Empowering Few-Shot Relation Extraction with The Integration of Traditional RE Methods and Large Language Models

Ye Liu[1,2], Kai Zhang[1,2(✉)], Aoran Gan[1,2], Linan Yue[1,2], Feng Hu[1,2], Qi Liu[1,2], and Enhong Chen[1,2]

[1] School of Data Science, School of Computer Science and Technology,University of Science and Technology of China, Hefei, China
[2] State Key Laboratory of Cognitive Intelligence, Hefei, China
{liuyer,gar,lnyue,fenghufh3}@mail.ustc.edu.cn,
{kkzhang08,qiliuql,cheneh}@ustc.edu.cn

Abstract. Few-Shot Relation Extraction (FSRE), a subtask of Relation Extraction (RE) that utilizes limited training instances, appeals to more researchers in Natural Language Processing (NLP) due to its capability to extract textual information in extremely low-resource scenarios. The primary methodologies employed for FSRE have been fine-tuning or prompt tuning techniques based on Pre-trained Language Models (PLMs). Recently, the emergence of Large Language Models (LLMs) has prompted numerous researchers to explore FSRE through In-Context Learning (ICL). However, there are substantial limitations associated with methods based on either traditional RE models or LLMs. Traditional RE models are hampered by a lack of necessary prior knowledge, while LLMs fall short in their task-specific capabilities for RE. To address these shortcomings, we propose a Dual-System Augmented Relation Extractor (DSARE), which synergistically combines traditional RE models with LLMs. Specifically, DSARE innovatively injects the prior knowledge of LLMs into traditional RE models, and conversely enhances LLMs' task-specific aptitude for RE through relation extraction augmentation. Moreover, an Integrated Prediction module is employed to jointly consider these two respective predictions and derive the final results. Extensive experiments demonstrate the efficacy of our proposed method.

Keywords: Relation Extraction · Few Shot · Large Language Models.

1 Introduction

Relation Extraction (RE) aims to determine the relation expressed between two entities within an unstructured textual context [23]. Few-Shot Relation Extraction (FSRE), as a subtask of RE, seeks to solve the RE problem by utilizing only K instances per relation (K-shot) in the training and validation phases [3,20].

The primary methodologies employed to address the FSRE task have been fine-tuning or prompt tuning techniques grounded on Pre-trained Language

M. Onizuka et al. (Eds.): DASFAA 2024, LNCS 14854, pp. 349–359, 2024.
https://doi.org/10.1007/978-981-97-5569-1_22

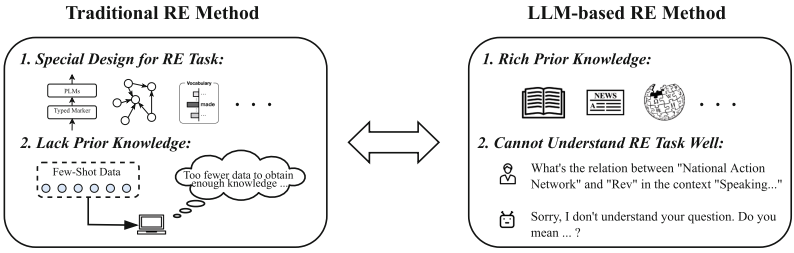

Fig. 1. The comparison between traditional RE methods and LLM-based RE methods.

Models (PLMs) [3,23]. Recently, with the emergence of Large Language Models (LLMs), numerous researchers have embarked on the exploration of FSRE through the In-Context Learning (ICL) technology [5,19,20]. However, there are substantial limitations associated with methods based on either traditional RE models or LLMs. As depicted in Figure 1, although most traditional RE methods are custom-built for the RE task, they still lack necessary prior knowledge that is crucial for resolving many domain-specific cases [3,6]. Acquiring such prior knowledge is particularly challenging in extremely low-resource settings, such as an 8-shot scenario. On the other hand, methods based on LLMs present a contrasting issue. With the scaling of model size and corpus size, LLMs possess an extraordinary amount of prior knowledge. Nevertheless, given that these LLMs are designed for general usage, they lack the task-specific ability for RE, which makes it difficult to fully harness their prior knowledge. This dichotomy between the strengths and weaknesses of traditional RE models and LLMs presents a novel perspective in the field of few-shot relation extraction.

To this end, this paper proposes a novel approach that amalgamates the traditional RE methods with LLMs. By doing so, we aim to address the aforementioned shortcomings by capitalizing on their respective strengths. Specifically, we develop a Dual-System Augmented Relation Extractor (DSARE) for few-shot relation extraction. DSARE consists of three key components: (a) A LLM-augmented RE module: This module designs prompts that enable LLMs to generate additional in-domain labeled data to boost the training of traditional RE models, thereby effectively injecting the prior knowledge of LLMs into the traditional RE methods. (b) A RE-augmented LLM module: This module utilizes the trained RE model to identify and retrieve the most valuable samples from the training data. These samples are subsequently employed as demonstrations for the In-Context Learning of LLMs, thereby enhancing their RE-specific aptitude. (c) An Integrated Prediction module: It takes into account the predictions of both the LLM-augmented RE and RE-augmented LLM modules. When the two predictions differ, a specially designed selector is activated to make a final decision. Finally, extensive experiments on three publicly available datasets demonstrate the effectiveness of our proposed method, and further indicate the necessity to integrate traditional RE models and LLMs.

Our code is available via https://github.com/liuyeah/DSARE.

2 Related Work

Few-shot Relation Extraction. Due to the large computation ability of pre-trained language models, existing few-shot relation extraction methods mainly adopt the fine-tuning method to solve the few-shot relation extraction problem [13,23]. In recent years, in order to bridge the gap between pre-training objectives and RE task, prompt tuning has been proposed and demonstrated remarkable capability in low-resource scenarios [3,6,7].

Currently, with the arise of large language models, many researchers attempt to tackle few-shot relation extraction via In-Context Learning technology [5,19,20]. However, these approaches simply apply LLMs to few-shot relation extraction tasks through straightforward queries, which fails to fully harness the potential of LLMs. More importantly, they overlook the possibility that LLMs and traditional RE models could mutually enhance each other's performance.

Large Language Models. The emergence of Large Language Models (LLMs) such as GPT-4, LLama-2 and others [14,15,17,18], represents a significant advancement in the field of natural language processing. By leveraging In-Context Learning, a novel few-shot learning paradigm was first introduced by [2]. Up to now, LLMs have demonstrated remarkable performance across a range of NLP tasks, such as text classification, named entity recognition, question answering and relation extraction [5,8,19,20].

Previous research efforts [5,19,20] have sought to solve few-shot relation extraction by directly asking LLMs or retrieving more suitable demonstrations. For instance, Wan et al. [19] attempted to introduce the label-induced reasoning logic to enrich the demonstrations. Meanwhile, Xu et al. [20] designed task-related instructions and a schema-constrained data generation strategy, which could boost previous RE methods to obtain state-of-the-art few-shot results.

3 Problem Statement

Let C denote the input text and $e_{sub} \in C$, $e_{obj} \in C$ denote the pair of subject and object entities. Given the entity type of e_{sub}, e_{obj}, and a set of pre-defined relation classes $\mathbb{R}$, relation extraction aims to predict the relation $y \in \mathbb{R}$ between the pair of entities (e_{sub}, e_{obj}) within the context C [19,23].

As for the few-shot settings, following the strategy adopted by [4,20], we randomly sample K instances per relation (K-shot) for the training and validation phases. The whole test set is preserved to ensure the effectiveness of evaluation.

4 DSARE Model

4.1 LLM-augmented RE

LLM Data Augmentation. In this part, we aim to implement the data augmentation via LLMs, anticipated to enrich the training data for relation extraction. Specifically, drawing inspiration from [20], we construct prompts to tell the

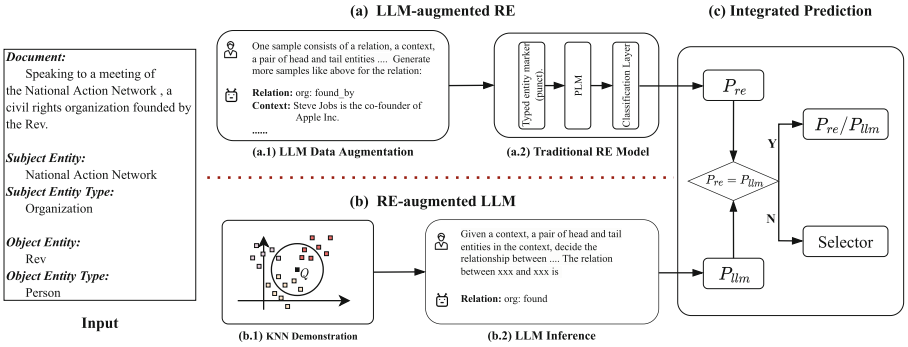

Fig. 2. The architecture of DSARE. It includes three parts: (a) LLM-augmented RE module; (b) RE-augmented LLM module; (c) Integrated Prediction module.

LLM the essential components of one RE training sample, i.e., context text, subject entity, object entity, subject entity type, object entity type and the relation. Then the LLM is guided to create more pseudo RE samples. Upon receiving the outputs from the LLM, we establish rules, such as regular expressions, to transform the augmented RE data into the desired format.

Traditional RE Model. With the augmented datasets, we obtain more diverse data to train a traditional RE model. Here we adopt the Typed Entity Marker (punct) method proposed by [23] to denote the entity and context text, and further train a relation extraction model. Specifically, we utilize the symbols "@" and "#" to denote the start/end of the subject and object entities, and further adopt the symbols "$*$" and "$\wedge$" to indicate the subject and object entity types, respectively. The processed text is then fed into the pre-trained language model to obtain the representations of the subject and object entities (h_{sub}, h_{obj}) via the special token "@" and "#". Finally, we pass (h_{sub}, h_{obj}) into a classification layer to derive the results.

4.2 RE-augmented LLM

KNN Demonstration. In Section 4.1, we train a traditional relation extraction model, which allows us to implement a k-nearest neighbors (KNN) search method to retrieve more valuable samples from the training set. Specifically, we utilize the obtained entity representation $H = [h_{sub}, h_{obj}]$ to represent each sample, and further obtain the representation and label pair (H_i, r_i) on the training set, which we denote as a datastore D.

When inferring a new sample j, we utilize its entity representation H_j to query D according to the euclidean distance to obtain the k nearest neighbors: $\mathcal{N} = \{(H_i, r_i)\}_{i=1}^{k}$, which we adopt as demonstrations for LLM inference.

Table 1. Data Statistics

Dataset	#Train	#Dev	#Test	#Rel
TACRED	8/16/32	8/16/32	15,509	42
TACREV	8/16/32	8/16/32	15,509	42
Re-TACRED	8/16/32	8/16/32	13,418	40

LLM Inference. After obtaining the effective demonstrations, we design prompts to provide the essential information to the LLM, thus generating the LLM results. Specifically, inspired by the various attempts about ICL [20], we first describe the target of the relation extraction task through a instruction. Then, the retrieved k nearest neighbors $\mathcal{N} = \{(H_i, r_i)\}_{i=1}^{k}$ of current sample are followed, which provide the most relevant information to the LLM. Finally, we ask the LLM to predict the relation of current sample.

4.3 Integrated Prediction

In Section 4.1 and 4.2, we apply traditional RE models and LLMs to conduct few-shot relation extraction from dual perspectives. In this part, we aim to obtain the final outputs by considering both the LLM-augmented RE inference result P_{re} and the RE-augmented LLM inference result P_{llm}.

More specifically, as illustrated in Figure 2, if the two results are equal (i.e., $P_{re} = P_{llm}$), our model directly yields the predicted relation. In circumstances where the two results diverge, we design a selector to further ask the LLM to make a choice between these two relations. In order to improve the effectiveness of the selector, we directly retrieve m samples labeled with these two relations from the training dataset, respectively. Subsequently, we ask the LLM via a similar way we introduced in **LLM Inference** to obtain the final results[1].

5 Experiments

5.1 Experimental Setup

Datasets and Evaluation Metrics. For extensive experiments, we conduct our experiments on three widely-used relation extraction datasets: TACRED [22], TACREV [1] and Re-TACRED [16]. More statistics about the datasets can be found in Table 1. Regarding the evaluation metrics, we adopt the micro-F1 scores of RE as the primary metric to evaluate models, considering that F1 scores can assess the overall performance of precision and recall [3,9,10,21].

[1] If the LLM does not make an inference or we are unable to convert the output into the pre-defined relations, we will conclude there is no relation between subject and object entities. Note that *no_relation* is also a relation category in these datasets.

Implementation Details. In this paper, we utilize the *zephyr-7b-alpha* [18] model on Huggingface as the LLM to conduct experiments.

In the Traditional RE Model part (Section 4.1), we adopt *roberta-large* [11] as the base architecture. The batch size is set to 4, and the model is optimized by AdamW [12] with a learning rate of $3e-5$. We train the model on the training set for 50 epochs and choose the best epoch based on the micro-F1 performance on the development set.

In the LLM Data Augmentation part (Section 4.1), we double the K-shot training set through LLMs. That is to say, for an 8-shot training set, we construct 8 pieces of pseudo data per relation, thereby creating the final augmented training set. In the KNN Demonstration part (Section 4.2), we set the number of retrieved nearest neighbors as $k = 8$. In the Integrated Prediction module (Section 4.3), we set the number of retrieved samples for each relation as $m = 4$.

Benchmark Methods. We compare our methods with the state-of-the-art few-shot relation extraction methods. According to the modeling architecture, es, including Traditional RE Methods (① ∼ ④), LLM-based Methods (⑤ ∼ ⑦) and Hybrid Methods (⑧).

- ① **TYP Marker** [23] proposes to incorporate entity representations with typed markers, which presents remarkable performance on the RE task.
- ② **PTR** [7] designs prompt tuning with rules for relation extraction tasks and applies logic rules to construct prompts with several sub-prompts.
- ③ **KnowPrompt** [3] innovatively injects the latent knowledge contained in relation labels into prompt construction with the learnable virtual type words and answer words.
- ④ **GenPT** [6] proposes a novel generative prompt tuning method to reformulate relation classification as an infilling problem, which exploits rich semantics of entity and relation types.
- ⑤ **GPT-3.5** [15], ⑥ **LLama-2** [17], ⑦ **Zephyr** [18] are the advanced LLMs. We leverage the API for GPT-3.5, while adopt the 7B version for LLama-2 (llama-2-7b-chat-hf) and Zephyr (zephyr-7b-alpha). We utlize the prompt from [20] to conduct In-Context Learning.
- ⑧ **Unleash** [20] proposes schema-constrained data generation methods[2] through LLMs, which boost previous RE methods (i.e., KnowPrompt) to obtain more competitive results.

It is worth noting that, for these LLM-based baselines (⑤ ∼ ⑦), due to the limitations of maximum tokens and the fact that these datasets have at least 40 relations, we utilize the one-shot demonstration per relation following the strategy proposed by [20]. In contrast, our DSARE method, as introduced in the Implementation Details part, requires a maximum of 16 demonstrations[3], which

[2] For fair comparison, we apply this data generation method to double the training set, which is the same as our settings introduced in the Implementation Details part.

[3] In the KNN Demonstration part (Section 4.2), the number of retrieved nearest neighbors is $k = 8$. And in the Integrated Prediction module (Section 4.3), we need a maximum of $2m = 8$ additional demonstrations.

Table 2. Experimental Results (%)

Methods	TACRED			TACREV			Re-TACRED		
	K=8	K=16	K=32	K=8	K=16	K=32	K=8	K=16	K=32
① TYP Marker	29.02	31.35	31.86	26.28	29.24	31.55	51.32	55.60	57.82
② PTR	28.34	29.39	30.45	28.63	29.75	30.79	47.80	53.83	60.99
③ KnowPrompt	30.30	33.53	34.42	30.47	33.54	33.86	56.74	61.90	65.92
④ GenPT	35.45	35.58	35.61	33.81	33.93	36.72	57.03	57.66	65.25
⑤ GPT-3.5	29.72			29.98			39.06		
⑥ LLama-2	22.68			21.96			34.31		
⑦ Zephyr	37.10			38.83			35.81		
⑧ Unleash	32.24	33.81	34.76	32.70	34.53	35.28	58.29	64.37	66.03
DSARE (ours)	**43.84**	**45.40**	**45.94**	**44.69**	**46.61**	**46.94**	**60.04**	**66.83**	**67.13**

is much fewer than the number of the one-shot demonstration per relation setting (>= 40), thus avoiding unfair comparison.

5.2 Experimental Result

The main results are illustrated in Table 2. Our proposed DSARE model outperforms all baselines across all metrics. Particularly on the TACRED and TACREV datasets, our method manifests a significant advantage. This demonstrates the effectiveness of our designs and the benefits of integrating traditional RE models and LLMs. Furthermore, there are also some interesting phenomena:

First, the vast majority of methods exhibit superior performance on the Re-TACRED dataset compared to the TACRED and TACREV datasets. This is reasonable as Re-TACRED is an improved version among these three datasets, which addresses some shortcomings of the original TACRED dataset, refactors its training set, development set and test set. The more precise labels contribute to the learning process of these models, thereby yielding superior performance. Second, among these LLM-based methods, Zephyr (7B) demonstrates competitive performance and significantly outperforms GPT-3.5 and LLama-2 on the TACRED and TACREV datasets. This proves its strong information extraction ability, as claimed in [18]. Third, Unleash introduces a schema-constrained data augmentation method through LLMs to enhance the Knowprompt baselines. It achieves a certain degree of improvement compared to Knowprompt, verifying the the feasibility of this line of thinking. And our DSARE model significantly surpasses Unleash, which further demonstrates the effectiveness of our designs from another perspective.

Table 3. Ablation Experiments (%)

Ablation Models	Re-TACRED		
	K=8	K=16	K=32
DSARE	**60.04**	**66.83**	**67.13**
ïż£ LLM-augmented RE	52.53	58.01	58.56
RE-augmented LLM	56.38	64.85	66.03
Pure RE	51.32	55.60	57.82
Pure LLM	35.81		

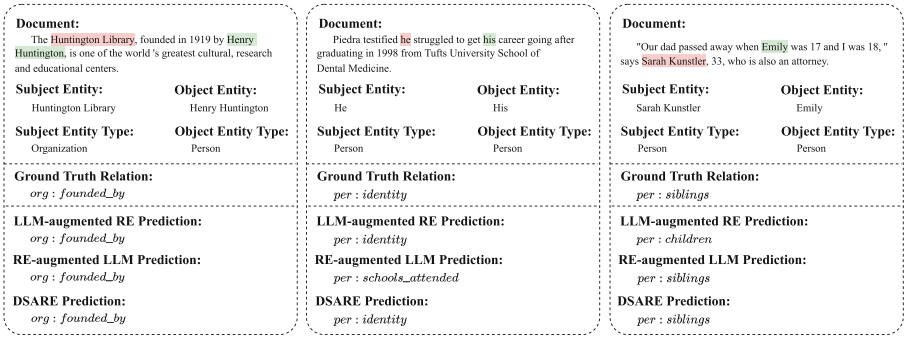

Fig. 3. The case study of the DSARE model. (a) is from the TACRED dataset (K=8), while (b) and (c) are from the Re-TACRED dataset (K=8).

5.3 Ablation Study

In this subsection, we carry out ablation experiments to validate the effectiveness of various components of DSARE model. Specifically, we first remove the Integrated Prediction module, consequently leading to two ablated variants: LLM-augmented RE and RE-augmented LLM. As shown in Table 3, there are obvious decreases between DSARE and its two variants, demonstrating the efficacy of the Integrated Prediction module.

Subsequently, we further remove the LLM Data Augmentation part from LLM-augmented RE and remove the KNN Demonstration part from RE-augmented LLM. This yields two other variants, i.e., Pure RE and Pure LLM[4]. From the results, both these variants perform inferiorly, especially the Pure LLM. These findings further demonstrate the validity and non-redundancy of our designs.

[4] Note that here Pure LLM is equivalent to the baseline ⑦ Zephyr.

5.4 Case Study

In this section, we conduct case study to more intuitively illustrate the effectiveness of integrating traditional RE models and LLMs. Specifically, as illustrated in Figure 3, we present the input information (i.e., document, subject/object entity, subject/object entity type), ground truth relation and the prediction of DSARE and its ablated variants, respectively.

In Figure 3 (a), both the LLM-augmented RE and the RE-augmented LLM make the correct prediction. In Figure 3 (b) and (c), the LLM-augmented RE and RE-augmented LLM correctly infer the relations (*per* : *identity* and *per* : *siblings*), respectively. And with the aid of the Integrated Prediction module, DSARE finally derives the correct predictions. These cases intuitively demonstrate the significant role of integrating traditional RE methods and LLMs, and further verify the validity of our DSARE model.

6 Conclusions

In this paper, we explored a motivated direction for empowering few-shot relation extraction with the integration of traditional RE models and LLMs. We first analyzed the necessity to joint utilize traditional RE models and LLMs, and further proposed a Dual-System Augmented Relation Extractor (DSARE). Specifically, we designed a LLM-augmented RE module, which could inject the prior knowledge of LLMs into the traditional RE models. Subsequently, a RE-augmented LLM module was proposed to identify and retrieve the most valuable samples from the training data, which provided more useful demonstrations for the In-Context Learning of LLMs. More importantly, we designed an Integrated Prediction module to joint consider the predictions of both LLM-augmented RE and RE-augmented LLM modules, thus taking advantages of each other's strengths and deriving the final results. Finally, extensive experiments on three publicly available datasets demonstrated the effectiveness of our proposed method. We hope our work could lead to more future studies.

Acknowledgement. This research was partially supported by grants from the National Natural Science Foundation of China (No. U20A20229), the Anhui Provincial Natural Science Foundation of China (No. 2308085QF229 and 2308085MG226) and the Fundamental Research Funds for the Central Universities.

References

1. Alt, C., Gabryszak, A., Hennig, L.: Tacred revisited: A thorough evaluation of the tacred relation extraction task. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1558–1569 (2020)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)

3. Chen, X., Zhang, N., Xie, X., Deng, S., Yao, Y., Tan, C., Huang, F., Si, L., Chen, H.: Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In: Proceedings of the ACM Web conference 2022. pp. 2778–2788 (2022)
4. Gao, T., Fisch, A., Chen, D.: Making pre-trained language models better few-shot learners. In: ACL-IJCNLP 2021. pp. 3816–3830 (2021)
5. Gutiérrez, B.J., McNeal, N., Washington, C., Chen, Y., Li, L., Sun, H., Su, Y.: Thinking about gpt-3 in-context learning for biomedical ie? think again. In: Findings of EMNLP 2022. pp. 4497–4512 (2022)
6. Han, J., Zhao, S., Cheng, B., Ma, S., Lu, W.: Generative prompt tuning for relation classification. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 3170–3185 (2022)
7. Han, X., Zhao, W., Ding, N., Liu, Z., Sun, M.: Ptr: Prompt tuning with rules for text classification. AI Open **3**, 182–192 (2022)
8. Liu, J., Shen, D., Zhang, Y., Dolan, W.B., Carin, L., Chen, W.: What makes good in-context examples for gpt-3? In: Proceedings of Deep Learning Inside Out (DeeLIO 2022). pp. 100–114 (2022)
9. Liu, Y., Wu, H., Huang, Z., Wang, H., Ning, Y., Ma, J., Liu, Q., Chen, E.: Techpat: technical phrase extraction for patent mining. ACM Transactions on Knowledge Discovery from Data **17**(9), 1–31 (2023)
10. Liu, Y., Zhang, K., Huang, Z., Wang, K., Zhang, Y., Liu, Q., Chen, E.: Enhancing hierarchical text classification through knowledge graph integration. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 5797–5810 (2023)
11. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
12. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
13. Lyu, S., Chen, H.: Relation classification with entity type restriction. In: Findings of ACL-IJCNLP 2021. pp. 390–395 (2021)
14. OpenAI: Gpt-4 technical report (2023)
15. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
16. Stoica, G., Platanios, E.A., Póczos, B.: Re-tacred: Addressing shortcomings of the tacred dataset. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 13843–13850 (2021)
17. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
18. Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., et al.: Zephyr: Direct distillation of lm alignment. arXiv preprint arXiv:2310.16944 (2023)
19. Wan, Z., Cheng, F., Mao, Z., Liu, Q., Song, H., Li, J., Kurohashi, S.: Gpt-re: In-context learning for relation extraction using large language models. arXiv preprint arXiv:2305.02105 (2023)
20. Xu, X., Zhu, Y., Wang, X., Zhang, N.: How to unleash the power of large language models for few-shot relation extraction? In: Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP). pp. 190–200 (2023)

21. Zhang, K., Zhang, K., Zhang, M., Zhao, H., Liu, Q., Wu, W., Chen, E.: Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 3599–3610 (2022)
22. Zhang, Y., Zhong, V., Chen, D., Angeli, G., Manning, C.D.: Position-aware attention and supervised data improve slot filling. In: Conference on Empirical Methods in Natural Language Processing (2017)
23. Zhou, W., Chen, M.: An improved baseline for sentence-level relation extraction. AACL-IJCNLP 2022 p. 161 (2022)