Fine-Tuning Large Language Models for Relation Extraction within a Retrieval-Augmented Generation Framework

Sefika Efeoglu[1] **Adrian Paschke**[1,2]
[1]Freie Universitaet Berlin, Takustraße 9, 14195 Berlin
[2]Data Analytic Center (DANA), Fraunhofer Institute FOKUS, Berlin, Germany
sefika.efeoglu@fu-berlin.de, adrian.paschke@fu-berlin.de

Abstract

Information Extraction (IE) plays a pivotal role in transforming unstructured data into structured formats, such as Knowledge Graphs. One of the main tasks within IE is Relation Extraction (RE), which identifies relations between entities in text data. This process enriches the semantic understanding of documents, enabling more precise information retrieval and query answering. Recent works leveraging pre-trained language models have demonstrated significant performance improvements in RE. In the current era of Large Language Models (LLMs), fine-tuning these LLMs can mitigate the limitations of zero-shot RE methods, particularly in overcoming the domain adaptation challenges inherent in RE. This work explores not only the effectiveness of fine-tuned LLMs but also their integration into a Retrieval-Augmented Generation (RAG)-based RE approach to address domain adaptation challenges when general-purpose LLMs serve as generators within the RAG framework. Empirical evaluations on the TACRED, TACRED-Revisited (TACREV), and Re-TACRED datasets reveal substantial performance improvements with fine-tuned LLMs, such as Llama2-7B, Mistral-7B, and Flan-T5 Large and surpass previous methods on these datasets.

1 Introduction

Information Extraction (IE) converts unstructured data into structured formats, such as Knowledge Graphs (KGs). A key IE task is Relation Extraction (RE), which identifies relationships between entities in text at sentence (See Figure 1) or document levels (Grishman, 2015). RE methods include supervised, unsupervised, and rule-based approaches (Aydar et al., 2020; Pawar et al., 2017). Supervised RE methods generally yield strong performance but require extensive labeled data. However, recent studies show that RE methods using pre-trained language models (PLMs) can surpass traditional supervised approaches (Zhou and Chen, 2022; Li et al., 2022; Wang et al., 2022). In the era of Large Language Models (LLMs), Retrieval-Augmented Generation (RAG) (Gao et al., 2023; Lewis et al., 2020) using zero-shot prompting settings, in-context learning (Pan et al., 2024), or simple vanilla prompting (Kai Zhang, 2023), have been utilized for RE without the need for additional model training.

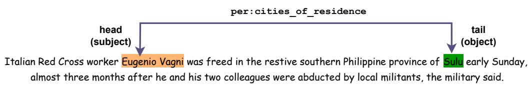

per:cities_of_residence
head
(subject)
tail
(object)
Italian Red Cross worker Eugenio Vagni was freed in the restive southern Philippine province of Sulu early Sunday, almost three months after he and his two colleagues were abducted by local militants, the military said.

Figure 1: Representation of a relation, per:cities_of_residence, between head and tail entities in a sentence from the TACRED dataset.

The RAG-based prompting approach performs well when entity relations are easily derived from sentence tokens but struggles when relation types are not introduced into LLMs (Efeoglu and Paschke, 2024). General-purpose LLMs, like Mistral (Jiang et al., 2023), Llama2 (Touvron et al., 2023), and Flan-T5 (Chung et al., 2022), also show shortcomings in RE tasks due to insufficient domain-specific relation knowledge (Efeoglu and Paschke, 2024; Kai Zhang, 2023; Xiong et al., 2023). Incorporating these relation types into LLMs could enhance RE through zero-shot prompting (Efeoglu and Paschke, 2024). To tackle this issue, we fine-tune language models on small sets of RE prompt datasets to enhance their ability to identify relations between entities at the sentence level. To evaluate the performance of fine-tuned LLMs, we conduct experiments using Llama2-7B [1] (Touvron et al., 2023), Mistral-7B-Instruct-v0.2 [2] (Jiang

[1] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf, accessed on 14.05.2025

[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2, accessed on 14.05.2025

Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025), pages 1–7
August 1, 2025 ©2025 Association for Computational Linguistics

et al., 2023), and Flan T5 Large (Chung et al., 2022) across three RE benchmark datasets: TACRED (Zhang et al., 2017), TACRED-Revisited (TACREV)(Alt et al., 2020) and Re-TACRED (Stoica et al., 2021). In this work, fine-tuning is used to overcome the limitations of zero-shot LLM prompting settings, such as RAG4RE (Efeoglu and Paschke, 2024), in identifying relations between entities across TACRED and its variants. The contributions of our approach are as follows:

- Fine-tuning greatly improved LLM performance, with Flan-T5 Large outperforming larger models like Mistral-7B-Instruct-v0.2 and Llama2-7B on TACRED and its variants.
- Our fine-tuned LLMs, evaluated within RAG4RE, showed strong results on these datasets.
- This study is the first to fine-tune LLMs for the RE task and to systematically compare smaller and larger models like Mistral-7B-Instruct-v0.2 and Llama2-7B by parameter count.

The rest of this paper first summarizes RE approaches using the language models in Section 2 and then introduces our proposed approach [3] in Section 3. Afterwards, the experimental setup and results are presented in Section 4 and discussed in Section 5. Lastly, all concluding remarks and future works are summarized in Section 6.

2 Related Works

Relation Extraction (RE), as a core task of Information Extraction (IE), plays a significant role in natural language processing. RE aims to identify or classify the relations between (head and tail) entities in a given text. In this work, we primarily focus on sentence-level RE approaches.

RE can be achieved through various methods: supervised, unsupervised, distant supervision, weak supervision, and rule-based (Pawar et al., 2017). Supervised methods require costly, annotated data (Pawar et al., 2017); distant supervision reduces data needs but risks noise (Aydar et al., 2020); weak supervision may lead to semantic drift (Agichtein and Gravano, 2000); and rule-based methods are limited by predefined rules (Pawar et al., 2017). In addition to the fundamental approaches, leading RE methods with fine-tuned LLMs include Cohen et al.'s span prediction for broader entity relations (Cohen et al., 2020), DeepStruct's structural enhancements, Zhou et al.'s entity-aware self-attention (Zhou and Chen, 2022), and Li et al.'s label graph for top-K prediction analysis (Li et al., 2022). Furthermore, Zhang et al. (Kai Zhang, 2023) used multiple-choice prompts, improving RE predictions with added context, though it does not surpass prior rule-based methods. Chen et al. (Chen et al., 2024) introduced context-aware prompt tuning, while RAG4RE (Efeoglu and Paschke, 2024) utilized retrieval-augmented prompting, all tested on TACRED and similar benchmarks.

In this work, we aim to fine-tune LLMs on RE prompt datasets to improve domain adaptation and evaluate their performance on benchmark datasets.

3 Methodology

This work addresses the challenge of sentence-level RE using general-purpose LLMs with zero-shot prompting. General-purpose LLMs struggle with domain-specific relation types, so we fine-tune them on a small RE prompt dataset to improve their ability to identify entity relations. We detail the fine-tuning process in Section 3.1 and describe the integration of fine-tuned LLMs into the RAG4RE approach in Section 3.2.

3.1 Fine-tuning Models on Prompt Datasets

We fine-tune both encoder-decoder models (such as Flan-T5) and decoder-only models, e.g., Llama2-7B and Mistral-7B, on RE prompt datasets using the Supervised Fine-Tuning Trainer (SFT) [4]. This fine-tuning process facilitates domain adaptation for general-purpose LLMs. The SFT approach, which requires labeled training data, is straightforward to implement and train. Additionally, we utilize the Low-Rank Adaptation for quantized language models (QLoRA) method (Dettmers et al., 2023) to fine-tune LLMs. QLoRA optimizes model parameters for text generation while minimizing memory usage on GPUs, which is crucial in scenarios with limited GPU memory.

3.1.1 Prompt Dataset Generation.

The RE prompt dataset is constructed following the template outlined in a previous study by (Efeoglu

[3] The source code: https://github.com/sefeoglu/fine-tuned-llm-relation-extraction

[4] SFT: https://huggingface.co/docs/trl/sft_trainer

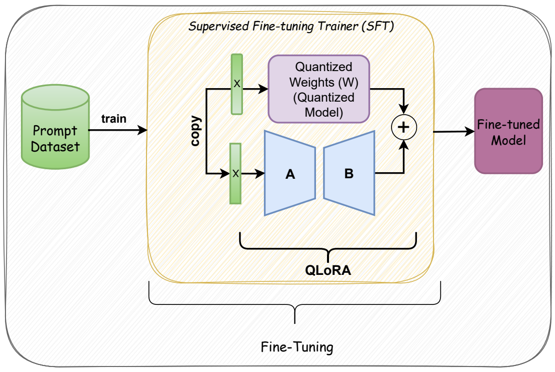

Supervised Fine-tuning Trainer (SFT)
Prompt Dataset
train
copy
x
x
Quantized Weights (W) (Quantized Model)
+
A
B
QLoRA
Fine-tuned Model
Fine-Tuning

Figure 2: Fine-tuning a pre-trained model on a prompt dataset alongside the QLoRA adapter and SFT.

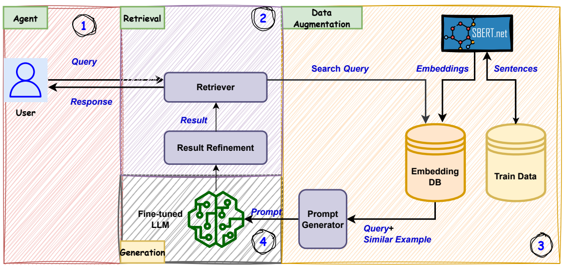
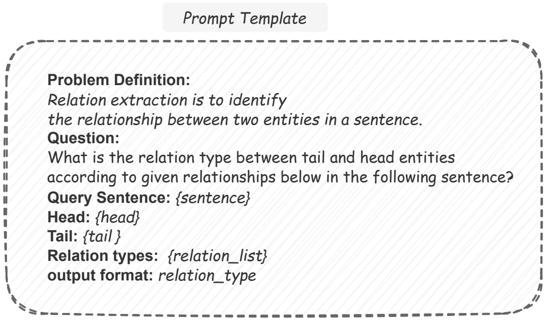


Figure 3: RAG with fine-tuned Large Language Models.

and Paschke, 2024). This dataset originates from a supervised dataset within a single domain and utilizes a specialized template for fine-tuning, as illustrated in Figure 4.


Prompt Template

Problem Definition:
Relation extraction is to identify
the relationship between two entities in a sentence.
Question:
What is the relation type between tail and head entities
according to given relationships below in the following sentence?
Query Sentence: *{sentence}*
Head: *{head}*
Tail: *{tail}*
Relation types: *{relation_list}*
output format: *relation_type*

Figure 4: A prompt template for fine-tuning a Large Language Model.

3.1.2 Parameter Efficient Fine-Tuning.

We utilize QLoRA, a parameter-efficient fine-tuning method that begins by applying quantization to a pre-trained language model. This technique reduces the model's high-precision floating-point representation to a lower precision, thus decreasing memory usage. In particular, we use the "4-bit NormalFloat (NF4)" format, which is optimized for normally distributed data and has been shown to outperform traditional 4-bit integers and floats (Dettmers et al., 2023). Following quantization, LoRA is applied to specific model modules. Fine-tuning is subsequently conducted using the SFT on a single-domain, task-specific prompt dataset. The entire process is illustrated in Figure 2.

3.2 Retrieval-Augmented Generation with Fine-Tuned Models

The Retrieval-Augmented Generation-based Relation Extraction (RAG4RE) approach, introduced by (Efeoglu and Paschke, 2024), comprises three modules: i.) Retrieval, ii.) Data Augmentation, and iii.) Generation. In our implementation, we integrate fine-tuned LLMs, trained on RE prompt datasets, into the generation module of the RAG4RE approach (Efeoglu and Paschke, 2024) to address the task of identifying relations between entities in sentences, as illustrated in Figure 3. Specifically, the LLM used in the generation module of RAG4RE is replaced with our fine-tuned LLMs, while all other components of RAG4RE remain unchanged.

4 Evaluation

We evaluate our approach using three benchmark datasets and language models. In Section 4.1, we detail the datasets, metrics, and experimental settings, including the fine-tuning of language models and the use of Retrieval-Augmented Generation with these fine-tuned models. Then, we present and analyze the experimental results, comparing them with those of previous high-performing RE methods in Section 4.2.

4.1 Experimental Setup

Through this section, we initially introduce the datasets utilized for evaluation, followed by a detailed settings used on the fine-tuning and the RAG4RE framework (Efeoglu and Paschke, 2024) leveraging our fine-tuned language model within its generation module.

Datasets. We utilize three RE benchmark datasets: TACRED (Zhang et al., 2017), TACREV (Alt et al., 2020), and Re-TACRED (Stoica et al., 2021) as detailed in Table 1. The prompt datasets are generated from the validation partitions of the benchmark datasets. The training datasets are utilized in the Embedding Database (DB) of RAG4RE (Efeoglu and Paschke, 2024), while the test splits are used for evaluation. We ensure a strict separation between the training and test splits across all benchmark datasets.

Table 1: The table gives the number of sentences in the test, train, and prompt datasets, as well as the number of relations per benchmark dataset.

Split	TACRED	TACREV	Re-TACRED
Train	68124	68124	58465
Test	15509	15509	13418
Validation	22631	22631	19584
Prompt Dataset (Generated from Validation)	22631	22631	19584
# of Relations	42	42	40

4.1.1 Metrics

The benchmark datasets used—TACRED and its variants—are imbalanced, with a high proportion of "no_relation" labels (Alt et al., 2020; Stoica et al., 2021), necessitating the use of micro metrics. For instance, in the TACRED test split, 12,184 out of 15,509 relations are labeled as "no_relation". We evaluate our experiments using the micro F1-score, precision, and recall across all three benchmark datasets.

4.1.2 Settings for Models

We employed the fine-tuning approach from Section 3.1, using a single GPU with 48 GB of memory and the parameters detailed below. Building on prior studies in RE with language models (Efeoglu and Paschke, 2024; Kai Zhang, 2023), we utilized the following LLMs:

- Flan T5 Large: An encoder-decoder model (Chung et al., 2022; Pan et al., 2024) with 770M parameters. LoRA parameters: alpha=32, dropout=0.01, r=4. Hyperparameters: learning rate=5e-5, batch size=8, one epoch.

- Mistral-7B (Jiang et al., 2023; Pan et al., 2024) and Llama2-7B (Pan et al., 2024; Touvron et al., 2023): Decoder-only models with 7B parameters, used in (Efeoglu and Paschke, 2024). We used Mistral-7B-Instruct-v0.2 [5]. LoRA parameters: alpha=16, dropout=0.1, r=64. Hyperparameters: learning rate=2e-4, batch size=4, one epoch, weight decay=0.001.

4.1.3 Settings for RAG4RE

Due to limited GPU resources, we were unable to fine-tune the Flan-T5 XL model used in the original RAG4RE (Efeoglu and Paschke, 2024). Therefore, all experimental settings are replicated from RAG4RE with Flan-T5 Large. We strictly adhere to the experimental setups established in RAG4RE for our study.

4.2 Results

We evaluated language models fine-tuned on prompt datasets detailed at Table 1 in Section 4.1. Furthermore, we integrated these fine-tuned language models into the RAG4RE (Efeoglu and Paschke, 2024). It is worth noting that due to constraints in GPU resources, we opted to utilize Flan-T5 Large instead of Flan-T5 XL or XXL for fine-tuning. Hence, we chose Flan-T5 Large and meticulously replicated the RAG4RE experiments within the confines of our work. In this section, we first introduce the results of our fine-tuned models and then the results of RAG4RE approach using our fine-tuned models.

With regard to evaluation of fine-tuned LLMs alongside LoRA on four different datasets, fine-tuned Mistral-7B models accomplish outstanding performance at Table 2. Notably, these fine-tuned Mistral-7B models achieve remarkable F1 scores of 89.64%, 94.61%, and 90.09% on TACRED, TACREV, and Re-TACRED, respectively (see Table 2). The Llama2-7B models fine-tuned on TACRED and TACREV follow the fine-tuned Mistral-7B models with micro-F1 scores of 88.20% and 93.75%. Unfortunately, the fine-tuned Llama2-7B models could not exhibit the same performance on Re-TACRED at Table 2. The fine-tuned Flan-T5 Large model takes second place with a F1 score of 86.94% on Re-TACRED dataset (see Table 2). Moreover, fine-tuning LLMs outperformed simple query prompting and the previously introduced RAG4RE method (Efeoglu and Paschke, 2024). Additionally, we integrated these fine-tuned LLMs into the RAG4RE approach (Efeoglu and Paschke, 2024) in order to explore their potential in addressing the limitations of general-purpose LLMs.

Remarkably, the integration of fine-tuned models into RAG4RE yielded significant improvements across all three datasets, including TACRED, TACREV and Re-TACRED, particularly when leveraging Flan-T5 Large at Table 2. While we observed enhancements in RAG4RE's perfor-

[5] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

Table 2: Experimental results on three benchmark datasets using different large language models (LLMs) and methods.

		TACRED			TACREV			Re-TACRED		
LLM	**Method**	**P (%)**	**R (%)**	**F1 (%)**	**P (%)**	**R (%)**	**F1 (%)**	**P (%)**	**R (%)**	**F1 (%)**
T5 Large	Simple Query	95.10	03.18	06.16	96.72	06.90	12.89	90.91	00.26	00.51
	RAG4RE	85.99	34.50	49.20	91.28	08.20	15.04	80.77	00.27	00.53
	Fine-tuning (QLoRA)	86.74	86.76	86.74	89.93	90.13	90.03	86.27	87.62	86.94
	RAG4RE + Fine-tuning	89.93	94.17	**92.00**	95.02	93.66	94.34	92.31	93.73	**93.01**
LLaMA2-7B	Simple Query (Efeoglu and Paschke, 2024)	84.97	01.21	02.38	74.64	00.44	00.87	80.20	00.94	01.86
	RAG4RE (Efeoglu and Paschke, 2024)	81.23	55.01	65.59	84.89	54.57	66.43	55.93	03.46	06.52
	Fine-tuning (QLoRA)	88.07	88.34	88.20	90.07	97.73	93.75	87.54	44.58	59.08
	RAG4RE + Fine-tuning	80.29	89.18	84.50	84.10	97.26	90.22	83.53	68.16	75.07
Mistral-7B	Simple Query (Efeoglu and Paschke, 2024)	94.67	11.96	21.23	92.34	05.15	09.75	64.64	05.48	10.11
	RAG4RE (Efeoglu and Paschke, 2024)	87.81	30.10	44.83	93.23	22.59	36.36	60.19	30.08	40.11
	Fine-tuning (QLoRA)	94.73	85.06	89.64	95.79	93.48	**94.61**	92.40	87.83	90.09
	RAG4RE + Fine-tuning	86.57	82.88	84.68	97.58	79.33	87.50	90.86	85.95	88.33

mance, as detailed in (Efeoglu and Paschke, 2024), with the integration of fine-tuned Llama-7B on Re-TACRED, it is noteworthy that this improvement was not observed on TACRED and TACREV. Regrettably, the results indicate that the use of Mistral-7B as the fined-tuned LLM did not yield improvements in the results of RE. The reason why the performance of the RAG4RE approach could not be improved when fine-tuned decoder-only models are used as a generator in its architecture (see Figure 3) might be related to catastrophic forgetting. Previous work fine-tuning language models on a single task is also dealing with the same forgetting problem (Feng et al., 2024).

As a result, the fine-tuned Flan-T5 Large models consistently achieved the highest F1 scores among all the experiments conducted in this work, particularly when integrated into the RAG4RE framework proposed in (Efeoglu and Paschke, 2024). However, fine-tuned Mistral is slightly better than RAG4RE using fine-tuned Flan-T5 Large on TACREV. In addition to the findings of the experiments using Flan-T5 Large, both fine-tuning language models on the dataset and integrating these fine-tuned models into RAG4RE outperformed zero-shot prompting approaches, such as simple queries and RAG4RE (Efeoglu and Paschke, 2024) (see Table 2).

5 Discussion

Our findings demonstrate significant improvements over the original RAG4RE (Efeoglu and Paschke, 2024) results on the TACRED, TACREV, and Re-TACRED datasets, as shown in Table 3, when fine-tuned Flan-T5 Large models are integrated into the RAG4RE approach. Fine-tuning language models, particularly in the context of domain adaptation, led to substantial performance enhancements for both general-purpose LLMs and RAG4RE (Efeoglu and Paschke, 2024) (see Table 3). The F1 scores of RAG4RE combined with fine-tuned LLMs surpassed those of previous approaches across all three datasets, as illustrated in Table 3. Similarly, the F1 scores of the fine-tuned LLMs exceeded those of prior approaches that employed both zero-shot prompting and pre-trained language models (PLMs) (see Table 3). The best-performing results in our experiments, reported in Table 3, surpassed those of approaches using both zero-shot prompting and PLMs on the TACRED, TACREV, and Re-TACRED datasets, achieving F1 scores of 92.00%, 94.61%, and 93.01%, respectively. Furthermore, our RAG4RE+Fine-tuning approach also outperformed the original RAG4RE utilizing general-purpose LLMs. Therefore, our fine-tuned LLMs achieved outstanding results on the TACRED, TACREV, and Re-TACRED datasets when integrated into the RAG4RE framework (Efeoglu and Paschke, 2024).

6 Conclusion

We address domain adaptation challenges in zero-shot relation extraction (RE) with general-purpose LLMs by fine-tuning Flan-T5 Large, Mistral-7B-Instruct-v0.2, and Llama2-7B on TACRED, TACREV, and Re-TACRED datasets. Our fine-tuned models outperformed previous methods, including RAG4RE (Efeoglu and Paschke, 2024). Integrating these fine-tuned LLMs into RAG4RE significantly enhanced its performance, especially with Flan-T5 Large. However, Llama2-7B and Mistral-7B showed inconsistent F1 scores, likely due to single-task fine-tuning issues. Future work will explore multi-task fine-tuning for RE and entity recognition to mitigate catastrophic forgetting (Feng et al., 2024; Liu et al., 2023; Yang et al., 2024).

Table 3: A comparison of our best-performing results with those of prior works in terms of F1-score.

Method Type	Method	TACRED	TACREV	Re-TACRED
PLM-based	DeepStruct (Wang et al., 2022)	76.8%	-	-
	EXOBRAIN (Zhou and Chen, 2022)	75.0%	-	91.4%
	KLG (Li et al., 2022)	-	84.1%	-
	SP (Cohen et al., 2020)	74.8%	-	
	GAP (Chen et al., 2024)	72.7%	82.7%	91.4%
Zero-Shot prompting	LLMQA4RE (Kai Zhang, 2023)	52.2%	53.4%	66.5%
	RationaleCL (Xiong et al., 2023)	80.8%	-	-
	RAG4RE (Efeoglu and Paschke, 2024)	86.6%	88.3%	73.3%
	RAG4RE+Fine-tuning (Ours)	**92.00%**	94.34%	**93.01%**
	Fine-tuning (Ours)	89.64%	**94.61%**	90.09%

Limitations

This approach requires an embedding database within the data augmentation module of the RAG and retrieves the most similar sentence for use in the RAG module. The most similar sentence with the sentence in the query might have low similarity score. The pre-trained language models may already be familiar with these datasets, as noted in (Efeoglu and Paschke, 2024), since they might be trained on these benchmark datasets.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, page 85–94, New York, NY, USA. Association for Computing Machinery.

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

Mehmet Aydar, Ozge Bozal, and Furkan Ozbay. 2020. Neural relation extraction: a survey. *arXiv preprint*.

Zhenbin Chen, Zhixin Li, Yufei Zeng, Canlong Zhang, and Huifang Ma. 2024. Gap: A novel generative context-aware prompt-tuning method for relation extraction. *Expert Systems with Applications*, 248:123478.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, and 12 others. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

Amir DN Cohen, Shachar Rosenman, and Yoav Goldberg. 2020. Relation classification as two-way span-prediction. *arXiv preprint arXiv:2010.04829*.

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

Sefika Efeoglu and Adrian Paschke. 2024. Retrieval-augmented generation-based relation extraction. *Preprint*, arXiv:2404.13397.

Wenfeng Feng, Chuzhan Hao, Yuewei Zhang, Yu Han, and Hao Wang. 2024. Mixture-of-loras: An efficient multitask tuning for large language models. *Preprint*, arXiv:2403.03432.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Ralph Grishman. 2015. Information extraction. *IEEE Expert*, 30(5):8–15.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Yu Su Kai Zhang, Bernal Jiménez Gutiérrez. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of ACL*.

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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Bo Li, Wei Ye, Jinglei Zhang, and Shikun Zhang. 2022. Reviewing labels: Label graph network with top-k prediction set for relation extraction. *Preprint*, arXiv:2212.14270.

Bingchang Liu, Chaoyu Chen, Cong Liao, Zi Gong, Huan Wang, Zhichao Lei, Ming Liang, Dajun Chen, Min Shen, Hailian Zhou, Hang Yu, and Jianguo Li.

2023. Mftcoder: Boosting code llms with multitask fine-tuning. *Preprint*, arXiv:2311.02303.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Sachin Pawar, Girish K. Palshikar, and Pushpak Bhattacharyya. 2017. Relation extraction : A survey. *arXiv preprint*.

George Stoica, Emmanouil Antonios Platanios, and Barnabas Poczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13843–13850.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022. DeepStruct: Pretraining of language models for structure prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 803–823, Dublin, Ireland. Association for Computational Linguistics.

Weimin Xiong, Yifan Song, Peiyi Wang, and Sujian Li. 2023. Rationale-enhanced language models are better continual relation learners. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15489–15497, Singapore. Association for Computational Linguistics.

Haoran Yang, Yumeng Zhang, Jiaqi Xu, Hongyuan Lu, Pheng Ann Heng, and Wai Lam. 2024. Unveiling the generalization power of fine-tuned large language models. *Preprint*, arXiv:2403.09162.

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 and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168, Online only. Association for Computational Linguistics.