.32	↑78.09
Eider (Xie et al., 2022)	60.51	62.48	60.42	62.47	†75.91	†76.99	†76.25	†77.13
Eider $_{+\text{Refiner}}$	↑60.81	↑62.90	↑60.71	↑62.88	↑76.43	↑77.52	↑76.84	↑77.61
DREEAM (Ma et al., 2023a)	63.47	65.30	63.31	65.30	†79.51	†80.66	79.66	80.73
DREEAM $_{+\text{Refiner}}$	↑**63.71**	↑**65.69**	↑**63.47**	↑**65.82**	↑80.62	↑81.58	↑80.45	↑81.69
AA (Lu et al., 2023)	61.31	63.38	60.84	63.10	80.04	81.15	80.12	81.20
AA $_{+\text{Refiner}}$	↑61.82	↑63.88	↑61.46	↑63.67	↑**80.92**	↑**82.01**	↑**80.93**	↑**82.03**

Table 4: Evaluation results of SLMs and refined with Refiner on DocRED and Re-DocRED, with best scores **bold**, and the effect of refine is indicated by arrows. The results of SLMs are from their original papers while others with † are obtained by our reproduction.

Model	Ign F1	F1
Refiner $_{+\text{DREEAM}}$	80.45	81.69
w/o self-supervision	80.35	81.48
w probability fusion$_{add}$	79.57	80.01
w/o probability fusion	78.51	79.24
w/o task distribution	45.32	46.78

Table 5: Ablation study on Re-DocRED test set. "w probability fusion$_{add}$" means that we simply add the probabilities.

and Re-DocRED datasets (average improvement of **0.4∼0.9** F1). Notably, the F1 scores of the DREEAM and AA refiner achieve **new state-of-the-art** performance compared with the existing SLMs, indicating that our refiner is effective for enhancing the performance of DocRE tasks.

5.3 Ablation Study

We investigate the effectiveness of modules in Refiner by removing them in turn. We show our results in Table 5:

(1) Task distribution is extremely effective for the use of LLMs in refined method. Removing it results in a dramatic decline in performance (**35.13** and **34.91 drop** in terms of Ign F1 and F1). It indicates that filtering a large number of negative samples (i.e., NA) further enhances the LLM's attention to the predicted relation labels.

(2) The use of balanced probability fusion indeed improves performance. Removing it (i.e., w/o probability fusion), F1 score decrease by **2.45**. Also, instead of our probability balance fusion method, we simply *add* the probability distributions F_{slm^k} and $F^{\tau}_{llm^k}$ together, which also leads to the decline of **1.68** F1. All of these highlight the effectiveness of probability balance fusion.

(3) Self-supervision technique brings a slight improvement by guiding LLMs to confirm whether the refined labels are correct, thereby enhancing the prediction accuracy.

5.4 Impact of Probability Balance Estimation

In order to further verify the impact of whether two distributions are balanced on model performance, we visualize the distribution difference of variances σ_{llm} and σ_{slm} for each entity pair at $\tau = 1.1$ and $\tau = 1.8$ in Figure 5. The results show that the variance difference of the LLM and SLM are more

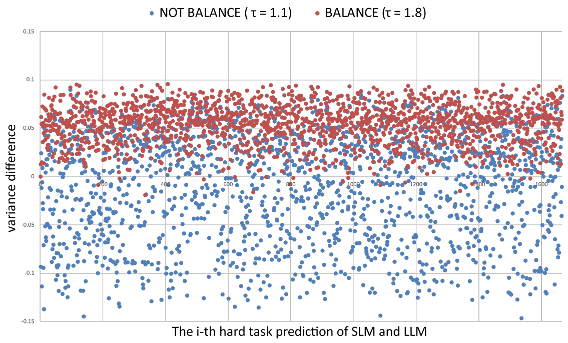


Figure 5: Visualization of balanced and imbalanced distributions on the Re-DocRED dataset.

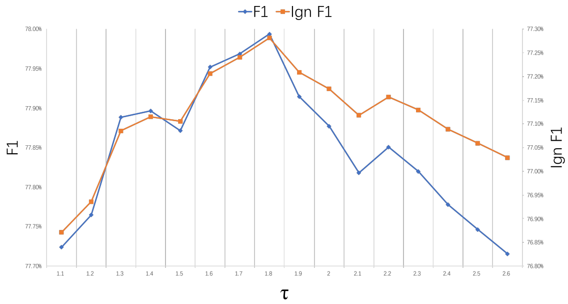


Figure 6: The impact of τ on F1 and Ign F1 scores for refined SLM ATLOP on Re-DocRED dataset.

discrete when τ is 1.1. As τ increases, the variance difference tend to converge. Further, Figure 6 demonstrates that as τ increases, F1 and Ign F1 exhibit an increasing trend followed by a decreasing trend. These indicate that better balance between the SLM and LLM prediction distributions indeed helps improve prediction outcomes.

5.5 Impact of δ and top-k in Task Distribution

δ is denoted as the hard task threshold in Eq. (7). As δ increases, tasks where $P_{slm}(r)$ is much lower than $P_{slm}(NA)$ will also be considered hard tasks. That is, the LLM will handle more tasks. Consequently, the number of correctly and incorrectly refined entity pairs increases, as shown in Figure 7.

We also evaluate the top-k results in Table 6, which show that, the proportion of the first k output labels containing the correct relation labels, increases significantly from top-1 to top-3 but more modest for top-4 and top-5.

	top-1	**top-2**	**top-3**	**top-4**	**top-5**
Hit(%)	72.51	89.02	93.69	95.47	96.98

Table 6: The impact of top-k in task distribution.

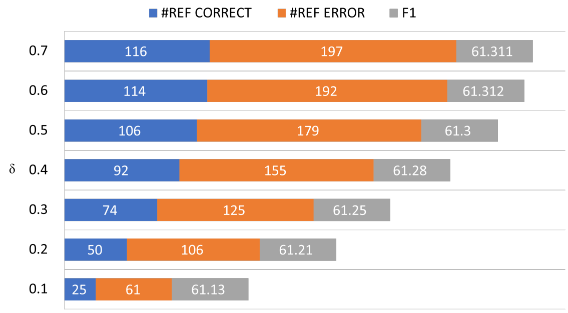


Figure 7: The impact of δ on F1 score, the number of correctly and incorrectly refined entity pairs (#Ref correct and #Ref error) for SLM ATLOP on DocRED.

Experiment	Time(Hour)	Cost(US$)
(a) *without task distribution*		
Only-LLM	24.21	100.91
Refiner$_{\neg TD\&PF}$	28.69	-
Self-supervision prompting	28.77	-
(b) *with task distribution*		
Refiner	1.21	-
Self-supervision prompting	1.07	-

Table 7: Analyze average time and cost of reasoning using different LLMs and templates with/without task distribution on DocRED and Re-DocRED. We use GPT-3.5-turbo-0613 for Only-LLM Template as mentioned in Section 3.1 and LLaMA3-8B for other Templates as mentioned in Section 5.1.

5.6 Cost Analysis and Case Study

Table 7 shows our method's efficiency. Without task distribution, inference time increases significantly due to the presence of a large number of negative samples (i.e., NA). Applying task distribution drastically reduces this time, highlighting the efficiency gains. For deep cost-benefit analysis, we provide a more detailed explanation of the computational requirements of our refiner framework in Appendix D.

Several interesting case studies illustrate the difference in probability fusion before and after balance adjustment. Due to space limitations, cases are provided in Appendix E.

Proportion(%)	**DocRED**	**Re-DocRED**
#train NA	96.81	92.80
#dev NA	96.90	91.06
#test Predicted as NA	33.02	24.20
#test Ign F1	63.48	79.63
#test F1	65.31	80.71

Table 8: Effect of proportion of NA in different dataset on SLM DREEAM.

Case	True Label	LLM Prediction
Robert Kingsbury Huntington (13 March 1921 – 5 June 1942), was a naval aircrewman and member of Torpedo Squadron 8 (or VT-8). ... Huntington was one of 29 from Torpedo Squadron 8 who gave their lives in this attack. ##QUESTION: Which of the following is right? A. Battle of Midway(MISC) isn't an administrative entity, and Battle of Midway(MISC) was a physical object or event in Japanese(LOC). ... D. None of the above options is correct.	D	A
"More" is a song by The Sisters of Mercy, from their album Vision Thing. ... The song has also been re-recorded by Meat Loaf for his 2016 album Braver Than We Are. ##QUESTION: Which of the following is right? ... B. MTV(MISC) is a production company, and Wuthering Heights(MISC) was produced by MTV(MISC). ... D. None of the above options is correct.	D	B

Table 9: Case study of LLM predictions conflicting with ground truth.

5.7 Relationship Validation between Dataset Characteristics and Model Behaviors

To enhance rigor of our validation between dataset characteristics and model behaviors, we provide additional examining experiments.

In Table 8, we observe that compared to DocRED, the proportion of NA in the Re-DocRED train set decreases by 4.01. With consistent NA proportions in train/dev sets, the same model structure, DREEAM, shows an 8.82 reduction in *incorrectly predicting relations as NA*. Performance improvement is also observed, suggesting that SLM NA prediction bias is related to label distribution.

Table 9 shows cases about *LLM predictions conflicting with ground truth*. From the cases, we observe that when certain information is not explicitly presented or requires additional evidence, the LLM prefers relation-indicating options over option D. The complete cases can be found in Appendix F.

6 Conclusion

In this paper, we present a noteworthy finding that LLMs tend to predict the existence of relations for all entity pairs, while SLMs often classify existing relations as NA. We propose a novel method utilizing LLMs as refiners, employing task distribution and self-supervised probability fusion. Our refiner demonstrates significant improvement over SLMs and LLMs methods and achieves state-of-the-art performance. Our methods do not require retraining of LLMs and can be flexibly integrated with various SLM approaches on DocRE, incurring an acceptable cost, while significantly enhancing the performance of all SLMs' original capabilities.

Limitations

We do not test our refiner on the graph-based models considering of LLMs and transformer-based SLMs have the same architecture. Moreover, we conduct experiments on LLMs (LLaMA3-8B for our refiner) without larger model parameters, and we may not have measured the upper limit of refiner. In addition, when we adjust the threshold δ, we increase both the number of correctly predicted entity pairs and the number of incorrectly predicted entity pairs by LLM. The incorrectly results may also affect the final prediction performance. In our future work, we will conduct more extensive research on graph-based SLMs, and try to select models with larger parameters for refinement.

Acknowledgments. The authors thank reviewers for their valuable comments. The work is supported by National Natural Science Foundation of China (62276057), and Sponsored by CAAI-MindSpore Open Fund, developed on OpenI Community.

References

Julien Delaunay, Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Georgeta Bordea, Nicolas Sidere, and Antoine Doucet. 2023. A comprehensive survey of document-level relation extraction (2016-2023). *arXiv e-prints*, pages arXiv–2309.

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 (NAACL)*.

Bayu Distiawan, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 229–240.

Chufan Gao, Xulin Fan, Jimeng Sun, and Xuan Wang. 2023. Promptre: Weakly-supervised document-level relation extraction via prompting-based data programming. *arXiv preprint arXiv:2310.09265*.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clay Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about gpt-3 in-context learning for biomedical ie? think again. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Junpeng Li, Zixia Jia, and Zilong Zheng. 2023. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. *Conference on Empirical Methods in Natural Language Processing 2023 (EMNLP)*.

Xue Lilong, Zhang Dan, Dong Yuxiao, and Tang Jie. 2024. Autore: Document-level relation extraction with large language models. *arXiv preprint arXiv:2403.14888*.

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. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *In International Conference on Learning Representations (ICLR)*.

Chonggang Lu, Richong Zhang, Kai Sun, Jaein Kim, Cunwang Zhang, and Yongyi Mao. 2023. Anaphor assisted document-level relation extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15453–15464.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023a. Dreeam: Guiding attention with evidence for improving document-level relation extraction. *The 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Yubo Ma, Yixin Cao, Yong Ching Hong, and Aixin Sun. 2023b. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yu Shi, Jiaming Shen, Yuchen Li, Naijing Zhang, Xinwei He, Zhengzhi Lou, Qi Zhu, Matthew Walker, Myunghwan Kim, and Jiawei Han. 2019. Discovering hypernymy in text-rich heterogeneous information network by exploiting context granularity. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 599–608.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-augmented language models. *In International Conference on Learning Representations (ICLR)*.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 30.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL)*.

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 *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL).*

Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. 2023a. Small models are valuable plug-ins for large language models. *arXiv preprint arXiv:2305.08848.*

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023b. How to unleash the power of large language models for few-shot relation extraction? *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL).*

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

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. *In International Conference for Learning Representation (ICLR).*

Shuang Zeng, Yuting Wu, and Baobao Chang. 2021. Sire: Separate intra-and inter-sentential reasoning for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (ACL)*, pages 524–534.

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

Kai Zhang, Bernal Jiménez Gutiérrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. *Findings of the Association for Computational Linguistics (ACL Findings).*

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, pages 14612–14620.

A Prompt Templates

A.1 Only-LLM Template

Our only-LLM prompt template used in Section 3.1 consists of *Instruction* + *Document* + *Question* + *Answer* as shown in Table 14. The Question is constructed by filling in the head and tail entities of the question through the question template of Table 17. NA is used as the last option. We aim to obtain the probability distribution of the first token after the Answer and select all the multiple-choice question letters greater than NA as the final relation prediction results.

A.2 Multi-choice Template

Our multiple-choice prompt template consists of *Instruction* + *Document* + *Question* + *Answer* as well. Unlike the only-LLM template, the prompt we use for task distribution selects only the most likely top-k relations predicted by the SLM. We obtain the probability distribution of the first token after the Answer and select all the multiple-choice letters with probabilities greater than NA as the final multi-label relation prediction results. The multiple-choice prompt is shown in Table 15.

A.3 Self-supervision Judgment Template

Our judgment prompt template consists of *Instruction* + *Document* + *Question* + *Answer* as shown in Table 16. Unlike the multiple-choice template, the self-supervision part does not use multiple-choice questions but judgment questions. We provide a document and a sentence (i.e., each multiple-choice with probabilities greater than NA as the final relation prediction) and let the LLM judge whether the sentence is T or F. If the probability of T is greater than the probability of F, we consider the result of this refinement acceptable; otherwise, we discard the result of this refinement.

B Hyper-Parameters of SLMs and LLMs

For the refinement of SLMs, we reproduce their prediction logits using the same parameter settings as outlined in their respective papers. For Refiner$_{\neg TD\&PF}$, we set the temperature to 0 to make the results more stable, and we choose different δ thresholds for different DocRE datasets. Detailed parameter selections of SLMs and LLMs are provided in Table 10.

C Impact of δ on Re-DocRED Dataset

δ is denoted as the hard task threshold in Eq. (7). As the equation defines, if δ increases, the hard task threshold will range up, meaning that tasks with $P_{slm}(r)$ much lower than $P_{slm}(NA)$ will also be considered as hard tasks. Consequently, the LLM will handle more tasks. Figure 8 shows the changes of F1, we can see that the best F1 score of the Refiner on Re-DocRED is 78.05 with δ set to 0.5.

Model	Parameters
ATLOP	adm ϵ = 1e-6
	learning_rate = 5e-5
	warmup_ratio = 0.06
	num_train_epochs = 30
Eider	learning_rate = 5e-5
	learning_rate(other) = 1e-4
	warmup_ratio = 0.06
	num_train_epochs = 30
DREEAM	learning_rate(encoder) = 5e-5
	learning_rate(classifier) = 1e-4
	warmup_ratio = 0.06
	num_train_epochs(teacher) = 30
	num_train_epochs(student) = 10
AA	learning_rate(encoder) = 5e-5
	learning_rate(classifier) = 1e-4
	warmup_ratio = 0.06
	num_train_epochs = 30
$\text{Refiner}_{\neg TD\&PF}$	max_new_tokens = 1
	top-k = 4
	top_p = 0.9
	temperature = 0
Refiner	top-k = 4
	temperature = 1.8
	ξ = 0.03
	δ(DocRED) = 0.6
	δ(Re-DocRED) = 0.5
	top_p = 0.9
	max_new_tokens = 1

Table 10: The hyper-parameters of SLMs and LLMs.

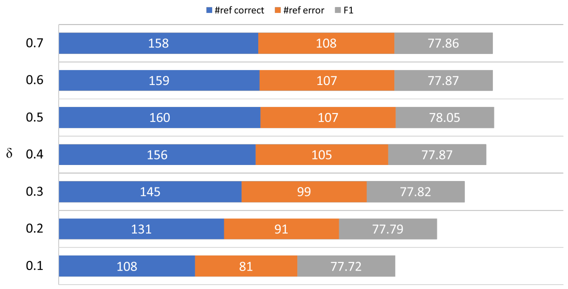


Figure 8: Impact of δ on F1 score, the number of correctly and incorrectly refined entity pairs for SLM ATLOP on Re-DocRED dataset.

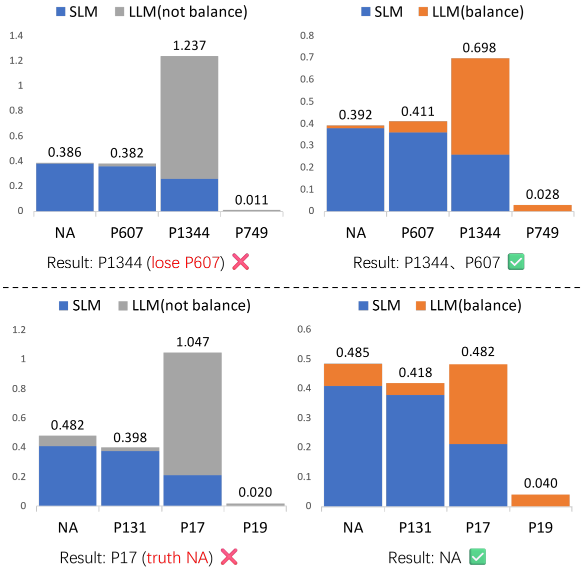


Figure 9: Case Study of Probability Fusion.

D Computational Requirements

Combined with our cost analysis in Section 5.6, we further provide a more detailed analysis of the computational requirements of our refiner framework. It is important to emphasize that the LLaMA3-8B model we used does not involve the fine-tuning process. We analyze the hour time taken for training the SLM and the inference for LLaMA3-8B on the refiner task. Our hardware environment consists of an A100 with 40GB GPU, and the LLaMA3-8B model runs and performs inference based on the Transformer library.

From Table 11, we can analyze that the LLaMA3-8B inference time does not significantly contribute to the overall inference cost. Therefore, when using LLMs for single entity pair prediction, the computational requirements increase in a reasonable way compared to SLM-only approaches.

Our proposed Task Distribution effectively enhances the cost-benefit of our model framework. This is because NA entities make up a very large proportion of the dataset. If all entity pairs are handed over to the LLM for relation judgment, requiring the LLM to analyze each entity pair individually would create a significant computational burden.

In Table 12, we can observe that due to the large number of NA entity pairs, direct inference incurs significantly higher inference costs compared to using Task Distribution, and the performance is not optimal. This further highlights that, for the DocRE task, considering Task Distribution for inference costs is a more reasonable approach to assess the real-world applicability.

E Case Study of Probability Fusion

In Figure 9, we present cases illustrating how the probability fusion affects the final decision. It is evident that LLM tends to be overly confident in its final prediction, leading to inaccuracy. However, when a better balance between SLM and LLM is achieved, which results in an accurate prediction.

	SLM Train	LLM Inference	Total	Inference Proportion(%)
ATLOP	2.23	-	2.23	-
ATLOP$_{+Refiner}$	2.23	0.44	2.67	16.48
Eider	1.61	-	1.61	-
Eider$_{+Refiner}$	1.61	0.49	2.1	23.33
DREEAM	2.15	-	2.15	-
DREEAM$_{+Refiner}$	2.15	0.52	2.67	19.48

Table 11: Analysis the SLM training time and LLM inference time (hours).

	GPT-3.5-Turbo			LLaMA3-8B		
	Cost ($)	Ign F1	F1	Time (hours)	Ign F1	F1
Direct Inference	100.91	24.17	27.65	28.69	30.48	31.92
With Task Distribution	4.96	33.34	34.01	1.21	41.67	42.43

Table 12: Analysis comparing the inference cost of GPT-3.5-Turbo and LLaMA3-8B under two scenarios: using Task Distribution and direct inference.

F Case study of LLM Predictions Conflicting with Ground Truth

As the LLM used in our study was not fine-tuned, we conducted a case study to illustrate that such conflicts are indeed worth investigating and may contribute to the suboptimal performance of LLMs in their current, non-fine-tuned state. Below, we provide examples highlighting the discrepancies between LLM predictions and the ground truth, underscoring the need for further exploration of this phenomenon. The case results are presented in Table 13.

From these cases, we observe that when certain information is not explicitly presented in the document or requires additional evidence for verification, the LLM tends to rely on its own knowledge to infer the correctness of a given option. This often leads the LLM to favor relation-indicating options over selecting option D. We hope these examples provide a clearer understanding of how LLM predictions can conflict with the ground truth.

G Question Relation Templates

By observing the description of each relation in DocRED, we can find that there are many relations, which only using the form of *[head + relationship + tail]* cannot express the true meaning of these relations. To help LLMs better understand each relation label, we design our own question relation templates for DocRED and Re-DocRED datasets.

Our design idea mainly includes two parts: first, the template needs to describe the entity (e.g, the entity should be an individual, or a region.), second, construct a grammatical sentence for the head and tail entities and relations. The question relation template we designed is shown in Table 17.

Table 13: Case study of LLM predictions conflicting with ground truth.

Case	True Label	LLM Prediction
Robert Kingsbury Huntington (13 March 1921 – 5 June 1942), was a naval aircrewman and member of Torpedo Squadron 8 (or VT-8). He was radioman / gunner to Ensign George Gay's TBD Devastator aircraft. Along with his entire squadron, Huntington was shot down during the Battle of Midway, on 4–5 June 1942. Born in Los Angeles, California, enlisted in the United States Navy 21 April 1941. He received the Distinguished Flying Cross for heroism and extraordinary achievement as rear gunner in a torpedo plane during an attack against enemy Japanese forces in the Battle of Midway 4 June 1942. Flying without fighter support and with insufficient fuel to return to their carrier, Huntington and his fellow crewmember pressed home their attack with utter disregard for their own personal safety, in the face of a tremendous antiaircraft barrage and overwhelming fighter opposition. Huntington was one of 29 from Torpedo Squadron 8 who gave their lives in this attack. ##QUESTION: Which of the following is right? A. Battle of Midway(MISC) isn't an administrative entity, and Battle of Midway(MISC) was a physical object or event in Japanese(LOC). B. Japanese(LOC) is a country, and Battle of Midway(MISC) isn't a person. Japanese(LOC) is the sovereign state of Battle of Midway(MISC). C. Battle of Midway(MISC) is owned by Japanese(LOC). D. None of the above options is correct.	D	A
"More" is a song by The Sisters of Mercy, from their album Vision Thing. It was the first single from the album, reaching number one on the Billboard Modern Rock Tracks chart for five weeks, starting 15 December 1990. The song was co-written and co-produced by Andrew Eldritch and Jim Steinman. It was covered by Shaaman on their album Reason, and Gregorian for their album The Dark Side. Steinman produced a cover of the song, by Mike Vogel and Erika Christensen, for the soundtrack of the MTV film Wuthering Heights. He also used the song's main guitar riff and the" I need all the love I can get" vocal in a song for his musical Batman. The song has also been re-recorded by Meat Loaf for his 2016 album Braver Than We Are. ##QUESTION: Which of the following is right? A. Wuthering Heights(MISC) was a radio or television show, and Wuthering Heights(MISC) was aired on or included by MTV(MISC). B. MTV(MISC) is a production company, and Wuthering Heights(MISC) was produced by MTV(MISC). C. MTV(MISC) is a class, and Wuthering Heights(MISC) is an individual member of MTV(MISC). D. None of the above options is correct.	D	B

Dollar General Corporation is an American chain of variety stores headquartered in Goodlettsville, Tennessee. As of July 2018, Dollar General operates 15,000 stores in 45 of the 48 contiguous United States (the exceptions being three states in the northwest : Idaho, Montana, and Washington). The company first began in 1939 as a family-owned business called J.L. Turner and Son in Scottsville, Kentucky by James Luther Turner and Cal Turner. In 1968, the name changed to Dollar General Corporation and the company went public on the New York Stock Exchange. Fortune 500 recognized Dollar General in 1999 and in 2018 reached # 123. Dollar General has grown to become one of the most profitable stores in the rural United States with revenue reaching around $ 21 billion in 2017. ##QUESTION: Which of the following is right? A. American(LOC) is a country, and Dollar General(ORG) isn't a person. American(LOC) is the sovereign state of Dollar General(ORG). B. American(LOC) is a country, and Dollar General(ORG) was originally made in American(LOC). C. American(LOC) is a country, and Dollar General Corporation(ORG) is a person. Dollar General Corporation(ORG) is a citizen of American(LOC). D. None of the above options is correct.	D	B
Live in New York was a 2-CD live album released by performance artist Laurie Anderson on Nonesuch Records in 2002. It was her ninth album of new recordings released since 1982. The front cover of the CD has the title Live at Town Hall, New York City September 19–20, 2001, however the official title of the album is just Live in New York. Recorded less than 10 days after the September 11, 2001, attacks on New York City, the album was produced during a tour Anderson gave of the United States in which she performed a mixture of older pieces from earlier in her career and newer works, including songs from her then-recent album Life on a String, as well as earlier albums such as United States Live, Big Science, Bright Red, Home of the Brave and Strange Angels. ##QUESTION: Which of the following is right? A. Big Science(MISC) is the prior version in series, followed by Bright Red(MISC). B. Big Science(MISC) and Bright Red(MISC) aren't in series, and Big Science(MISC) is replaced by Bright Red(MISC), so Big Science(MISC) will never take place again. C. Big Science(MISC) is the next version in series, following Bright Red(MISC). D. None of the above options is correct.	D	C

Only-LLM Prompt	**Example**
## INSTRUCTION: [The instruction for the LLM] ## DOCUMENT: [The related document of the entity pair] ## QUESTION: [The statements for LLM to choose] ## ANSWER:	## INSTRUCTION: Read the ##DOCUMENT and answer the ##QUESTION. Write the answers in ##ANSWER. ## DOCUMENT: Skai TV is a Greek free-to-air television network based in Piraeus. It is part of the Skai Group, one of the largest media groups in the country. It was relaunched in its present form on 1st of April 2006 in the Athens metropolitan area, and gradually spread its coverage nationwide. Besides digital terrestrial transmission, ..., all foreign shows. ## QUESTION: Which of the following is right? 1. Greece(LOC) is a country, and Skai TV(ORG) isn't a person. Greece(LOC) is the sovereign state of Skai TV(ORG). 2. Greece(LOC) is an administrative entity, and Skai TV(ORG) is located on the territory of Greece(LOC). 3. Greece(LOC) is a country, and Skai TV(ORG) was originally made in Greece(LOC). 4. .... [all relations] 97. None of the above options is correct. ## ANSWER: [TOKEN]

Table 14: The only-LLM prompt template and an example for an entity pair.

Multiple-choice Prompt	**Example**
## INSTRUCTION: [The instruction for the LLM] ## DOCUMENT: [The related document of the entity pair] ## QUESTION: [The statements for LLM to choose] ## ANSWER:	## INSTRUCTION: Read the ##DOCUMENT and answer the ##QUESTION. Write the answers in ##ANSWER. ## DOCUMENT: Skai TV is a Greek free-to-air television network based in Piraeus. It is part of the Skai Group, one of the largest media groups in the country. It was relaunched in its present form on 1st of April 2006 in the Athens metropolitan area, and gradually spread its coverage nationwide. Besides digital terrestrial transmission, ..., all foreign shows. ## QUESTION: Which of the following is right? A. Greece(LOC) is a country, and Skai TV(ORG) isn't a person. Greece(LOC) is the sovereign state of Skai TV(ORG). B. Greece(LOC) is an administrative entity, and Skai TV(ORG) is located on the territory of Greece(LOC). C. Greece(LOC) is a country, and Skai TV(ORG) was originally made in Greece(LOC). D. None of the above options is correct. ## ANSWER: [TOKEN]

Table 15: The multiple choice prompt template and an example for an entity pair.

Judgment Prompt	**Example**
## INSTRUCTION: [The instruction for the LLM] ## DOCUMENT: [The related document of the entity pair] ## QUESTION: [The statements for LLM to choose] ## ANSWER:	## INSTRUCTION: Read the ##DOCUMENT and answer the ##QUESTION. Write the answers in ##ANSWER. ## DOCUMENT: Skai TV is a Greek free-to-air television network based in Piraeus. It is part of the Skai Group, one of the largest media groups in the country. It was relaunched in its present form on 1st of April 2006 in the Athens metropolitan area, and gradually spread its coverage nationwide. Besides digital terrestrial transmission, ..., all foreign shows. ## QUESTION: True or False? Only return T or F. Greece(LOC) is a country, and Skai TV(ORG) isn't a person. Greece(LOC) is the sovereign state of Skai TV(ORG). ## ANSWER: [TOKEN]

Table 16: The judgment prompt template and an example for an entity pair.

Table 17: Question Relation Templates, where {head} and {tail} are the placeholders for subject and object.

ID	Relation	Template
P6	head of government	{tail} is a person, and {head} is a governmental body. {tail} is the head of {head}.
P17	country	{tail} is a country, and {head} isn't a person. {tail} is the sovereign state of {head}.
P19	place of	birth {tail} is a specific location, and {head} was born in {tail}.
P20	place of	death {tail} is a specific location, and {head} died in {tail}.
P22	father	{head} and {tail} are both people, {tail} is {head}'s biological father.
P25	mother	{head} and {tail} are both people, {tail} is {head}'s biological mother.
P26	spouse	{head} and {tail} are spousal.
P27	country of citizenship	{tail} is a country, and {head} is a person. {head} is a citizen of {tail}.
P30	continent	{tail} is a continent, and {head} is part of {tail}.
P31	instance of	{tail} is a class, and {head} is an individual member of {tail}.
P35	head of state	{tail} is a person, and {head} is a country or state. {tail} is the head of {head}.
P36	capital	{head} is an administrative territorial entity, and {tail} is a captial of {head}.
P37	official language	{head} 's official language is {tail}.
P39	position held	{head} is a person, and {head} holds {tail} position.
P40	child	{head} has {tail} in their family as their child.
P50	author	{tail} isn't a person, and the author of {tail} is {head}.
P54	member of sports team	{tail} is a sports team or club, and {head} plays for {tail}.
P57	director	{head} is a work directed by {tail}.
P58	screenwriter	{tail} is the author of the script for {head}.
P69	educated	{tail} is an educational institution, and {head} was educated in {tail}.
P86	composer	{head} is the music wrote by {tail}.
P102	member of political party	{tail} is a political party, and {head} has been a member of {tail}.
P108	employer	{tail} is a person or organization, and {head} worked for {tail}.
P112	founded by	{head} is a organization, religion or place, and {tail} is a founder or co-founder of {head}.
P118	league	{head} is a team, and {head} is in {tail} league.
P123	publisher	{tail} is organization or person, and {head} is published by {tail}.
P127	owned by	{head} is owned by {tail}.
P131	located in the administrative territorial entity	{tail} is an administrative entity, and {head} is located on the territory of {tail}.
P136	genre	{tail} is a work's genre in which {head} worked.
P137	operator	{tail} is a person or organization, and {tail} is the operator of {head}.
P140	religion	{tail} is a religion, and {tail} is the religion of {head}.

P150	contains administrative territorial entity	{head} is administrative territorial entity, and {tail} is the direct subdivision of {head}.
P155	follows	{head} is the next version in series, following {tail}.
P156	followed by	{head} is the prior version in series, followed by {tail}.
P159	headquarters location	{head} is a organization, and {head} is based in {tail}.
P161	cast member	{tail} is a cast member of {tail}.
P162	producer	{tail} is a person, and {tail} is the producer of {head}.
P166	award received	{head} received the award called {tail}.
P170	creator	{head} is a work or fictional object, created by {tail}.
P171	parent taxon	{tail} is the closest parent taxon of {head}.
P172	ethnic group	The ethnic group of {head} is {tail}.
P175	performer	{tail} is the performer of {head}.
P176	manufacturer	{head} are made by manufacturer {tail}.
P178	developer	{tail} is a person or organisation, and is the developer of {head}.
P179	series	{tail} is a series, and {head} is part of this series.
P190	sister city	{head} is twinned with {tail}, so they 're sister cities to each other.
P194	legislative body	{tail} is a political institution, as the legislative body governing {head}.
P205	basin country	{tail} is a country, and {head} is the drainage or body of water created by {tail}.
P206	located in or next to body of water	{tail} is sea, lake or river, and {head} is located in or next to body of {tail}.
P241	military branch	{tail} is a military branch, and {head} belongs to {tail}.
P264	record label	{tail} is a record company, and {head} was released on {tail}.
P272	production company	{tail} is a production company, and {head} was produced by {tail}.
P276	location	{head} isn't an administrative entity, and {head} was a physical object or event in {tail}.
P279	subclass of	{head} and {tail} are two classes, and {head} is a subclass of {tail}.
P355	subsidiary	{head} is a company or organization, and {tail} is the subsidiary of {head}.
P361	part of	{head} is a part of {tail}.
P364	original language of work	{head} is a film or performance work, and {tail} is the original language of {head}.
P400	platform	{tail} is a released platform, and {head} was released for {tail}.
P403	mouth of the watercourse	{head} is a watercourse, and {head} drains into the body of {tail}.
P449	original network	{head} was a radio or television show, and {head} was aired on or included by {tail}.
P463	member of	{tail} isn't an ethinc or social groups, and {head} belongs to {tail}.
P488	chairperson	{head} is an organization, group or body, and {tail} is the chairperson of {head}.
P495	country of origin	{tail} is a country, and {head} was originally made in {tail}.
P527	has part	{tail} is a part of {head}.

P551	residence	{tail} is a place, and {head} is a person. The {tail} is the place where {head} is, or has been, resident.
P569	date of birth	{tail} is a time, and {head} was born on {tail}.
P570	date of death	{tail} is a time, and {head} was died on {tail}.
P571	inception	{tail} is a time point, and {head} was firstly founded on {tail}.
P577	publication date	{tail} is a time point, and {head} was a work firstly published on {tail}.
P580	start time	{tail} is a time, and {head} started being valid in {tail}.
P582	end time	{tail} is a time, and {head} stopped being valid in {tail}.
P585	point in time	{tail} is a time point, and {head} took place at this point in {tail}.
P607	conflict	{tail} is a battle, was or other military engagement, and {head} participated in {tail}.
P674	characters	{head} is a work, and {tail} is one of characters in {head}.
P676	lyrics by	{head} is a song, and the lyrics of {head} were written by {tail}.
P706	located on terrain feature	{tail} is a specified landform, and {head} is located on {tail} according to the terrain feature.
P710	participant	{head} is an event, and {tail} participated in {head}.
P737	influenced by	{head} was a person or idea, etc, and {head} was influenced by {tail}.
P740	location of formation	{tail} is a location, and {head} is a group or organization formed in {tail}.
P749	parent organization	{head} is a company or organization, and {tail} is the parent organization of {head}.
P800	notable work	{tail} is a notable work, and {tail} is one of {head}'s works.
P807	separated from	{head} emerged after the collapse or separation of {tail}.
P840	narrative location	{head} is a work or story, and {head} is about what happened in {tail}.
P937	work location	{tail} is a location, and {head} worked in the past or is working now.
P1001	applies to jurisdiction	{head} has the territorial jurisdiction of {tail}.
P1056	product or material produced	{tail} was the material or product produced by {head}.
P1198	unemployment rate	{tail} as the best competition record of {head} in some event.
P1336	territory claimed by	{head} is an area, and {head} is administered by {tail}.
P1344	participant of	{head} is a person or an organization, and {tail} is an event. {head} participated in {tail}.
P1365	replaces	{head} and {tail} aren't in series, and {head} replaces {tail}, so {tail} will never take place again.
P1366	replaced by	{head} and {tail} aren't in series, and {head} is replaced by {tail}, so {head} will never take place again.
P1376	capital of	{tail} is an administrative division, and {head} is the capital of {tail}.
P1412	languages spoken, written or signed	{tail} is a person, and {head} is the language that {tail} speaks or writes.
P1441	present in work	{head} is a fictional entity or historical person, and {head} is present in the work named {tail}.
P3373	sibling	{head} and {tail} are siblings.