Do Instruction-tuned Large Language Models Help with Relation Extraction?

Xue Li, Fina Polat and Paul Groth

University of Amsterdam

Abstract

Information extraction and specifically relation extraction are key tasks in knowledge base construction. With in-context learning, Large Language Models (LLMs) often demonstrate impressive generalization on unseen information extraction tasks, even with limited examples. However, when using in-context learning for relation extraction, LLMs are not competitive with fully supervised baselines that employ smaller language models. To address this, we explore the potential of instruction-tuning as a mechanism to improve relation extraction performance while preserving in-context capabilities. Our preliminary results demonstrate that instruction-tuned LLMs have the potential to achieve comparable performance with fully supervised smaller LMs. We instruction-tuned a Dolly-v2-3B model using the parameter-efficient approach LoRA on a challenging silver standard relation extraction dataset comprising 1,079 relations. Results show that the instruction-tuned model can achieve a 28.5 micro-F1 and a 27.3 macro-F1 score under a strict matching evaluation strategy. Additionally, manual evaluation with two evaluators shows an average of 66.5% accuracy with 0.760 inter-agreement. You can find access to code and dataset at https://github.com/INDElab/KGC-LLM.git .

1. Introduction

Large language models (LLMs) have exhibited impressive performance across various NLP tasks. Using the in-context learning (ICL) paradigm, wherein models are shown demonstrations to handle new tasks without updating any model parameters, LLMs have showcased performance on par with fully supervised smaller language models (LMs)[1] (such as BERT-based models) while using a limited number of examples[1].

Despite this notable achievement, previous studies have shown that LLMs using ICL still significantly underperform when compared to fully supervised smaller LMs, particularly for relation extraction (RE) [2, 3, 4]. RE represents a fundamental and challenging building block within Information Extraction (IE) pipelines, as it requires semantic understanding of sentences to extract subject-predicate-object triples, which are essential for knowledge base construction (KGC). This limited performance might stem from the low incidence of RE tasks in the dataset used to train the LLMs [4].

To overcome performance deficits when using ICL with LLMs, instruction-tuning [5] can present a different approach to harness the capabilities of LLMs. This involves fine-tuning LLMs on datasets where tasks are described using instructions.

KBC-LM'23: Knowledge Base Construction from Pre-trained Language Models workshop at ISWC 2023

x.li3@uva.nl (X. Li); f.yilmazpolat@uva.nl (F. Polat); p.t.groth@uva.nl (P. Groth)

[1] We refer smaller LMs to LMs that have under 1B parameters

However, as the number of parameters grows, updating these model parameters while working with resource constraints becomes increasingly impractical. To address this, the development of parameter-efficient fine-tuning (PEFT) techniques allows models to adapt to different domains or tasks without consuming excessive time or space. PEFT methods typically modify only a smaller number of additional tunable parameters while keeping the primary model parameters frozen. Two widely adopted PEFT approaches are prefix-tuning [6] and low-rank adaptation (LoRA) [7].

To this end, we explore how instruction tuning can help improve LLMs on RE tasks, using the LoRA technique. Importantly, the injectable learned lower-rank matrices allow us to efficiently adapt to a new task or domain both in time and space. Besides being efficient, this also allows for the retention of properties that are frequently found useful in instruction-tuned models, e.g. the ability to respond to chat-style conversational input or to answer factoid questions.

Specifically, we fine-tune an open source LLM called *Dolly-v2-3B* [8], using the LoRA approach, on a silver standard RE dataset [9] that has been transformed into an instruction-based dataset. The dataset contains around 1,079 different types of relations, making it challenging for smaller-sized LMs.

We applied two evaluation approaches. First, we evaluated the model using an exact match with silver standard labels. Results show the model achieves a 28.5% micro-F1 score and a 27.3 macro-F1 score. After qualitative investigation, we observed that the model generates a substantial number of correct triples that are not included in the dataset annotations. To better assess the model's true performance, we randomly sampled 100 instances from the test set and manually evaluate the triples produced by our model. Furthermore, we also observed that some generations contain correct triples but cannot be derived from the input text (we refer to them as out-of-scope triples). Our hypothesis is that Dolly-v2-3B was fine-tuned on Wikipedia-related data; hence, the model contains related knowledge. Therefore, we apply two criteria for our manual evaluation: (1) if the triple is correct; (2) if the triple can be derived from the input. Our results show an average accuracy of 66.5% with a 0.742 Cohen's Kappa inter-agreement. In addition, the results also indicate that 8.5 % of triples are out-of-scope triples.

In summary, our contributions are threefold:

- Code to transform existing RE datasets into instruction-based datasets [2].
- An instruction-tuned Dolly-v2-3B model capable of performing relation extraction.
- An evaluation of this model performance using both an existing relation extraction baseline dataset complimented with a manual analysis. These initial results indicate that instructed models can potentially be competitive with fully supervised models using less annotated data.

2. Related Work

Relation Extraction: Over the years, several different approaches have been framed for RE. Early approaches treated RE as a pipeline involving named entity recognition followed by relation classification [10]. More recently, end-to-end approaches leveraging the transformer

[2]https://github.com/INDElab/KGC-LLM.git

Instruction: A triple has three components: (subject, relations, object).
Extract triples from the given text.
Input:
Eri-TV is a state-owned Eritrean television station.
Output:
Extracted triples are: (['Eri-TV', 'country', 'Eritrea'],)

Figure 1: One example of a demonstration of the transformed RE dataset.

architectures have been used [11]. Additionally, attempts to employ seq2seq models for RE have gained attention and led to significant improvements [9, 12].

A key challenge is to extract large numbers of entities and relations. For example, the REBEL dataset [9] contains over a thousand types of relations. To tackle this scale, generative models have been employed. A state-of-the-art example is GenIE [12] which frames RE as a Generative Information Extraction task and employs a constrained decoding strategy. Training on the REBEL dataset, GenIE achieved a 68.93 micro-F1 score and 30.46 macro-F1 score.

LLMs with ICL for RE: Despite the high performance of LLMs on various tasks, previous work attempted to explore their performance on RE using ICL [3, 2, 4]. The results indicate that LLMs are not good few-shot learners when it comes to RE. For instance, Jimenez Gutierrez et al. showed that LLMs underperform smaller LMs for biomedical RE.

Instruction-tuning: Recently, supervised fine-tuning on a large number of tasks represented with demonstrations has shown improvements in LLMs' capacity to generalise to unseen tasks [13]. To better exploit knowledge learned by LLMs during pre-training, different adaptation strategies have been developed to make fine-tuning LLMs more practical. For example, prefix-tuning [6] updates only a small part that is the prefix of pre-trained transformers while keeping the rest of the model parameters frozen. LoRA [7] proposes a low-rank adaptation fine-tuning strategy that does not modify the model itself but instead trains injectable lower-rank matrices. An additional advantage of LoRA is that it can used with other strategies, such as prefix-tuning.

Our work differs from the aforementioned in that we transform classic RE datasets to instruction-based datasets and then instruction-tune an LLM using LoRA. Importantly, our data transformation strategy allows any RE dataset to be transformed and fine-tuned with any LLM.

3. Methods

Data Transformation: We convert the REBEL dataset to an instruction-based dataset for fine-tuning LLMs. Unlike building an instruction-based dataset with different tasks[13, 14], our transformed dataset focuses solely on one task: extracting triples. The prompt template we utilize is adapted from [14], which comprises three components: *Instruction*, *Input* and *Output*. An example instruction can be seen in Figure 1. The REBEL training set consists of 3,120,296 samples and 1,079 different types of relations, which we convert.

Instruction-tuning: We proceed to instruction-tune a Dolly-v2-3B model with LoRA. Dolly-v2 [8] is a series of open-sourced large language models based on Pythia-12b, which were instruction-tuned on 15k instructions generated by employees of Databricks Dolly-v2 models

are available in different parameter sizes, ranging from 3B to 12B. Considering computational constraints, we select the 3B model for our experiments. Dolly-v2 has been instruction-tuned on Wikipedia data with questions that required contributors to refer to specific information from given Wikipedia paragraphs. This attribute makes the model well-suited for our RE dataset, given that the texts in the REBEL dataset are also sourced from Wikipedia.
Evaluation: We evaluate the performance of the instruction-tuned model on RE using two distinct methods:

- **Traditional evaluation.** For the traditional evaluation, we strictly match the subject, relation and object triple with gold labels. Subsequently, we calculate the precision, recall and F1 score, both micro and macro, under the assumption that the labels in the REBEL dataset are fully complete and correct.
- **Post-hoc human evaluation.** For the post-hoc human evaluation, we follow the evaluation methods presented by Groth et al. and Wadhwa et al., where human annotators judge the output of model Each triple is assessed based on two criteria: (1) whether the triple is correct or not; (2) whether the triple is correct but cannot be inferred from the provided sentence. The first criterion assesses the precision of the model's generation, while the second one gauges the model's ability to generate correct triples from its background knowledge. We term such correct triples "out-of-scope" triples.

4. Experiments and Results

In our experiments, we employed specific hyperparameters, namely the number of epochs and the ranks of the matrices in LoRA. Ultimately, we conducted our experiments with 3 epochs and a rank of 4, which aligns with the numbers used by Hu et al.. To determine the best-performing model, we evaluated the models on a validation set containing 50 samples due to inference time constraints. Subsequently, we selected the best-performing model to evaluate and report final performance results.

The results of the strict evaluation can be found in Table 1. Notably, the state-of-the-art model outperforms the instruction-tuned model under the assumption that the provided annotations are complete and correct. However, when we assess the precision as evaluated by humans (as shown in Table 2), we observe that the precision is around 66.5%. In both human evaluation criteria, the inter-agreement between the two evaluators exceeds 0.7, indicating a substantial level of agreement between evaluators. We also note that 8.5% of the triples in the human evaluation what we term out-of-scope, namely, they were correct but not entailed by the given sentence.

5. Discussion

Training Data Amount: It is important to note that the reported model is based on 800 steps of fine-tuning, equivalent to 102,400 samples, making up only 33% of the training set, while GenIE is trained on the full dataset. Using a higher rank of adaptation model might be able to improve the performance further.

	Micro			Macro			
	Precision	Recall	F1	Precision	Recall	F1	# of instances for training
GenIE	68.02	69.87	68.93	33.9	30.48	30.46	3,120,296
Instruct-tuned Dolly-v2-3b	36.6	23.3	28.5	36.7	22.6	27.3	102,400

Table 1
Results on of strict evaluation of instruct-tuned model v.s. the state-of-the-art.

	Value	Cohen's kappa
Precision	66.5	0.760
Out-of-scope rate	8.5	0.724

Table 2
Results for human evaluation with two evaluators on randomly sampled 100 instances from the test set.

Disparity between Micro and Macro Measurements: An intriguing observation is the significant performance disparity between micro and macro measurements for GenIE. This indicates that the model performs poorly on certain types of relations but better on others, which could be attributed to the existence of long-tail relations. It is likely that GenIE performs well on dominant and frequent relation types but not so well on less frequent relation types. In contrast, the instruct-tuned model exhibits consistent performance between micro and macro measurements suggesting that it struggles less with long-tail relations.

Performance Increase for Human Evaluation: During the analysis, we noticed that many triples generated by the instruction-tuned model are correct but not included in the dataset annotations. Out of 100 random samples, the instruction-tuned model generates 453 triples. Among these 453 triples, both evaluators agree that 274 triples are correct. Interestingly, when comparing these triples with the dataset annotations, we found that 184 of the human-evaluated correct triples are not included in the REBEL annotations, accounting for 67.2% of the correct triples generated by the model. This finding indicates that the current evaluation approach might have limitations when applied to instruction-tuned models. Evaluating generative LLMs remains a challenge, also for tasks such as mention detection [17]. Moreover, it is noteworthy that the model demonstrates the ability to generate out-of-scope triples, indicating that its generation process relies on both the input context and the knowledge learned from pre-training.

6. Conclusion

Our findings demonstrate the potential of instruct-tuned models for RE, especially when dealing with a substantial number of relations. Even with a 3B model (considerably smaller than the 176B parameters of GPT-3), fine-tuned on a relatively small amount of data, the model already exhibits good performance for RE. We anticipate that further exploration and fine-tuning will likely lead to even better performance. Furthermore, the instruction-tuned model displays the ability to generate out-of-scope triples to a certain extent, indicating that the model retains knowledge acquired during its pre-training, which holds promise for unifying LLMs and

knowledge graphs. Finally, our methods are generalizable and can be applied to any existing RE datasets, underscoring their applicability and potential for future research.

Acknowledgments

This research is supported by Dutch Research Council (NWO) through grant MVI.19.032 and the by the European Union's Horizon Europe research and innovation programme within the ENEXA project (grant Agreement no. 101070305).

References

[1] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, W. Chen, What makes good in-context examples for GPT-3?, in: Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Association for Computational Linguistics, Dublin, Ireland and Online, 2022, pp. 100–114. doi:10.18653/v1/2022.deelio-1.10.

[2] B. Jimenez Gutierrez, N. McNeal, C. Washington, Y. Chen, L. Li, H. Sun, Y. Su, Thinking about GPT-3 in-context learning for biomedical IE? think again, in: Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 4497–4512.

[3] Z. Wan, F. Cheng, Z. Mao, Q. Liu, H. Song, J. Li, S. Kurohashi, Gpt-re: In-context learning for relation extraction using large language models, 2023. arXiv:2305.02105.

[4] K. Zhang, B. J. Gutiérrez, Y. Su, Aligning instruction tasks unlocks large language models as zero-shot relation extractors, 2023. arXiv:2305.11159.

[5] S. Mishra, D. Khashabi, C. Baral, H. Hajishirzi, Cross-task generalization via natural language crowdsourcing instructions, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3470–3487. doi:10.18653/v1/2022.acl-long.244.

[6] X. L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 4582–4597. doi:10.18653/v1/2021.acl-long.353.

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, CoRR abs/2106.09685 (2021). URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[8] M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, R. Xin, Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

[9] P.-L. Huguet Cabot, R. Navigli, REBEL: Relation extraction by end-to-end language

generation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021.

[10] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao, Relation classification via convolutional deep neural network, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 2335–2344.

[11] Y. Wang, B. Yu, Y. Zhang, T. Liu, H. Zhu, L. Sun, TPLinker: Single-stage joint extraction of entities and relations through token pair linking, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 1572–1582. doi:10.18653/v1/2020.coling-main.138.

[12] M. Josifoski, N. D. Cao, M. Peyrard, R. West, Genie: Generative information extraction, CoRR abs/2112.08340 (2021). URL: https://arxiv.org/abs/2112.08340. arXiv:2112.08340.

[13] Y. Wang, S. Mishra, P. Alipoormolabashi, et al., Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 5085–5109.

[14] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language model with self generated instructions (2022). doi:10.48550/arXiv.2212.10560.

[15] P. Groth, M. Lauruhn, A. Scerri, R. Daniel Jr., Open information extraction on scientific text: An evaluation, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 3414–3423.

[16] S. Wadhwa, S. Amir, B. C. Wallace, Revisiting relation extraction in the era of large language models, 2023. arXiv:2305.05003.

[17] D. Daza, M. Cochez, P. Groth, SlotGAN: Detecting mentions in text via adversarial distant learning, in: Proceedings of the Sixth Workshop on Structured Prediction for NLP, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 32–39. doi:10.18653/v1/2022.spnlp-1.4.