AutoRE: Document-Level Relation Extraction
with Large Language Models

Lilong Xue[*]
Tsinghua University
xll21@mails.tsinghua.edu.cn

Dan Zhang[*]
Tsinghua University
zd21@mails.tsinghua.edu.cn

Yuxiao Dong
Tsinghua University
yuxiaod@tsinghua.edu.cn

Jie Tang[†]
Tsinghua University
jietang@tsinghua.edu.cn

Abstract

Large Language Models (LLMs) have demonstrated exceptional abilities in comprehending and generating text, motivating numerous researchers to utilize them for Information Extraction (IE) purposes, including Relation Extraction (RE). Nonetheless, most existing methods are predominantly designed for Sentence-level Relation Extraction (SentRE) tasks, which typically encompass a restricted set of relations and triplet facts within a single sentence. Furthermore, certain approaches resort to treating relations as candidate choices integrated into prompt templates, leading to inefficient processing and suboptimal performance when tackling Document-Level Relation Extraction (DocRE) tasks, which entail handling multiple relations and triplet facts distributed across a given document, posing distinct challenges. To overcome these limitations, we introduce AutoRE, an end-to-end DocRE model that adopts a novel RE extraction paradigm named RHF (Relation-Head-Facts). Unlike existing approaches, AutoRE does not rely on the assumption of known relation options, making it more reflective of real-world scenarios. Additionally, we have developed an easily extensible RE framework using a Parameters Efficient Fine Tuning (PEFT) algorithm (QLoRA). Our experiments on the RE-DocRED dataset showcase AutoRE's best performance, achieving state-of-the-art results, surpassing TAG by 10.03% and 9.03% respectively on the dev and test set. The code is available[1] and the demonstration video is provided[2].

1 Introduction

The rise of LLMs, such as GPT-4 (Achiam et al., 2023) and Llama2 (Touvron et al., 2023), has significantly propelled the progress of natural language processing due to their strong capabilities
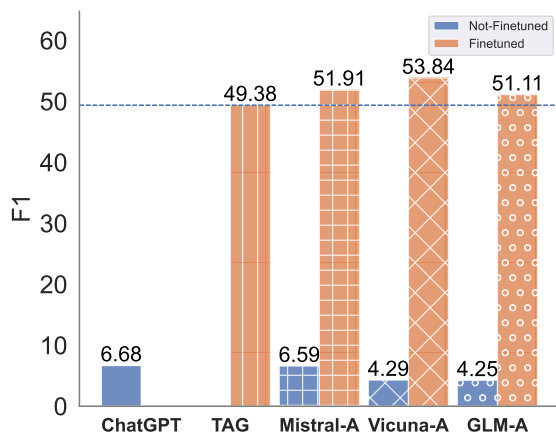

Figure 1: The result on the test set of Re-DocRED. AutoRE (-A) achieves SOTA for different LLMs.

in text understanding, generation, scientific reasoning (Zhang et al., 2024a,b), social bot detection (Zhou et al., 2024), and generalization (Zhao et al., 2023). There has been an increasing interest in using LLMs to generate structured information for IE tasks (Xu et al., 2023; Wadhwa et al., 2023) and making impressive progress. Typical IE tasks using LLMs include Named Entity Recognition (NER) (Wang et al., 2023a), Relation Extraction (RE) (Zhou et al., 2023), and Event Extraction (EE) (Xu et al., 2023). Despite the outstanding result, the performance of current LLMs in RE is still far from satisfactory.

Underperformance on DocRE Tasks. We evaluated several high-performing LLMs on the document-level RE (DocRE) task, specifically using the test set of Re-DocRED (Tan et al., 2022). These models included GPT-3.5-turbo[3] (ChatGPT), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) (Mistral-7B), Vicuna-7B-v1.5 (Chiang et al., 2023) (Vicuna-7B), and ChatGLM3-6B (Du et al., 2022). Our results indicate that, without specific fine-tuning, the performance of these language

---

[1]https://github.com/THUDM/AutoRE
[2]https://www.youtube.com/watch?v=IhKRsZUAxKk
[*]LX and DZ contributed equally.
[†]JT is the corresponding author.

[3]https://chat.openai.com/chat

models on DocRE tasks is suboptimal, as shown in the blue bars in Figure 1.

Inefficacy in Multi-Relations. Incorporating relations directly into the prompt template as candidates is a common strategy for LLM-based models (Wang et al., 2023b; Wei et al., 2023; Xiao et al., 2023). This method is effective for tasks that involve a relatively small number of relation types. However, the number of relation types can easily exceed 100 in real-world scenarios. Dealing with multiple relations, as seen in the Re-DocRED dataset with its 96 relation types, presents a significant challenge for most existing models. Embedding such many relations directly into the prompt template is often impractical (Wadhwa et al., 2023).

Limitations of Current RE Paradigms. The current paradigms in RE exhibit significant limitations in their effectiveness. Modern generative methods typically operate by either directly producing triplet facts from the input text in a singular step (Wang et al., 2023b), or by initially identifying a set of relations and subsequently generating triplet facts based on these relations (Wei et al., 2023). Earlier approaches prioritized the extraction of the head entity before the derivation of triplet facts (Li et al., 2019). However, these methodologies fall short of handling DocRE tasks that involve multiple relations and plenty of triplet facts. For instance, a single instance in the Re-DocRED dataset might encompass as many as 27 different relations and include up to 142 distinct triplet facts.

To address challenges identified in existing RE paradigms, we innovate a new pipeline RE paradigm, Relation-Head-Facts (RHF). We comprehensively redefined the 96 relation descriptions and crafted simplified relation extraction templates, developing an instruction-tuning dataset based on Re-DocRED. Utilizing the Mistral-7B model with Parameter Efficient Fine-Tuning (PEFT), QLoRA (Dettmers et al., 2023), our model achieved state-of-the-art (SOTA) performance on the Re-DocRED test set. Key contributions of our work include:

Various RE Paradigms. We conducted experiments across a variety of RE paradigms and revealed that a pipeline RE approach is especially potent for DocRE, particularly RHF. This paradigm prioritizes the extraction of relations, followed by the identification of subjects, thereby significantly enhancing the model's capacity to efficiently and accurately uncover triplet facts.

Efficient DocRE Model. Adopting the RHF paradigm for DocRE and refined relation descriptions, we have meticulously crafted an instruction-finetuning dataset based on Re-DocRED. This dataset was utilized to fine-tune the Mistral-7B with QLoRA, culminating in the creation of AutoRE, which achieved SOTA results across multiple pre-trained LLMs (PLMs), demonstrating the generality and effectiveness of this model architecture.

Easy Enhancement of Capabilities. We have incorporates three distinct QLoRA modules within the RHF framework, where each module is exclusively responsible for a specific task: one for relation extraction, another for head entity identification, and the third for triple fact extraction, ensuring specialized handling for each aspect. This strategy effectively lays the groundwork for future advancements while ensuring a minimal rise in computational demands and avoiding interference between subtasks.

2 Related work

DocRE refers to the task of extracting relations between entities at the document level, we follow the definition in (Zheng et al., 2023): Given a document $\mathcal{D}$ with a set of sentences containing a set of entities $\mathcal{V} = \{e_i\}_{i=1}^{N}$. The DocRE task is to predict the relation types between an entity pair $(e_h, e_t)_{h,t \in \{1,\cdots,N\}, h \neq t}$, where h stands for the head (subject) and t stands for the tail (object).

LLMs for DocRE. Researchers have been employing LLMs to tackle RE tasks. For example, ChatIE (Wei et al., 2023) deconstructs the complex RE process into assembling the outputs from multiple rounds of Question-Answer into a final structured format. PromptRE (Gao et al., 2023) integrates LLM prompting with data programming to deal with DocRE. However, the performance of LLMs on RE tasks still lags behind SOTA models. Han et al. concluded that *ChatGPT does not adequately comprehend the subject-object relationships in RE tasks*. Similarly, Li et al. noted that in Standard-IE settings, *ChatGPT's performance is generally not as effective as BERT-based models*. Moreover, most models are tested on SentRE. To test LLMs for DocRE, we conducted tests on Chat-GPT, Mistral-7B, Vicuna-7B, and ChatGLM3-6B and revealed that the current performance is far from satisfactory, as illustrated in Figure 1. This aligns with findings reported by (Li et al., 2023b),

indicating that current models still fall significantly short in performance on DocRE.

RE Prompt Template. These models fine-tuned on LLMs for RE operate on a prompt-based or instruction-driven mechanism (Beurer-Kellner et al., 2023), engaging in a question-and-answer format to execute RE tasks. ChatIE (Wei et al., 2023), InstructUIE (Wang et al., 2023b), and YAYI (Xiao et al., 2023) while demonstrating formidable capabilities in IE, exhibit considerable limitations in their RE prompt templates. A common method in their RE process involves embedding a list of relations into the model's prompt template as alternatives. However, this approach becomes impractical when dealing with DocRE, such as the 96 relations in the Re-DocRED. This limitation is acknowledged by Wadhwa et al., who concludes that *"for datasets with long texts or a large number of targets, it is not possible to fit detailed instructions in the prompt"*.

RE Paradigms. Within the context of LLMs, RE paradigms are primarily categorized into two types: Pipeline and Joint. The Pipeline approach involves first identifying relations and then extracting triplet facts, or initially extracting a head entity followed by its corresponding relation and tail entity. This approach deviates from the traditional methodology of first extracting entities and then determining their interrelations (Chen and Guo, 2022; Jiang et al., 2020). The main drawbacks is that applying the conventional Pipeline approach to LLMs can be extremely time-consuming, particularly when many entities lack interrelations. On the other hand, the Joint paradigm, which inputs a text and directly outputs all triplet facts as seen in (Zhang et al., 2023), aligns more closely with traditional practices. However, as illustrated in Table 1, these paradigms encounter significant challenges when applied to DocRE, particularly due to the complexity of handling samples that may contain multiple relations and a multitude of triplet facts.

In summary, current LLMs still exhibit significant gaps in performance for DocRE, indicating a need for further fine-tuning. Additionally, the existing RE templates, which treat relations as candidates, struggle to handle scenarios involving multiple relations. Coupled with the underwhelming effectiveness of current RE paradigms, there is a need for a paradigm shift.

Paradigm	TP	FP	R	P	F1
D-F	735	3824	4.21	16.12	6.68
D-RS-F	867	4811	4.97	15.27	7.50
D-R-F	1674	93741	9.59	1.75	2.97
D-R-H-F	3201	333226	18.35	0.95	1.81

Table 1: The result of four RE paradigms with ChatGPT. Here, TP denotes True Positive, FP is False Positive, R for Recall, P means Precision, and F1 references Micro F1. All paradigms perform poorly.

3 Methodology

3.1 RE Paradigms

We summarized the existing paradigms of RE and designed a unique extraction paradigm, different RE paradigms are illustrated in Figure 2.

Document-facts (D-F). Fed with a document, the model directly outputs all triplets facts. This method is brute-force and requires the shortest inference time. It directly inputs relation types as candidates into the prompt and then let the model generate all triplet facts in one step as InstructIE (Wang et al., 2023b) did.

Document-relations-facts (D-RS-F). In this paradigm, the model extracts the relations present within the document and embeds all the predicted relations into the prompt to obtain triplet facts.

Document-relation-facts (D-R-F). In this framework, the model identifies relations within a given sentence and systematically traverses these relations to acquire triplet facts that correspond to each identified relation, which is similar to the approach taken by (Wei et al., 2023).

Document-relation-head-facts (D-R-H-F). In our newly designed paradigm, the model specifically focuses on each relation to identify an appropriate set of entities that will function as the `head` in the triplet facts. Subsequently, the relevant triplets facts corresponding to these relations are extracted.

We test these paradigms with ChatGPT and the results are displayed in Table 1. We provide testing prompts for the Re-DocRED dataset under different paradigms using ChatGPT in Table 5. For brevity, we only provide two representative relation extraction prompt words. The rest are similar to these. We arrived at the following conclusion:

- LLMs still perform poorly in DocRE tasks involving the extraction of multiple relations

While the City Sleeps , We Rule the Streets is the debut studio album by Cobra Starship . It was released on October 10 , 2006 in the US , and on October 17 , 2006 in Canada . A rough clip of " Send My Love to the Dancefloor , I 'll See You In Hell (Hey Mister DJ) " , a finished version of " Snakes on a Plane (Bring It) " , and " The Church of Hot Addiction " were uploaded onto Cobra Starship 's PureVolume site . " The Church of Hot Addiction " was also used as the theme song for the WWE 's Great American Bash 2007 . It has sold more than 69,000 copies to date .

Document
D-F
[Cobra Starship,country,US]
[Hey Mister DJ,performer,Cobra Starship]
[The Church of Hot AddictionStarship,follows,Bring It]

Document
D-RS-F
country / performer / follows
[Cobra Starship,country,US]
[Hey Mister DJ,performer,Cobra Starship]
[The Church of Hot AddictionStarship,follows,Bring It]

Document
D-R-F
performer / country / follows
[Cobra Starship,country,US]
[WWE,country,US]
[Great American Bash,country,US]

Document
D-R-H-F
performer / country / follows
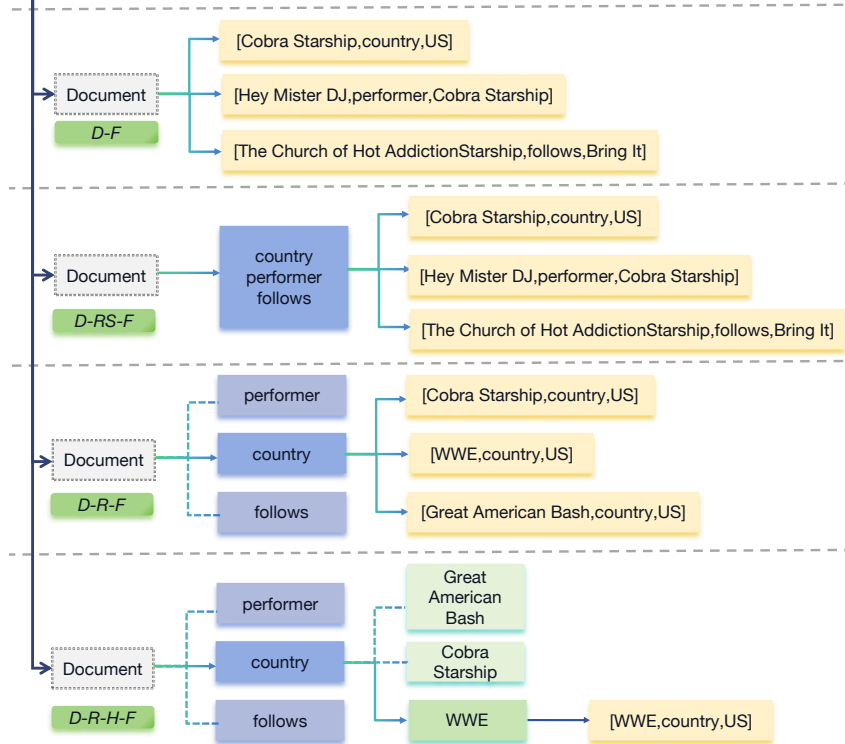Great American Bash / Cobra Starship / WWE
[WWE,country,US]

Figure 2: Processing steps of different RE paradigms.

and triplet facts, achieving only single-digit scores. As of now, fine-tuning the model is still necessary.

- By extending the thought chain to derive final triplet facts, we can obtain more accurate triplet facts, though this approach does introduce a higher number of erroneous facts.

- Harnessing the last paradigm, which we name RHF, the model can find triplets facts more accurately in a step-by-step mode with finer-grained tasks, thereby enhancing recall rates.

3.2 Dataset Processing

We used the Re-DocRED dataset for fine-tuning, refining it by removing duplicates and ensuring factual accuracy. This involved adjusting reciprocal relations like "follows" and "followed by" to accurately represent inversion, enhancing the dataset's robustness and precision.

In earlier experiments with ChatGPT, we discovered that providing the model with descriptions

of relations enhances its capability to extract factual information. Nevertheless, incorporating Wikidata relation descriptions[4] led to diminished performance, likely due to their occasional lack of clarity and precision, as exemplified by:

"located in the administrative territorial entity": "The item is located on the territory of the following administrative entity. Use P276 for specifying locations that are non-administrative places and for items about events. Use P1382 if the item falls only partially into the administrative entity."

Addressing this, we systematically rewrote all 96 relation descriptions, markedly improving model performance in Table 2. An example of our revised description is as follows. Details of all relation descriptions are presented in Table 6.

"located in the administrative territorial entity": "This relation indicates that a subject (e.g., a place, event, or item) is situated within an administrative region, the object. Example: (Harvard University, located in the administrative territorial entity,

---

[4]https://www.wikidata.org/

Paradigm	TP	FP	R	P	F1
D-R-F-no$_{desc}$	1952	27584	11.19	6.61	8.31
D-R-H-F-no$_{desc}$	4005	123631	22.95	3.14	5.52
D-R-F-wiki$_{desc}$	1296	21482	7.43	5.69	6.44 ↓
D-R-H-F-wiki$_{desc}$	3283	137462	18.82	2.33	4.15↓
D-R-F-new$_{desc}$	3508	29002	20.11	10.79	**14.04** ↑
D-R-H-F-new$_{desc}$	4200	118243	24.07	3.43	**6.00** ↑

Table 2: The result of two RE paradigms, we skip the step of extracting relations and instead use the correct relation as prior knowledge.

Module	TP	FP	Recall	Precision	F1
QLoRA-relation-dev	3190	657	63.81	82.92	72.12
QLoRA-head-dev	11269	1910	65.38	85.51	74.10
QLoRA-fact-dev	14439	2628	83.77	84.60	84.18
QLoRA-relation-test	3073	686	64.44	81.75	72.06
QLoRA-head-test	12820	2771	73.48	82.23	77.60
QLoRA-fact-test	14439	2628	82.75	84.60	83.66
AutoRE-dev	7588	3805	44.02	66.60	53.01
AutoRE-test	7445	3794	42.67	66.24	51.91

Table 3: The results of AutoRE on the Re-DocRED dev and test sets for the three subtasks of RHF.

Model	dev F1	test F1
TAG	49.34	49.38
AutoRE-ChatGLM3-6B	**49.86**	**51.11**
AutoRE-Mistral-7B	**53.01**	**51.91**
AutoRE-Vicuna-7B	**54.29**	**53.84**

Table 4: The results of AutoRE for different PLMs. Compared with TAG, all AutoRE models achieve the best performance.

Cambridge, Massachusetts)."

Finally, in line with the RHF paradigm, we crafted instruction fine-tuning templates, breaking down the RE process of each sample into three distinct steps. The specific details of these templates can be found in Table 7, we provide a display of how we constructed our training data using simple prompt templates, the extraction of relations, the extraction of the head entity, and finally the triplet extraction.

3.3 QLoRA Tuning

Mistral-7B was selected as the foundation for fine-tuning because it demonstrated the best performance among the several open-source models tested when evaluating LLMs on the Re-DocRED task. To facilitate efficient training, we opted for PEFT's QLoRA. The key advantage of QLoRA is its ability to combine the benefits of quantization and Low-Rank Adaptation (Hu et al., 2021), resulting in efficient fine-tuning. Specifically, quantization reduces data complexity, allowing for more efficient storage and processing, which is particularly valuable for deploying large models on resource-constrained devices.

We leveraged three distinct QLoRA modules, each tailored to a specific stage of the RHF steps. This implementation was critical in enhancing RE efficiency. With the data volume varying across the intertwined tasks, a one-size-fits-all approach

could have compromised performance. However, the modular structure of QLoRA facilitated smooth integration with the underlying base model. As a result, we instituted three distinct QLoRA modules, each meticulously fine-tuned to its specific dataset. This meticulous approach resulted in the creation of the AutoRE, which amalgamates these modules for amplified DocRE performance.

4 Experiment

4.1 Experimental Setup

Test set. In our evaluation, we utilized the refined Re-DocRED test set consisting of 499 articles and 17,448 triplet facts, and a validation set encompassing 498 articles with 17,236 triplets, ensuring a comprehensive and precise assessment.

Evaluation Metric. We adopted the strict Micro F1 criterion, recognizing a prediction as correct only if it precisely captures the entire relation, along with both the head and tail entities. It's important to highlight that within the Re-DocRED dataset, a triplet fact may encompass multiple aliases (mentions) for both the head and tail entities. Consequently, our evaluation protocol deems a prediction accurate if it correctly identifies any valid triplet pair. If a prediction aligns with any alias pair of the head and tail entity, it's counted as correct, but alternate accurate aliases aren't tallied in the correct statistics. Conversely, all incorrect predictions are flagged as false positives. This method ensures a stringent and statistically valid evaluation, lending robust credibility to the final results.

4.2 Overall Result

After training three distinct QLoRA modules, we test the performance on the Re-DocRED and then combine three QLoRAs to get the final performance on the dev set and test set, the result is shown in Table 3. When compared with TAG (Zhang et al., 2023) as a baseline which firstly reported the
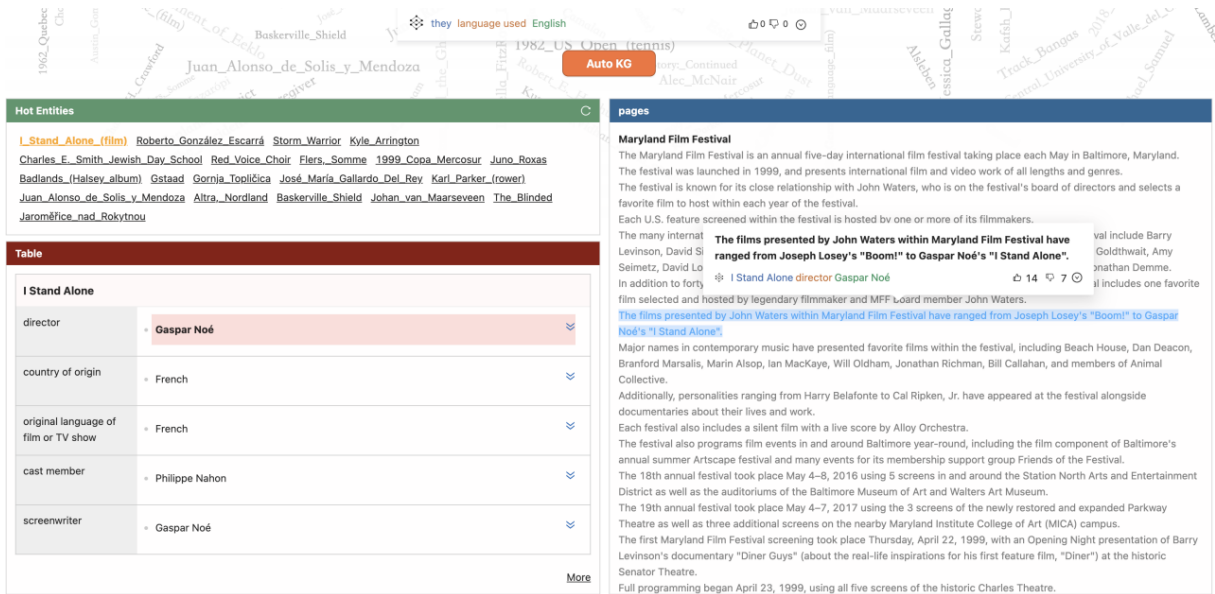
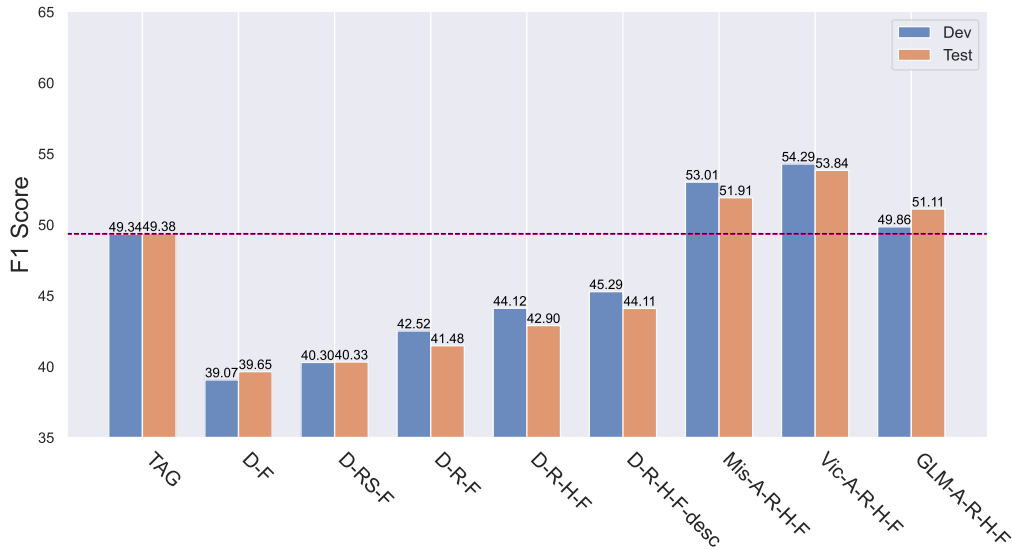Figure 3: The homepage of online AutoRE.

Figure 4: Performance of different paradigms and AutoRE (-A) for different PLMs.

end-to-end RE on Re-DocRED, our method has achieved SOTA results, as shown in Table 4. In both the dev set and test set, the performance improvement of AutoRE finetuned with Mistral-7B over TAG is approximately 7.44% on the dev set, and about 5.12% on the test set demonstrating the effectiveness of our approach. Furthermore, by decomposing the task into three subtasks and training with three LoRA modules, we not only achieved excellent results but also naturally acquired an easily extendable trait. This allows for targeted improvement of a specific module's performance without impacting the performance of other subtasks. Additionally, it is worth noting that our work is the

first to utilize large language models for processing the Re-DocRED dataset. AutoRE can serve as a reference for subsequent research in this field.

4.3 Ablation

In the ablation study, we employed Mistral-7B to fine-tune the paradigms mentioned before, revealing that the RHF model yields the best performance when solely utilizing one QLoRA module. This finding substantiates our initial hypothesis during paradigm selection: employing a step-by-step approach enhances the extraction of triplet facts while significantly reducing erroneous triplets through fine-tuning. Building on this, we compared the im-

pact of including descriptions versus omitting them. The results confirmed that incorporating proper relation descriptions indeed benefits the model, as shown in Figure 4. Additionally, we explored the effectiveness of training the entire dataset with one QLoRA versus independently training different stages of RHF with three distinct QLoRAs. The latter approach demonstrated superior performance. We believe this is due to the data imbalance among predicting relations, predicting head entities, and predicting factual triples in the RHF paradigm, with the data volume for the three subtasks begin approximately 2.8%, 24.23%, and 72.97%, respectively. When combined, the model tends to favor the prediction of triples, while its capability to predict relations is relatively insufficient.

Additionally, we have applied this framework to Vicuna-7B and ChatGLM3-6B, and both models surpassed the current SOTA levels, demonstrating the universality of the AutoRE framework. The comparative results of these experiments are illustrated in the accompanying Figure 4. Vicuna-7B scored the highest, surpassing TAG by 10.03% and 9.03% respectively on the dev and test set, whereas ChatGLM3-6B was somewhat lower. This might be due to ChatGLM3-6B having a higher proportion of Chinese in its pre-training, while it was tested on an English task. We have deployed the system on the online platform[5] for users to access and experience, as shown in Figure 3.

5 Conclusion

In this paper, we introduce RHF, a new paradigm for RE, alongside AutoRE, an advanced DocRE model. AutoRE represents a cutting-edge approach to the DocRE task, utilizing LLMs combined with QLoRA. This innovative model establishes a new standard, achieving SOTA results on the Re-DocRED dataset. AutoRE proficiently addresses the intricate task of extracting multiple relations from document-level texts, a significant challenge that has stymied existing models. Our future goal is to create a comprehensive, unified framework for RE, fully leveraging the capabilities and promise of this paradigm.

Limitations

Insufficient Number of Relations. In real-world applications, the number of relations can reach

thousands, significantly surpassing the 96 relations present in the Re-DocRED dataset. To better adapt to these real-world scenarios, it is imperative to gather more extensive datasets and expand the range of relations.

Limitation to In-Domain Data. AutoRE is not equipped to handle unseen relations. This limitation underscores the method's inadequate generalizability, primarily due to the limited scope of data it has been trained on.

References

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

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. 2023. Prompting is programming: A query language for large language models. *Proceedings of the ACM on Programming Languages*, 7(PLDI):1946–1969.

Zheng Chen and Changyu Guo. 2022. A pattern-first pipeline approach for entity and relation extraction. *Neurocomputing*, 494:182–191.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *ArXiv*, abs/2305.14314.

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

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*.

---

[5]https://models.aminer.cn/neptune/

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *ArXiv*, abs/2305.14450.

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

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Ming Jiang, Jennifer D'Souza, Sören Auer, and J Stephen Downie. 2020. Targeting precision: A hybrid scientific relation extraction pipeline for improved scholarly knowledge organization. *Proceedings of the Association for Information Science and Technology*, 57(1):e303.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023a. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *ArXiv*, abs/2304.11633.

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

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. *ArXiv*, abs/1905.05529.

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 *Conference on Empirical Methods in Natural Language Processing*.

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

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2023. Revisiting relation extraction in the era of large language models. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2023:15566–15589.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023a. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Xiao Wang, Wei Zhou, Can Zu, Han Xia, Tianze Chen, Yuan Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, J. Yang, Siyuan Li, and Chunsai Du. 2023b. Instructuie: Multi-task instruction tuning for unified information extraction. *ArXiv*, abs/2304.08085.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with chatgpt. *ArXiv*, abs/2302.10205.

Xinglin Xiao, Yijie Wang, Nan Xu, Yuqi Wang, Hanxuan Yang, Minzheng Wang, Yin Luo, Lei Wang, Wenji Mao, and Daniel Zeng. 2023. Yayi-uie: A chat-enhanced instruction tuning framework for universal information extraction. *ArXiv*, abs/2312.15548.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2023. Large language models for generative information extraction: A survey. *arXiv preprint arXiv:2312.17617*.

Dan Zhang, Ziniu Hu, Sining Zhoubian, Zhengxiao Du, Kaiyu Yang, Zihan Wang, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024a. Sciglm: Training scientific language models with self-reflective instruction annotation and tuning. *arXiv preprint arXiv:2401.07950*.

Dan Zhang, Sining Zhoubian, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024b. Rest-mcts*: Llm self-training via process reward guided tree search. *arXiv preprint arXiv:2406.03816*.

Ruoyu Zhang, Yanzeng Li, and Lei Zou. 2023. A novel table-to-graph generation approach for document-level joint entity and relation extraction. In *Annual Meeting of the Association for Computational Linguistics*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. 2023. A survey of large language models. *ArXiv*, abs/2303.18223.

Hanwen Zheng, Sijia Wang, and Lifu Huang. 2023. A survey of document-level information extraction. *ArXiv*, abs/2309.13249.

Huixue Zhou, Mingchen Li, Yongkang Xiao, Han Yang, and Rui Zhang. 2023. Llm instruction-example adaptive prompting (leap) framework for clinical relation extraction. *medRxiv*, pages 2023–12.

Ming Zhou, Dan Zhang, Yuandong Wang, Yangli-ao Geng, Yuxiao Dong, and Jie Tang. 2024. Lgb: Language model and graph neural network-driven social bot detection. *arXiv preprint arXiv:2406.08762*.

Paradigms	Prompt
D-R-F	Given a passage: {sentences}, and relation list: {relation_list} Check the passage, and find which relations can be derived from the passage. Your output format is as following: relation1 relation2 ... one example like: country of citizenship father The relations must be in the relation list. If no relation in the sentence, you should only output: no relation Given a relation: {relation}. Provided a passage: {sentences}. Derive all the triplet facts from the passage according to the given relations. Your output format is as following: [subject,{relation}, object] [subject,{relation}, object] ... The subject and object should be an entity from the passage.
D-R-H-F	Given a passage: {sentences}, and relation list: {relation_list} Check the passage, and find which relations can be derived from the passage. Your output format is as following: relation1 relation2 ... one example like: country of citizenship father The relations must be in the relation list. If no relation in the sentence, you should only output: no relation Given the relation: {relation}. Now the passage is: {sentences}. Derive all the entities from the passage that can serve as the subject of the {relation}. Your output format is as following: entity1 entity2 ... The entities should all be from the passage. Given the relation: {relation}. Now the passage is: {sentences}. Derive all the triplet facts from the passage that takes {subject} as a subject. Your output format is as following: [{subject},{relation},object] [{subject},{relation},object] ... The object should be an entity from the passage.

Table 5: ChatGPT prompt template for RE on Re-DocRED.

Relation	Description
located in the administrative territorial entity	In the "located in the administrative territorial entity" relation, the subject, a place, event, or item, resides or takes place in the object, an administrative region. Example: (Harvard University, located in the administrative territorial entity, Cambridge, Massachusetts).
country	For the "country" relation, the subject pertains to a non-human entity, such as an organization, place, or event. The object signifies the sovereign state where the subject is based or occurs. Example: (Amazon Inc, country, United States).
country of citizenship	The "country of citizenship" relation denotes that the subject, an individual, is recognized as a citizen by the object, a country. Example: (Elon Musk, country of citizenship, United States).
contains administrative territorial entity	The relation "contains administrative territorial entity" involves a subject, an administrative territory, encompassing the object, a subdivision or part of this administrative territory. Example: (California, contains administrative territorial entity, Los Angeles).
has part	The "has part" relation reflects that the subject, an entity or whole, comprises the object, a part or component of the subject. Example: (A car, has part, engine).
date of birth	In the "date of birth" relation, the subject, a person, was born on the object, the specified date. Example: (John Doe, date of birth, January 1, 1990).
part of	In the "part of" relation, the subject, a component or section, belongs to the object, a larger whole or aggregate. Example: (Engine, part of, a car).
notable work	The "notable work" relation indicates a significant work assigned to the subject, a creator, while the object is that noted scientific, artistic, or literary work itself. Example: (Jane Austen, notable work, Pride and Prejudice).
publication date	The "publication date" relation marks when the subject, a work, was first published or released, with the object being that specific date. Example: (Pride and Prejudice, publication date, 1813).
inception	In the "inception" relation, the subject, an event, or an item (not a person), came into existence at the object, a specific date or point in time. Example: (Google, inception, September 4, 1998).
date of death	The "date of death" relation specifies when the subject, a once-living person, died. The object is the particular date of demise. Example: (Albert Einstein, date of death, April 18, 1955).

Table 6: New designed relation descriptions. We only present part of the descriptions of 96 relations. The whole relation descriptions can be found via this link: https://github.com/THUDM/AutoRE.

Submission	Instruct Tuning Template
relation_template	Given a passage: {sentences}, list any underlying relations.
entity_template	Given a relation {relation}, and its description: {description} and a passage: {sentences}, list entities that can be identified as suitable subjects for the relation.
fact_template	Given relation {relation} and relation description: {description}. Provided a passage: {sentences}, list all triple facts that take {relation} as the relation and {subject} as the subject.

Table 7: Instruction tuning template for RHF.