LLM with Relation Classifier for Document-Level Relation Extraction

Anonymous ACL submission

Abstract

Large language models (LLMs) create a new paradigm for natural language processing. Despite their advancement, LLM-based methods still lag behind traditional methods in document-level relation extraction (DocRE), a critical task for understanding complex entity relations. To address this issue, this paper first investigates the causes of the performance gap, identifying the dispersion of attention by LLMs due to entity pairs without relations as a primary factor. We then introduce a novel classifier-LLM approach to DocRE. The proposed approach begins with a classifier specifically designed to select entity pair candidates exhibiting potential relations and thereby feeds them to LLM for the final relation extraction. This method ensures that during inference, the LLM's focus is directed primarily at entity pairs with relations. Experiments on DocRE and Re-DocRE benchmarks reveal that our method significantly outperforms recent LLM-based DocRE methods.

1 Introduction

Document-level Relation Extraction (DocRE) aims to extract relations between entity pairs within crossing sentences in one document. Prior DocRE models emulate the process of reading and reasoning on entity pairs throughout the entire document using advanced neural network architectures, including self-attention networks (Tan et al., 2022a), and GNNs (Li et al., 2020), and achieved a SOTA performance (Ma et al., 2023).

Recently, Sun et al. tried to utilize LLM to simulate DocRE by using a chain-of-retrieval prompt. However, the LLM-based method still lags behind traditional approaches in DocRE. We observed that following the definition of the DocRE task, all possible entity pairs (referred to as candidate space) are constructed and fed into LLMs, and within this extensive array of entity pairs, only a select few harbor relations. Our preliminary experiments on two widely used DocRED (Yao et al., 2019) and Re-DocRED (Tan et al., 2022b) datasets showed that this phenomenon leads to an imbalance in the candidate space, which may make LLMs focus more on *NA* entity pairs that do not express any relation. Consequently, the identified factor is regarded as one of the main causes of the performance deficiencies observed in LLMs on DocRE.

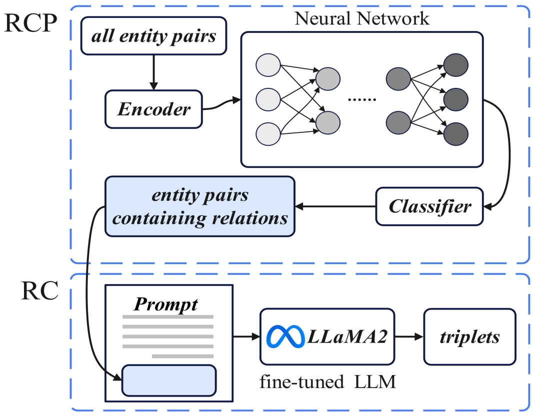


Figure 1: **Illustration of LMRC**. Relation Candidate Proposal(RCP) leverages localized context pooling (Zhou et al., 2021) in the construction of a pre-processing classifier, focusing on selecting relation-expressing entity pairs. Relation Classification(RC) takes the results from the previous stage to create a prompt that guides fine-tuned LLaMA2 to accomplish multi-classification tasks.

Based on this finding, this paper introduces a novel method LMRC (shown in Figure 1) to narrow the performance gap between the LLM-based DocRE methods and traditional methods. Specifically, LMRC conceptualizes DocRE as a workflow comprising two key stages: **Relation Candidate Proposal**(RCP) and **Relation Classification**(RC). The former constructs a pre-processing classifier that explicitly leverages the attention mechanism to filter out *NA* entity pairs. The latter uses

LLMs to accomplish multi-classification on the reduced candidate space. Experimental results on the DocRED and Re-DocRED benchmarks showed that the proposed LMRC gains significant improvement over other LLM-based DocRE methods, suggesting its viability as a strategy for future DocRE.

2 Preliminary Experiment

For analysing why current LLMs under-performed in DocRE, we fine-tune LLaMA2-13B-Chat for DocRE and report the results realized by this approach as well as the finding on DocRED and Re-DocRED.

Fine-tuning LLaMA2 To construct prompts for this task, we use the instruction: *Your task is to determine whether there are relations between the entity pairs based on the information in the text. If there exist relations, select relations for the entity pairs from the relation set; if there is no relation, return None.*, followed by an input consisting of a predefined relation set, the text corresponding to the document, and the entity pairs that need to be classified. To prevent ambiguities and reduce token usage, we use *None* to represent *NA* and require the model to label entity pairs with no relation as *None*. The complete prompt format is provided in Appendix E.

Each document in DocRED involves a large number of tokens, frequently surpassing the maximum token length. To address this, we conduct relation extraction for each document D via $\frac{n\times(n-1)}{k}$ inputs, where n denotes the number of entities in document D, and the variable k represents the maximum number of entity pairs that can be accommodated in each input. We integrate all entity pairs into the inputs according to the above rules to perform LoRA (Hu et al., 2022) fine-tuning and testing[1] on LLaMA2-13B-Chat.

Results Statistics of DocRED and Re-DocRED are shown in Table 1. *NA* entity pairs constitute a significant proportion in both datasets, leading to an imbalance in the candidate space. Further to the empirical observations by Lilong et al. (2024), our analysis investigates the model's outputs from a distribution perspective, supported by experiments, aiming to identify the fundamental reasons behind the observed underperformance. As demonstrated

[1]Entities in the triplets returned by LLaMA2-13B-Chat are aligned to the dataset using thefuzz, and the relations generated not in the predefined relation set are considered incorrect.

Description	DocRED Dev	DocRED Test	Re-DocRED Dev	Re-DocRED Test
Candidate Space	395,572	392,158	193,232	198,670
# *NA* Entity Pairs	384,949	-	179,870	185,043
# Relation Entity Pairs	10,623	-	13,362	13,627
# Annotated Triples	12,275	-	17,284	17,448

Table 1: Statistics on DocRED and Re-DocRED

Metrics	DocRED Dev	DocRED Test	Re-DocRED Dev	Re-DocRED Test
$Precision$	69.00	-	84.88	83.94
$Recall$	27.43	-	38.06	38.14
F_1	39.25	38.66	52.56	52.45
Ign F_1	38.62	38.09	52.29	52.15
# Extracted Triples	4,925	4,932	7,787	7,979

Table 2: Results of preliminary experiment.

in Table 2, the number of triples generated by LLaMA2-13B-Chat is far less than that annotated in the dataset. This phenomenon indicates that LLMs (e.g., LLaMA2) tend to label relation-expressing entity pairs as *NA*, resulting in lower recall and subsequently lowering the F_1 score.

3 LMRC

To prevent LLMs from prioritizing *NA* entity pairs, LMRC initially uses traditional neural networks for **Relation Candidate Proposal** to identify relation-expressing entity pairs. Then, LLMs rely on these proposals for **Relation Classification**.

3.1 Relation Candidate Proposal

In this stage, we build a simple model to conduct a binary classification task, with the outcome being entity pairs expressing relations. As prior works (Tan et al., 2022a; Ma et al., 2023) have shown that contextual information is indispensable for the relation extraction task, our model adapts localized context pooling from Zhou et al. (2021).

Entity Representation Following the entity marker technique (Zhang et al., 2017; Shi and Lin, 2019), a special token "*" is inserted at the start and end position of each entity mention. Then, tokens $T = \{t_i\}_{i=1}^{l}$ within document D are encoded by a Transformer-based(Vaswani et al., 2017) pretrained language model (PLM) to generate contextualized embeddings $\mathbf{H}$ along with their attentions $\mathbf{A}$:

$$\mathbf{H}, \mathbf{A} = PLM(T), \quad (1)$$

where $\mathbf{H} \in \mathbb{R}^{l\times d}$, $\mathbf{A} \in \mathbb{R}^{H\times l\times l}$, d is the hidden dimension of the PLM and H is the number of attention heads. We take the embedding of "*" at the start of mentions as their embeddings. The entity embedding $h_{e_i} \in \mathbb{R}^d$ for each entity e_i with mentions $M_{e_i} = \{m_j^i\}_{j=1}^{N_{e_i}}$ is computed by logsumexp pooling (Jia et al., 2019):

$$h_{e_i} = log \sum_{j=1}^{N_{e_i}} exp(h_{m_j^i}). \quad (2)$$

Localized Context Representation For each entity e_i, we aggregate the attention output for its mentions by mean pooling $A_{e_i} = \sum_{j=1}^{N_{e_i}} (a_{m_j^i})$, where $a_{m_j^i} \in \mathbb{R}^{H \times l}$ is the attention weight at the position of mention m_j^i from the last layer. Then given an entity pair (e_s, e_o), its localized context embedding $c^{(s,o)} \in \mathbb{R}^d$ can be obtained by:

$$q^{(s,o)} = \sum_{i=1}^{H} (A_{e_s}^i \circ A_{e_o}^i), \quad (3)$$

$$c^{(s,o)} = \mathbf{H}^\top q^{(s,o)}, \quad (4)$$

where $q^{(s,o)} \in \mathbb{R}^l$ is the mean-pooled attention weight for entity pair (e_s, e_o) and $\mathbf{H}$ is the contextualized embedding in Eq.(1).

Binary Classification To predict whether entity pair (e_s, e_o) expresses relation, we first generate context-enhanced entity representations:

$$z_s^{(s,o)} = tanh(\mathbf{W}_s h_{e_s} + \mathbf{W}_c c^{(s,o)}), \quad (5)$$

where $\mathbf{W}_s, \mathbf{W}_c \in \mathbb{R}^{d \times d}$ are trainable parameters. We obtain the object representation $z_o^{(s,o)}$ in the same manner. Then, a bilinear classifier is applied on the representations to compute the probability:

$$P(NA|e_s, e_o) = \sigma(z_s^{(s,o)\top} \mathbf{W} z_o^{(s,o)} + b), \quad (6)$$

where $\mathbf{W} \in \mathbb{R}^{d \times d}$ is a trainable parameter matrix, σ is the sigmoid function, $P(NA|e_s, e_o)$ is the probability that entity pair (e_s, e_o) does not express any relation. We choose Binary Cross Entropy as our loss function.

3.2 Relation Classification

After identifying entity pairs in the RCP stage, we apply the method from Section 2, driving LLaMA2-13B-Chat to complete relation classification with supervised fine-tuning. We slightly modify the previous prompt by removing the *None* category and changing some expressions. These changes aim to sharpen the model's focus on the classification task. Additionally, the number of inputs for each document is greatly reduced ($\frac{n \times (n-1)}{k} \rightarrow 1$) due to the elimination of *None*. Detailed changes can be found in Appendix E.

4 Experiments

4.1 Experiment Settings

Dataset We conduct experiments on DocRED (Yao et al., 2019) and Re-DocRED (Tan et al., 2022b), two large-scale crowd-sourced benchmark datasets for document-level RE. In DocRED, over 40.7% of relational facts require multi-sentence extraction. Although DocRED is a widely used benchmark, the annotations of the dataset remain incomplete. Tan et al. (2022b) proposed Re-DocRED, a more reliable benchmark for DocRE that revises DocRED to mitigate the false negative issue within it.

Configuration In the RCP stage, we select RoBERTa$_{\text{large}}$ (Liu et al., 2019) as the foundational encoder. We implement early stopping based on the F_1 score obtained from the development set. In the RC stage, we fine-tune LLMs with the RC-specific prompt using LoRA. Details regarding hyperparameters are provided in Appendix A.

Evaluation In alignment with other SOTA models, we utilize the standard evaluation metrics: F_1 and Ign F_1. Ign F_1 is calculated by excluding triplets that are already present in the training set from both the development and test sets.

4.2 Main Results

We compare our LMRC with pretrained BERT-based and LLM-based methods on both datasets. BERT-based methods, known for achieving state-of-the-art (SOTA) performance, utilize BERT family pretrained models as encoders. Recently introduced LLM-based methods employ fine tuning, in-context learning, or retrieval augmented generation (RAG, Lewis et al. (2020)) to enhance the performance of LLMs on relation extraction. The experimental results are presented in Table 3.

As shown in Table 3a, the performance of directly fine-tuned LLaMA2 and other LLM-based methods exhibits inefficient processing and suboptimal performance, highlighting the challenges in utilizing LLMs for DocRE. Our results also corroborate the findings of Lilong et al. (2024). However, after task division, our LMRC achieves substantial enhancement, significantly increasing F_1 on LLaMA2 at both 7B and 13B scales. Table 3b compares the performance of LMRC against existing methods on the Re-DocRED test set. We observe that LMRC outperforms other LLM-based methods. Moreover, LMRC narrows the gap with the state-of-the-art

Method	Dev		Test	
	Ign F_1	F_1	Ign F_1	F_1
BERT-based				
HIN-BERT$_{base}$(Tang et al., 2020)	54.29	56.31	53.70	55.60
CorefBERT$_{base}$(Ye et al., 2020)	55.32	57.51	54.54	56.96
CorefRoBERTa$_{large}$(Ye et al., 2020)	57.35	59.43	57.90	60.25
SSAN-RoBERTa$_{large}$(Xu et al., 2021)	60.25	62.08	59.47	61.42
KD-RoBERTa$_{large}$(Tan et al., 2022a)	65.27	67.12	65.24	67.28
DREEAM-RoBERTa$_{large}$(Ma et al., 2023)	65.52	67.41	65.47	67.53
LLM-based				
CoR(Sun et al., 2024)	-	38.4 ± 10.6	-	38.5 ± 9.1
GenRDK(Sun et al., 2024)	-	42.5 ± 10.6	-	41.5 ± 8.7
Our Methods				
LoRA FT LLaMA2-7B-Chat†	33.95	34.32	33.99	34.34
LoRA FT LLaMA2-13B-Chat†	38.62	39.25	38.09	38.66
LMRC-LLaMA2-7B-Chat	52.40	54.10	52.81	54.73
LMRC-LLaMA2-13B-Chat	58.16	59.97	58.49	60.52

(a) Results on the development and test set of DocRED.

Method	Ign F_1	F_1
BERT-based		
KD(Tan et al., 2022a)	77.60	78.28
DREEAM(Ma et al., 2023)	79.66	80.73
LLM-based		
CoR(Sun et al., 2024)	-	37.1 ± 9.2
GenRDK(Sun et al., 2024)	-	41.3 ± 8.9
AutoRE(Lilong et al., 2024)	-	51.91
Our Methods		
LoRA FT LLaMA2-13B-Chat†	52.15	52.45
LMRC-LLaMA2-13B-Chat	74.08	74.63

(b) Results on the test set of Re-DocRED

Table 3: Evaluation results on the DocRED and Re-DocRED datasets. The scores of prior methods are borrowed from corresponding papers. Results marked with † are our baselines.

Method	Intra	Inter
BERT-RE$^{\dagger}_{base}$	61.61	47.15
RoBERTa-RE$^{*}_{base}$	65.65	50.09
LSR-BERT$^{\dagger}_{base}$	65.26	52.05
GAIN-BERT$^{*}_{base}$	67.10	53.90
LoRA FT LLaMA2-13B-Chat	45.43	31.67
LMRC	65.88	52.66

Table 4: Intra- and Inter-F_1 on the development set of DocRED. † denotes results from Nan et al. (2020), and * denotes results from Zeng et al. (2020).

Settings	F_1 of RCP	Ign F_1	F_1
RCP stage			
LMRC	64.64	58.16	59.97
w/o pre-classifer *w* LLM	31.30	23.22	24.59
RC stage			
relation classification	-	86.09	86.75

Table 5: Ablation studies evaluated on DocRED dev set.

method, DREEAM, positioning it as a promising paradigm for future DocRE. Additionally, we report Intra-F_1/Inter-F_1, which consider either intra- or inter-sentence relations respectively. LSR (Nan et al., 2020) and GAIN (Zeng et al., 2020) are both graph-based methods. As Table 4 illustrates, LMRC not only surpasses selected baselines in Intra- and Inter-F_1 but also remains competitive when compared with graph-based models like GAIN-BERT$_{base}$.

4.3 Ablation studies

We explore the effectiveness of RCP and RC stages on DocRED dev set. In the RCP stage, we fine-tune LLaMA2-13B-Chat to replace the pre-classifier for binary classification. In the RC stage, we input entity pairs annotated in the ground truth, mask their relation tags, and then employ task-specific fine-tuned LLaMA2-13B-Chat to classify them into the predefined relation set.

As shown in Table 5, the F1 score drops significantly when substituting our pre-classifier, indicating that the fine-tuned LLM still struggles to distinguish the presence of relations. This may be attributed to DocRE involving multiple relations and triplet facts distributed across a document, posing distinct challenges for LLMs. This emphasizes the significant role of a pre-processing classifier. Furthermore, the ablation result of the RC stage highlights that the RC-specifically fine-tuned LLM excels in relation classification, laying effective groundwork for future advancements.

5 Conclusion

In this work, we investigate the underlying reasons for LLM's limited effectiveness in document-level relation extraction and introduce a new approach, the LLM with Relation Classifier (LMRC), for DocRE. Our method comprises two main stages: relation candidate proposal and relation classification. Through experiments conducted on DocRED and Re-DocRED, we demonstrate the effectiveness of our proposed LMRC approach. The results further reveal that LMRC holds strong competitive advantages over other existing LLM-based methods. Our innovative model establishes a new standard, indicating its potential as a viable framework for future DocRE research.

6 Limitations

Despite our efforts, this study has some limitations: **LLMs**: We only fully tested our method with LLaMA2 due to time constraints. Given budget limitations, we randomly sampled 100 documents from the DocRED dev set to test the performance of GPT-4-turbo, with results presented in Appendix C. In the future, we plan to evaluate GPT4's performance on the entire dataset and explore the applicability of our method on other freely accessible LLMs, such as Mistral and Vicuna, to understand its effectiveness across different LLMs. **Other Methods**: The imbalance of the dataset may affect the accuracy of directly fine-tuned LLMs. In future work, we aim to address this issue by employing imbalanced training techniques, such as down-sampling. **Other Relation Extraction Tasks**: Our model could be suitable for various levels of relation extraction, including sentence-level and document-level tasks. Our next experiments will investigate the performance of LMRC on these tasks to demonstrate its generalization ability.

References

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 *International Conference on Learning Representations*.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3693–3704.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Bo Li, Wei Ye, Zhonghao Sheng, Rui Xie, Xiangyu Xi, and Shikun Zhang. 2020. Graph enhanced dual attention network for document-level relation extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1551–1560, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

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*.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. DREEAM: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1971–1983, Dubrovnik, Croatia. Association for Computational Linguistics.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1546–1557, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Qi Sun, Kun Huang, Xiaocui Yang, Rong Tong, Kun Zhang, and Soujanya Poria. 2024. Consistency guided knowledge retrieval and denoising in llms for zero-shot document-level relation triplet extraction. *arXiv preprint arXiv:2401.13598*.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland. Association for Computational Linguistics.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting DocRED - addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. 2020. Hin: Hierarchical inference network for document-level relation extraction. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia*

Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I, page 197–209.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14149–14157.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186, Online. Association for Computational Linguistics.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640, Online. 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*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14612–14620.

A Hyperparameter settings

The hyperparameters for the RCP stage, as well as the settings for LLaMA2's LoRA (Hu et al., 2022) fine-tuning during the preliminary experiment and the RC stage, can be found in Tables 6 and 7, respectively. In the RCP stage, we adopt AdamW as the optimizer (Loshchilov and Hutter, 2019) and apply a linear warmup for the learning rate at the first 6% steps. We use development set to manually tune the optimal hyperparameters for the RCP stage, based on the F1 score. The value of hyperparameters we finally adopted are in bold.

Hyperparam	DocRED	Re-DocRED
batch size	4	4
# Epoch	20, **30**, 40	**30**, 40
lr for encoder	{5, **3**, 1}e-5	{**3**, 1}e-5
lr for classifier	1e-4	1e-4
max gradient norm	1.0	1.0

Table 6: Settings for the RCP stage.

Hyperparam	Pre		RC stage	
	DocRED	Re-DocRED	DocRED	Re-DocRED
batch size	4	4	4	4
# Epoch	2	2	8	8
learning rate	1e-4	1e-4	1e-4	1e-4
warmup steps	200	200	100	100
lora r	8	8	8	8
lora alpha	16	16	16	16

Table 7: Settings for LoRA fine-tuning. (Pre stands for preliminary experiment)

B Out-of-Domain Relations Studying

In the aforementioned evaluation, we simply categorize all relations generated by LLaMA2 that do not fall within the predefined relation set as erroneous outcomes. However, previous work (Wadhwa et al., 2023) has pointed out that evaluating LLM-based models cannot entirely rely on exact matches to targets. For example, although "works at" from the result is semantically similar to "work for" in the target, strict evaluation criteria would count it as a misclassification.

To delve into this phenomenon thoroughly, we revisit the out-of-domain relations that generated by LLaMA2-13B-Chat. We leverage SBERT[2] (Reimers and Gurevych, 2019) to align out-of-domain relations into the predefined relation set R. This process involves the computation of cosine similarity. For each out-of-domain relation r_i, we

[2] https://www.sbert.net

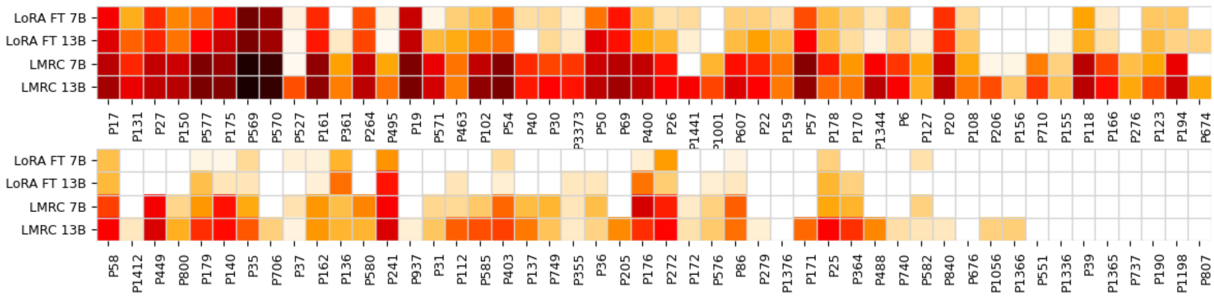
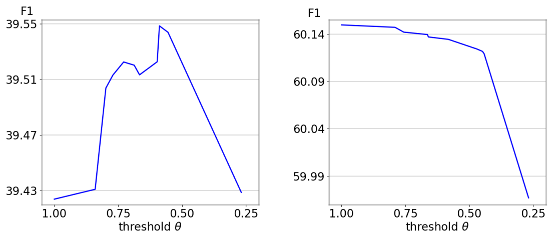


Figure 2: F_1 scores per relation type in the DocRED development set results (darker = better). White color means that no correct predictions were made for this relation. The relations are arranged in descending order by the number of triples.



(a) LoRA FT

(b) LMRC

Figure 3: The impact of threshold θ for cosine similarity on the F1 score. Both methods are conducted on the DocRED dev set.

choose the relation in R with the highest similarity $s^i_{\max}$ as the final result. Intuitively, some out-of-domain relations may be meaningless, and aligning all of them to R is not appropriate. Therefore, we introduce a heuristic threshold θ, where alignment is performed when $s^i_{\max} \geq \theta$; otherwise, the triplets containing r_i are discarded. After alignment, we recalculate the F_1 of our methods on the DocRED dev set.

Figure 3 shows that LLaMA2 directly fine-tuned on DocRED results in its peak F_1 score when threshold θ is set around 0.55. While, even with semantic alignment, the out-of-domain relations generated by LLaMA2 fine-tuned within the LMRC method remain incorrect. We analyze these relations in depth and conclude that they mainly suffer from the following two problems:

(1) The model outputs "-" appearing in the entity pair input format as a relation.

(2) Some out-of-domain relations can be mapped onto similar relations within the predefined set R. But, the classification is incorrect.

We notice that the two methods only generated 74 and 72 out-of-domain relations, respectively. This explains why threshold θ has little impact on F_1 score.

C Performance of GPT-4-turbo

To ensure a comprehensive assessment, we employed a 3-shot learning format with GPT-4-turbo, utilizing examples meticulously curated from the human-annotated DocRED dataset. These examples were carefully chosen to include both relation-expressing and *NA* entity pairs, thereby mirroring the complexity and variability inherent in real-world documents.

Given the constraints of budget, our initial analysis was conducted on a randomly selected sample of 100 documents from the development set. The preliminary results are shown in Table 8.

Metrics	Value
Precision	7.11
Recall	34.49
F_1	11.79
Ign F_1	10.85
Intra F_1	15.85
Inter F_1	8.17

Table 8: Performance of GPT-4-turbo on sampled documents from the DocRED dev set.

These outcomes underscore the inherent challenges of DocRE, even when utilizing advanced open-API LLMs such as GPT-4-turbo.

D Overall Performance

Tables 9 and 10 provide detailed results of relation extraction for all relations by our methods on the DocRED dev set. Figure 2 provides a more straightforward visual representation of the enhancement effects.

E Prompts

Table 11 shows the prompt designed for document-level relation extraction task, and Table 12 shows the prompt for relation classification task. The primary distinction between the two lies in "Instruction" and "Entity Pairs". The former encompasses all constructible entity pairs, while the latter's entity pairs are obtained by the RCP stage.

Relation ID	Relation Name	LMRC 7B	LMRC 13B	LoRA FT 7B	LoRA FT 13B
P17	country	51.87	56.91	64.52	69.89
P131	located in the administrative territorial entity	22.51	35.27	44.24	54.68
P27	country of citizenship	43.38	44.34	62.68	63.46
P150	contains administrative territorial entity	29.72	32.73	60.93	65.51
P577	publication date	34.18	50.16	77.03	77.76
P175	performer	47.46	61.28	71.95	75.82
P569	date of birth	79.33	78.01	94.39	95.24
P570	date of death	72.38	70.36	88.80	90.25
P527	has part	3.31	2.23	1.87	38.03
P161	cast member	44.10	46.58	73.46	74.21
P361	part of	0.00	5.88	25.45	31.15
P264	record label	38.49	39.77	62.08	64.52
P495	country of origin	3.54	1.81	24.45	33.54
P19	place of birth	61.86	62.50	76.98	78.05
P571	inception	2.53	19.32	54.43	59.68
P463	member of	13.53	23.13	32.80	31.09
P102	member of political party	18.02	29.75	62.92	73.79
P54	member of sports team	32.51	33.33	75.88	77.88
P40	child	9.30	0.00	43.48	46.32
P30	continent	10.29	10.22	39.48	49.80
P3373	sibling	4.32	5.71	42.06	46.96
P50	author	32.48	55.90	60.83	64.08
P69	educated at	47.14	47.95	65.14	70.09
P400	platform	20.00	24.39	62.50	63.69
P26	spouse	8.51	20.18	47.50	50.67
P1441	present in work	9.84	5.00	0.00	51.81
P1001	applies to jurisdiction	0.00	0.00	20.69	39.74
P607	conflict	13.11	18.71	47.83	57.26
P22	father	12.50	25.58	44.76	50.37
P159	headquarters location	8.70	18.18	32.21	31.88
P57	director	38.60	50.77	75.65	73.10
P178	developer	18.37	18.37	46.07	34.15
P170	creator	13.64	9.52	27.27	31.88
P1344	participant of	6.78	3.45	51.13	64.81
P6	head of government	15.63	11.32	42.02	50.94
P127	owned by	0.00	4.88	24.30	22.22
P20	place of death	42.67	43.24	61.54	64.08
P108	employer	10.34	15.15	24.00	32.38
P206	located in or next to body of water	0.00	0.00	2.35	38.41
P156	followed by	0.00	0.00	2.82	14.63
P710	participant	0.00	0.00	30.95	41.67
P155	follows	0.00	2.86	12.50	17.39
P118	league	25.29	22.22	64.52	65.00
P166	award received	5.71	8.45	40.38	48.21
P276	location	0.00	0.00	18.18	24.00
P123	publisher	15.58	18.82	24.53	39.39
P194	legislative body	15.38	12.31	54.05	61.11
P674	characters	0.00	12.50	0.00	23.84

Table 9: F_1 scores on each relation by our methods. The relations are arranged in descending order by the number of triples.

Relation ID	Relation Name	LMRC 7B	LMRC 13B	LoRA FT 7B	LoRA FT 13B
P58	screenwriter	17.78	19.23	40.45	50.63
P1412	languages spoken, written or signed	0.00	0.00	0.00	6.35
P449	original network	0.00	0.00	52.17	57.83
P800	notable work	0.00	0.00	11.49	21.98
P179	series	3.08	17.72	27.27	43.28
P140	religion	2.38	7.06	47.30	48.65
P35	head of state	10.53	7.27	23.68	36.96
P706	located on terrain feature	0.00	0.00	0.00	12.99
P37	official language	4.00	0.00	7.89	3.17
P162	producer	3.85	3.92	27.03	27.16
P136	genre	21.05	33.33	18.18	20.51
P580	start time	0.00	0.00	29.89	20.34
P241	military branch	27.45	47.22	50.88	57.43
P937	work location	0.00	0.00	0.00	4.55
P31	instance of	0.00	0.00	11.32	15.87
P112	founded by	0.00	7.14	10.53	35.56
P585	point in time	0.00	0.00	15.09	37.89
P403	mouth of the watercourse	9.52	5.00	34.67	40.58
P137	operator	0.00	0.00	18.52	31.75
P749	parent organization	0.00	0.00	20.00	16.67
P355	subsidiary	0.00	6.25	7.14	5.66
P36	capital	0.00	6.67	18.18	12.12
P205	basin country	0.00	0.00	0.00	29.27
P176	manufacturer	4.35	32.73	59.34	42.98
P272	production company	26.42	13.95	45.33	49.32
P172	ethnic group	0.00	0.00	5.71	8.51
P576	dissolved, abolished or demolished	0.00	5.00	12.50	14.81
P86	composer	3.39	6.56	35.85	33.33
P279	subclass of	0.00	0.00	0.00	4.35
P1376	capital of	0.00	0.00	0.00	0.00
P171	parent taxon	0.00	0.00	0.00	34.29
P25	mother	12.50	20.00	24.00	50.00
P364	original language of work	0.00	12.50	20.90	42.62
P488	chairperson	0.00	0.00	0.00	29.41
P740	location of formation	0.00	0.00	0.00	8.70
P582	end time	8.33	0.00	12.50	7.14
P840	narrative location	0.00	0.00	0.00	6.45
P676	lyrics by	0.00	0.00	0.00	0.00
P1056	product or material produced	0.00	0.00	0.00	12.50
P1366	replaced by	0.00	0.00	0.00	13.33
P551	residence	0.00	0.00	0.00	0.00
P1336	territory claimed by	0.00	0.00	0.00	0.00
P39	position held	0.00	0.00	0.00	0.00
P1365	replaces	0.00	0.00	0.00	0.00
P737	influenced by	0.00	0.00	0.00	0.00
P190	sister city	0.00	0.00	0.00	0.00
P1198	unemployment rate	0.00	0.00	0.00	0.00
P807	separated from	0.00	0.00	0.00	0.00

Table 10: F_1 scores on each relation by our methods. The relations are arranged in descending order by the number of triples. (Continued)

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:
Your task is to determine whether there are relations between the entity pairs based on the information in the text. If there exists relations, select relations for the entity pairs from the relation set; if there is no relation, return None.
The format of the input entity pair is ’(head entity| -| tail entity)’.
Your output format is ’(head entity| relation/None| tail entity)’.

Relation set:
{predefined relation set}

Text:
{text}

{number of entity pairs} Entity pairs:
{entity pairs}

Response:

Table 11: Prompt for document-level relation extraction

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:
This is a relation classification task. we will provide entity pairs that require relation classification. Your task is to select relations for each entity pair from the given relation set based on the information in the text. There may be multiple relations between an entity pair.
The format of the input entity pair is ’(head entity| -| tail entity)’.
Your output format is ‘(head entity| relation| tail entity)’.

Relation set:
{predefined relation set}

Text:
{text}

{number of entity pairs} Entity pairs:
{entity pairs}

Response:

Table 12: Prompt for relation classification