Relation-Aware Prompting Makes Large Language Models Effective Zero-shot Relation Extractors

Mahdi Rahimi, Razvan-Gabriel Dumitru, Mihai Surdeanu
Department of Computer Science
University of Arizona, Tucson, Arizona, USA
{marahimi, rdumitru, msurdeanu}@arizona.edu

Abstract

While supervised relation extraction (RE) models have considerably advanced the state-of-the-art, they often perform poorly in low-resource settings. Zero-shot RE is vital when annotations are not available either due to costs or time constraints. As a result, zero-shot RE has garnered interest in the research community. With the advent of large language models (LLMs) many approaches have been proposed for prompting LLMs for RE, but these methods often either rely on an accompanying small language model (e.g., for finetuning on synthetic data generated by LLMs) or require complex post-prompt processing. In this paper, we propose an effective prompt-based method that does not require any additional resources. Instead, we use an LLM to perform a two-step process. In the first step, we perform a targeted summarization of the text with respect to the underlying relation, reduce the applicable label space, and synthesize examples. Then, we combine the products of these processes with other elements into a final prompt. We evaluate our approach with various LLMs on four real-world RE datasets. Our evaluation shows that our method outperforms the previous state-of-the-art zero-shot methods by a large margin. This work can also be considered as a new strong baseline for zero-shot RE that is compatible with any LLM[1].

1 Introduction

Relation extraction (RE) aims to identify semantic relations between two entities from unstructured text. With the recent advances in large language models (LLMs), studies show that LLMs perform well in various downstream tasks without any training or fine-tuning. But it is unclear whether they are effective for zero-shot RE. A recent line of research shows that such zero-shot approaches for relation extraction are ineffective or continue to lag behind supervised methods (Ma et al., 2023, Wang et al., 2023, Ye et al., 2023, Jimenez Gutierrez et al., 2022, Li et al., 2023a, Xu et al., 2023b, Han et al., 2024, Swarup et al., 2025). However, other line of LLM-based research reports results comparable or outperforming state-of-the-art. These methods fall into three groups. In a first group, these methods use LLMs to some extent, but eventually rely on fine-tuning a small language model (Xu et al., 2023b, Zhou et al., 2024, Xu et al., 2023a, Tang et al., 2023). For instance, an LLM is used to generate/augment synthetic data and then a model from BERT family is fine-tuned on the generated data. The second group fine-tunes a large language model (Wadhwa et al., 2023, Sainz et al., 2024, Li et al., 2024, Wang et al., 2023). Fine-tuning LLMs requires specialized hardware, significant compute resources, and is expensive. The third group does not require fine-tuning but requires complex post-prompt computations, e.g., Li et al. (2023c) performs a complex computation on the LLM answers using an uncertainty-based active learning method to estimate output probabilities of the LLM. A few other methods, e.g., (Wei et al., 2024, inter alia), that do not belong to the above groups are evaluated on limited benchmarks. Therefore, it is unclear whether prompt based zero-shot RE is effective without any finetuning or complex post-prompt computations.[2]

In this work, we present an effective prompt-based method to RE that does not require either fine-tuning or complex computations. Our approach only requires an API access to an LLM. This simplifies the zero-shot RE process and makes it more accessible and faster to deploy which is important for developing zero-shot systems. We achieve this by a novel prompt-based method we

[1] Code and data are available at https://github.com/mahrahimi1/relation-aware-prompting

[2] For a detailed review of the literature on zero-shot relation extraction see Appendix A.

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 280–292

call Relation-Aware Prompting. Formally:

(1) We perform a targeted summarization of instances with respect to the underlying relations to bring out the relations in the texts and discard unrelated facts.

(2) We reduce the applicable relation labels using annotation guidelines and through a method inspired by the process of elimination. We use entity type constraints for this purpose when they are available as well.

(3) We propose a method using subject-verb-object (SVO) structure to generate synthetic examples that will be used as demonstrations.

(4) We combine the results of the above processes with other elements such as relation definitions into a final prompt.

We evaluate our approach using various LLMs on four real-world and challenging relation extraction datasets. The evaluation shows that our method outperforms the previous state-of-the-art zero-shot methods by a large margin. We also perform an ablation study where we investigate the effectiveness and usefulness of our prompt elements that will demonstrate the effectiveness of the proposed method.

This work can also be considered as a new strong baseline for relation extraction. Any LLM-based work in RE (such as finetuning LLMs, or other methods) can use our method as a strong baseline for evaluating their respective approach.

2 Problem Statement

In the RE task, the goal is to classify a sentence containing two marked entities (a *head* and a *tail*) into a set of predefined relations, or determine that none of the relations apply (referred to as none-of-the-above or NoTA). This work focuses on zero-shot RE where no RE training data is provided to models prior to inference time.

3 Methodology

Our method is a two-step process. In the first step, (a) relations between the head and tail entities in instances are summarized; (b) the applicable label space is reduced; and (c) examples are synthesized. In the second step, the results of the first step are combined with other prompting elements into a final comprehensive prompt. The following subsections describe each step. Figure 1 demonstrates an overview of our approach.

3.1 Targeted Summarization

We summarize the relation between the head and tail entities in instances (Li et al., 2023c) in order to bring out the relation in the text and discard unrelated facts and misleading cues. The goal is not to summarize the complete sentence, but to summarize the relation between the entities in the sentence. In the prompt, we emphasize this and instruct the model to ignore everything else for the summary. A concrete example as well as our prompt is provided in Appendix C.

3.2 Reducing the Label Space

Relation Extraction usually involves classification between many classes. This is an overly difficult task for LLMs. When entity types are available in the data, we use them to filter out the relation types that are impossible. In case entity typing is unavailable or not applicable,[3] we propose an approach to reduce the number of candidate relations through a method inspired by the process of elimination. As the first step, we ask an LLM to reduce the number of classes down to 3 for each instance given the relation definitions and annotation guidelines. For example, Figure 4 (left) in Appendix F shows our prompt for SemEval 2010 Task 8 dataset. We select the parts of annotation guidelines that we believe are helpful for the LLM to differentiate between relation types given the entities. For instance, guidelines may have a "Restrictions" section in relation definitions that can help the LLM narrow down candidate relations based on entities of the test examples. Figure 4 (right) in Appendix F shows what we selected for Instrument-Agency relation in the aforementioned dataset.

After the candidate relations are narrowed down to three, we add NoTA (if not already included), and then prompt the LLM to select the best option as explained in subsection 3.4 and shown in Figure 6 (bottom) in Appendix F.

If the relations are undirected, i.e., it is not provided which entity in the sentence is the head and which entity is the tail, such as in SemEval 2010 Task 8 dataset, one extra step is required to determine the direction of the relation. Further details are provided in Appendix D.

[3] Such as SemEval 2010 Task 8 dataset where entities are not named entities, but rather common nouns.

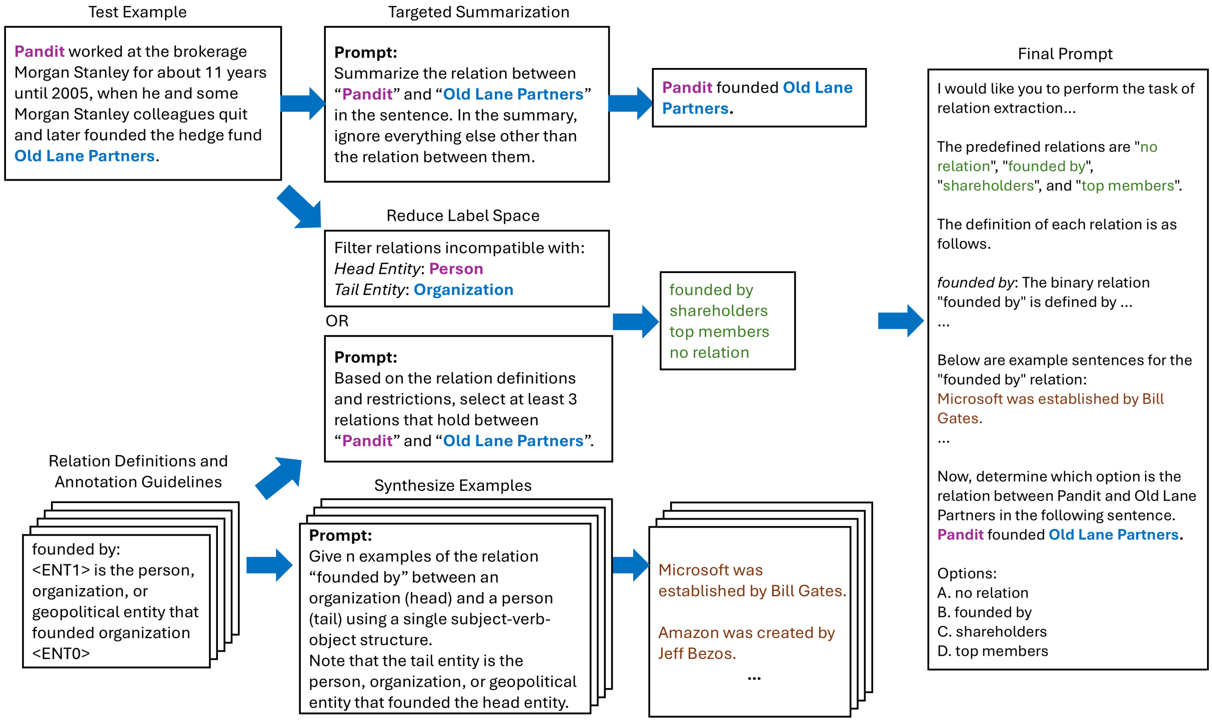


Figure 1: Overview of Relation-Aware Prompting.

3.3 Synthesizing Examples

Some annotation guidelines have examples for each relation type, but we do not use them in our prompt to emulate a no-supervision scenario. Instead, we *synthesise examples*: we prompt an LLM to generate examples using a subject-verb-object (SVO) structure. We generate the examples based on relation definitions, entity types, and relation labels. Figure 5 in Appendix F shows our prompt. After generating the examples, we use them in our final prompt as demonstrations similar to in-context learning (explained in the next subsection).

3.4 Final Prompt

The results of previous processes, i.e, "the targeted summary", "the reduced applicable labels", and "synthesized examples" are combined with other prompting elements to form our final prompt shown in Figure 6 (bottom) in Appendix F. These other prompting elements are entity tagging (Zhou et al., 2024) and relation definitions and annotation guidelines (Zhou et al., 2024). Furthermore, we pose the final classification as multiple-choice question answering (Zhang et al., 2023a) where options are relation labels. Additionally, we turn the labels into a more human-readable form before using them as the options. For example, we change "org:founded_by" to "founded by".

4 Experiments

4.1 Experimental Setup

Datasets We evaluate our method on four relation extraction datasets: TACRED (Zhang et al., 2017), TACREV (Alt et al., 2020), RETACRED (Stoica et al., 2021), and SemEval-2010 Task 8 (Hendrickx et al., 2010) (henceforth SemEval). The statistics of the datasets are provided in Appendix B. We follow previous work (Sainz et al., 2021, Lu et al., 2022, Zhang et al., 2023a, inter alia) to report micro F1 with NoTA relation excluded. Following previous work (Zhang et al., 2023a, Li et al., 2023c) and to keep OpenAI API costs under control, we randomly select 1,000 examples from each dataset's test partition to serve as our test set.

Baselines For small language model-based methods, we selected two low-resource state-of-the-art methods: $\text{NLI}_{\text{DeBERTa}}$ (Sainz et al., 2021) and $\text{SuRE}_{\text{PEGASUS}}$ (Lu et al., 2022). For LLMs baselines, we selected QA4RE (Zhang et al., 2023a) and SUMASK (Li et al., 2023c). We also evaluated the performance of a Vanilla prompting method. Further details of the baselines are as follows.

- $\text{NLI}_{\text{DeBERTa}}$ (Sainz et al., 2021) reformulates RE as a natural language inference (NLI) task and uses a DeBERTa model that is finetuned on MNLI dataset as the entailment engine.

Method	TACRED			TACREV			Re-TACRED			SemEval			Avg
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	F1
$\text{NLI}_{\text{DeBERTa}}$†	42.9	76.9	55.1	43.3	84.6	57.2	71.7	58.3	64.3	22.0	25.7	23.7	50.1
$\text{SuRE}_{\text{PEGASUS}}$†	13.8	51.7	21.8	13.5	54.1	21.6	16.6	34.6	22.4	0.0	0.0	0.0	16.4
Vanilla‡	35.7	51.6	37.9	42.5	77.8	55.0	62.0	81.5	70.5	57.1	63.2	60.0	55.9
QA4RE‡	38.0	82.1	52.0	38.7	85.8	53.4	65.2	91.9	76.3	40.1	48.2	43.8	56.4
SUMASK‡	42.5	76.8	54.7	43.9	80.2	56.8	62.5	73.4	67.5	16.0	15.7	15.8	48.7
Ours‡	56.0	72.5	**63.2**	58.9	82.6	**68.8**	75.5	79.8	**77.6**	80.6	83.8	**82.1**	**72.9**

Table 1: Main results on four RE datasets. We mark the best results in bold. † marks re-implemented results from Zhang et al. (2023a). ‡ denotes our runs with GPT4.1.

LLM	TACRED	TACREV	RETACRED	SemEval	Avg
Gemma 3 27B	57.4	63.6	67.4	65.2	63.4
Llama 3.1 70B	57.3	63.9	73.1	73.8	67.0
Mistral Large 2411	63.2	69.1	72.4	73.2	69.5
GPT4o mini	56.1	66.2	69.0	64.0	63.8
GPT4.1	63.2	68.8	77.6	82.1	72.9

Table 2: Evaluation of our method using various open source and proprietary LLMs on the four RE datasets.

- $\text{SuRE}_{\text{PEGASUS}}$ (Lu et al., 2022) reformulates RE as a summarization task and utilizes $\text{PEGASUS}_{\text{Large}}$ obtaining competitive results in few-shot and fully-supervised settings.

- QA4RE (Zhang et al., 2023a) reformulates RE as multiple-choice question answering in order to take advantage of QA's higher prevalence in instruction-tuning training data of LLMs.

- SUMASK (Li et al., 2023c) for each relation type, generates a set of summarizations and yes/no questions, and then asks a LLM to answer the yes/no questions based on the summarizations. Then performs a computation on the answers using an uncertainty based active learning method to estimate output probabilities of the LLM.

- Vanilla Prompt (Zhang et al., 2023a) is a simple and direct prompt strategy. We use the version from QA4RE authors.

We ran all LLM baselines as well as our method with the same LLM: GPT4.1. We also evaluated our method with various LLMs, namely Gemma 3 27B, Llama 3.1 70B, Mistral Large 2411 (123B), and GPT4o mini. The details of the implementation of our method are provided in Appendix E.

4.2 Results

Our evaluation of zero-shot relation extraction on the four RE datasets is shown in Table 1. Our Relation-Aware Prompting technique outperform SOTA methods in all four datasets. Our method provides significant improvements of 8.5 F1 points on TACRED, 12 points on TACREV, 1.3 points on RETACRED, and 22.1 points on SemEval. The improvements on SemEval are important because the dataset has been known to be more challenging for zero-shot methods due to (1) lack of entity typing, (2) relations being undirected, and (3) overlapping relations between the same entity mentions. These results are highly encouraging considering that our method relies solely on off-the-shelf LLMs and no additional components. We also evaluated our method with various open source and proprietary LLMs shown in Table 2. While bigger models perform slightly better, our method works across all LLMs. Even our method evaluated on Gemma 27B outperforms prompting baselines such as QA4RE ans SUMASK that are evaluated with GPT4.1 on three out of four datasets, even though Gemma is orders of magnitude smaller than GPT4.1.

4.3 Ablation Study

We conduct an ablation study to analyze the effectiveness of the proposed elements of our method. The experiments were run on a subset of the development partitions of TACRED and SemEval. We randomly sampled 1000 examples from the development sets. We selected GPT4o mini and Gemma 3 27B to conduct the experiments. In each experiment we remove an element of our main prompt and report the results. In each experiment the number of synthesized examples is a hyperparameter chosen from {0, 1, 5, 10} via hyperparam-

Prompts	TACRED		SemEval	
	4o mini	Gemma3	4o mini	Gemma3
Main Prompt	**61.9**	**65.5**	**62.6**	**66.1**
w/o Rel. Defs.	60.9	62.6	56.4	58.9
w/o Targeted Sum.	57.2	58.1	–	–
w/o Reduc. Label Space	45.1	46.5	58.5	63.1

Table 3: Ablation study on TACRED and SemEval.

eter search.

Table 3 shows the results. We observe that removing "Relation Definitions and Annotation Guidelines", "Targeted Summarization", and "Reducing the Label Space" from our final prompt decreases the performance considerably (as mentioned before, we do not do targeted summarization for SemEval), reaffirming the effectiveness of the proposed components.

5 Conclusion

In this work, we present Relation-Aware Prompting, an effective prompt-based method for zero-shot relation extraction. We propose targeted summarization of instances with respect to the underlying relations to bring out the relations in the texts, reducing the applicable relations through a method inspired by the process of elimination, synthesizing examples using subject-verb-object structure, and other prompting elements. We evaluate our approach on four RE datasets. Our approach significantly outperforms current zero-shot LLM prompt-based methods. Our approach can also be considered as a new strong baseline for zero-shot RE that is compatible with any LLM.

Limitations

We conduct comprehensive experiments exclusively on zero-shot RE and showed that our approach is a new, robust state-of-the-art method. However, we did not engage in few-shot RE, domain-specific explorations, or other languages. Thus, the performance of our method on these settings is still unclear. We acknowledge these matters and leave answering these questions for future work.

Acknowledgments

The authors would like to thank Guozheng Li and other authors of SUMASK prompting work for providing us the source code of their original work.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Alberto Cetoli. 2020. Exploring the zero-shot limit of FewRel. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1447–1451, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Chih-Yao Chen and Cheng-Te Li. 2021. ZS-BERT: Towards zero-shot relation extraction with attribute representation learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3470–3479, Online. Association for Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6250–6255, Hong Kong, China. Association for Computational Linguistics.

Jiaying Gong and Hoda Eldardiry. 2021. Prompt-based zero-shot relation classification with semantic knowledge augmentation. *arXiv preprint arXiv:2112.04539*.

Jiale Han, Shuai Zhao, Bo Cheng, Shengkun Ma, and Wei Lu. 2022a. Generative prompt tuning for relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3170–3185, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ridong Han, Chaohao Yang, Tao Peng, Prayag Tiwari, Xiang Wan, Lu Liu, and Benyou Wang. 2024. An empirical study on information extraction using large language models. *Preprint*, arXiv:2305.14450.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2022b. Ptr: Prompt tuning with rules for text classification. *AI Open*, 3:182–192.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 326–336, Toronto, Canada. Association for Computational Linguistics.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical IE? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 178–181, Barcelona, Spain. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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. *Preprint*, arXiv:2304.11633.

Bo Li, Dingyao Yu, Wei Ye, Jinglei Zhang, and Shikun Zhang. 2023b. Sequence generation with label augmentation for relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13043–13050.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023c. Revisiting large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6877–6892, Singapore. Association for Computational Linguistics.

Guozheng Li, Peng Wang, Jiajun Liu, Yikai Guo, Ke Ji, Ziyu Shang, and Zijie Xu. 2024. Meta in-context learning makes large language models better zero and few-shot relation extractors. *Preprint*, arXiv:2404.17807.

Junpeng Li, Zixia Jia, and Zilong Zheng. 2023d. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5495–5505, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. 2022. Summarization as indirect supervision for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6575–6594, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shengfei Lyu and Huanhuan Chen. 2021. Relation classification with entity type restriction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 390–395, Online. Association for Computational Linguistics.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10572–10601, Singapore. Association for Computational Linguistics.

Chaoxu Pang, Yixuan Cao, Qiang Ding, and Ping Luo. 2023. Guideline learning for in-context information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15372–15389, Singapore. Association for Computational Linguistics.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Mahdi Rahimi and Mihai Surdeanu. 2023. Improving zero-shot relation classification via automatically-acquired entailment templates. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 187–195, Toronto, Canada. Association for Computational Linguistics.

Ofer Sabo, Yanai Elazar, Yoav Goldberg, and Ido Dagan. 2021. Revisiting few-shot relation classification: Evaluation data and classification schemes. *Transactions of the Association for Computational Linguistics*, 9:691–706.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. Gollie: Annotation guidelines improve zero-shot information-extraction. *Preprint*, arXiv:2310.03668.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13843–13850.

Xiaofei Sun, Linfeng Dong, Xiaoya Li, Zhen Wan, Shuhe Wang, Tianwei Zhang, Jiwei Li, Fei Cheng, Lingjuan Lyu, Fei Wu, and Guoyin Wang. 2023. Pushing the limits of chatgpt on nlp tasks. *Preprint*, arXiv:2306.09719.

Anushka Swarup, Tianyu Pan, Ronald Wilson, Avanti Bhandarkar, and Damon Woodard. 2025. LLM4RE: A data-centric feasibility study for relation extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6670–6691, Abu Dhabi, UAE. Association for Computational Linguistics.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *Preprint*, arXiv:2303.04360.

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

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context learning for relation extraction using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547, Singapore. Association for Computational Linguistics.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2021a. Zero-shot information extraction as a unified text-to-triple translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1225–1238, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021b. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418, Online. Association for Computational Linguistics.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *Preprint*, arXiv: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. 2024. Chatie: Zero-shot information extraction via chatting with chatgpt. *Preprint*, arXiv:2302.10205.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2361–2364.

Benfeng Xu, Quan Wang, Yajuan Lyu, Dai Dai, Yongdong Zhang, and Zhendong Mao. 2023a. S2ynRE: Two-stage self-training with synthetic data for low-resource relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8186–8207, Toronto, Canada. Association for Computational Linguistics.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023b. 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)*, pages 190–200, Toronto, Canada (Hybrid). Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *Preprint*, arXiv:2303.10420.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. 2023a. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 794–812, Toronto, Canada. Association for Computational Linguistics.

Wenjie Zhang, Xiaoning Song, Zhenhua Feng, Tianyang Xu, and Xiaojun Wu. 2023b. Labelprompt: Effective prompt-based learning for relation classification. *arXiv preprint arXiv:2302.08068*.

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.

Sizhe Zhou, Yu Meng, Bowen Jin, and Jiawei Han. 2024. Grasping the essentials: Tailoring large language models for zero-shot relation extraction. *Preprint*, arXiv:2402.11142.

A Related Work

A.1 Pre-LLM Works

Prior to the advent of large language models, most recent approaches for supervised relation extraction use pretrained masked language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) or adapt sequence-to-sequence models to the task, such as T5 (Raffel et al., 2020) and BART (Lewis et al., 2020). Traditional RE methods needed a large amount of labeled data for training models from scratch (Kambhatla, 2004, Zeng et al., 2014). The pre-LLM recent approaches outperform traditional approaches by finetuning a masked language model (Wu and He, 2019, Joshi et al., 2020, Yamada et al., 2020, Wang et al., 2021b, Lyu and Chen, 2021, Paolini et al., 2021, Wang et al., 2022, Li et al., 2023b) or prompting a masked language model (Han et al., 2022b, Han et al., 2022a, Zhang et al., 2023b).

As for low-resource RE, several approaches have been proposed for relation extraction with few training examples (Han et al., 2018, Gao et al., 2019, Baldini Soares et al., 2019, Sabo et al., 2021). For the problem of zero-shot RE, approaches leverage techniques such as similarity based Siamese architectures (Chen and Li, 2021) and indirect supervision as task reformulation. In the literature, zero-shot RE has been reformulated as other tasks such as reading comprehension (Levy et al., 2017), textual entailment (Sainz et al., 2021, Rahimi and Surdeanu, 2023), summarization (Lu et al., 2022), span-prediction (Cohen et al., 2020), question answering (Cetoli, 2020), triple generation (Wang et al., 2022, Wang et al., 2021a), and prompting (Gong and Eldardiry, 2021).

A.2 LLM-based Works

Since our work is a LLM-based approach, we focus the rest of the section on similar LLM-based methods for RE with special focus on zero-shot methods. A common approach is prompting LLMs for data generation and then use the generated data to finetune a small language model (Xu et al., 2023b, Zhou et al., 2024, Xu et al., 2023a, Tang et al., 2023). Another common approach is finetuning a retriever to retrieve relevant training examples to be used as in-context learning demonstrations (Sun et al., 2023, Wan et al., 2023). We do not use finetuning or retrieval in our approach. The rest of the common approaches is as follows.

LLM Prompt-Based Methods: In addition to vanilla prompting (Ye et al., 2023, Li et al., 2023a, Ma et al., 2023, Jahan et al., 2023), several approaches have been proposed. Li et al. (2023c) is a zero-shot method that for each relation type, generates a set of summarizations and yes/no questions, and then asks a LLM to answer the yes/no questions based on the summarizations. Then performs a complex computation on the answers using an uncertainty based active learning method to estimate output probabilities of the LLM. Wei et al. (2024) turns zero-shot IE tasks including entity-relation triple extraction into an interactive dialogue-like multiple turns QA. Zhang et al. (2023a) reformulates RE as multiple-choice question answering in order to take advantage of QA's higher prevalence in instruction-tuning training data of LLMs. In this method, manually-constructed relation verbalization templates are used to generate the options of multiple-choice questions. Agrawal et al. (2022) uses a guided prompt design to direct the LLM towards a structured output for clinical relation extraction. Our approach is different from these approaches as we do not require complex post-prompt computations or interactive dialogue-like QA or guided prompt design.

Methods That Use Annotation Guidelines: Zhou et al. (2024) uses annotation guidelines to prompt a LLM to generate synthetic data and then trains a small language model with this data for zero-shot RE. Sainz et al. (2024) uses annotation guidelines to finetune a large language model for IE tasks. They puts annotation guidelines, input and gold output in the prompt to finetune the LLM. Then use the LLM to perform zero-shot IE on unseen datasets. Pang et al. (2023) does not use guidelines, but rather learns them and then use them for prompting LLMs. They automatically synthesize a set of guidelines based on a few error cases, and during inference retrieve helpful guidelines for better classification. Li et al. (2023d) integrates a LLM and a natural language inference (NLI) module to generate relation triples. They use relation descriptions to construct hypotheses for NLI and to guide NLI to output expected relations. Our approach is different from these approaches as we only use guidelines for our prompt without finetuning or using NLI.

Methods That Finetune LLMs: Wadhwa et al. (2023) finetunes a T5 model using Chain of Thought style explanations generated by GPT-3.

Li et al. (2024) uses a meta-training framework for zero and few-shot RE by tuning a LLM to perform in-context learning on 12 RE datasets, and then evaluate it on unseen RE benchmarks. Wang et al. (2023) proposes a unified information extraction framework, and reformulates IE tasks to the sequence-to-sequence form and solves them through fine-tuning LLMs. Our approach is different from these methods as we don't finetune LLMs.

Summarization: Li et al. (2023c) produces k targeted summarizations, questions, and answers for each relation type. Then the vector representations of these items are generated and used to estimate the conditional probabilities for each relation type. Instead, we use targeted summarization once and place it directly in our final prompt. Lu et al. (2022) reformulates RE as a summarization task. They convert input sentences with an entity information verbalization technique and convert output relations with label verbalization templates. Then with the converted inputs and outputs that suit a summarization model, they adopt such a model. The model is pretrained on summarization tasks and then simply finetuned with the converted inputs and outputs. This method is different from ours as: (a) it requires finetuning of a summarization model whereas ours is zero-shot; and (b) the summary output are the verbalization templates whereas ours are more natural.

B Dataset Statistics

The statistics of the datasets are shown in Table 4.

Dataset	# train	# dev	# test	# rel.
TACRED	68,124	22,631	15,509	42
TACREV	68,124	22,631	15,509	42
RETACRED	58,465	19,584	13,418	40
SemEval	8,000	-	2,717	19

Table 4: Statistics of TACRED, TACREV, RETACRED and SemEval.

C Targeted Summarization Example

Complex sentences can confuse LLMs due to misleading cues. Figure 3 (top) shows an example. For this example, ChatGPT predicts the relation of "other family" (a family relation other than immediate family), but the gold label is "no relation". The presence of some cues in the sentence such as the word "family" may have confused the model. The prediction of the model on the summarized version, however, is correct.

D Determining the Direction of a Relation

If relations are undirected, as in SemEval 2010 Task 8 dataset, one extra step is required to determine the direction of the relation. To this end, we prompt the LLM to choose the directionality of the relation from two options that are created from a template. The template is chosen from the very first sentence of the relation definitions. For instance, for "Instrument-Agency" we use the following sentence as the template: `X is the instrument (tool) of Y or, equivalently, that Y uses X`. We create two sentences with the template. For one sentence, we replace "X" with the head entity and replace "Y" with the tail entity. For the other sentence, we swap the entities. Finally, we use the two sentences as options of a multiple-choice question in the prompt. Figure 2 shows the prompt.

If the relation between "master" and "stick" or between "stick" and "master" is "Instrument-Agency", then which one of the following options is correct?

Options:
A. master is the instrument (tool) of stick or, equivalently, that stick uses master.
B. stick is the instrument (tool) of master or, equivalently, that master uses stick.

Respond with a single letter without any additional things including your notes and explanations.

Figure 2: Our prompt for selecting the direction of a relation in SemEval dataset.

E Implementation Details of Our Method

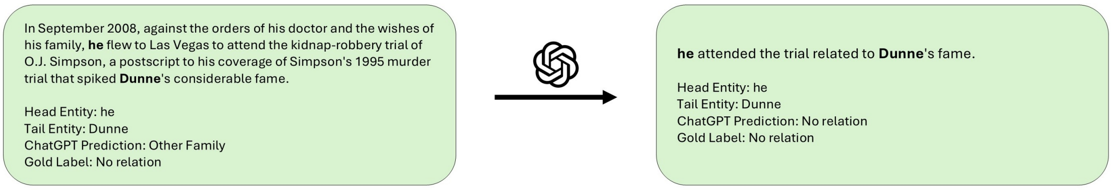
TACRED, TACREV, and RETACRED datasets provide entity types. Therefore, we use entity type constraints to reduce applicable label space. SemEval dataset, however, is focused on common nouns. For SemEval we use our proposed prompt-based method to reduce applicable label space.

SemEval sentences are short. Therefore, we do not use Targeted Summarization for SemEval. For TACRED, TACREV, and RETACRED we use it. However, there are some examples in these datasets where the head and tail entities have identical text in a case-insensitive way (e.g. "He" and "he" in the sentence "He told the Times he no longer is active in the Church of Scientology"). For these instances, we skip the summarization as we thought it could confuse the models.

In our experiments, we set temperature to zero. Our hyperparameters are every element of our prompt, such as the number of synthesized examples, whether to use summarization, whether to use entity tagging, etc. These hyperparameters are selected using a small set equal to 1% of development set. This set contains a few examples per relation. This setting is comparable to using examples in the annotation guidelines as development.

F Prompts

In this section, we present our prompts for Targeted Summarization (Figure 3), Reducing Label Space (Figure 4), Synthesizing Examples (Figure 5), and our final prompt (Figure 6).



Summarize the relation between the words "he" and "Dunne" in the following sentence. In the summary, ignore everything else other than the relation between "he" and "Dunne". Your summary must include the words "he" and "Dunne". In your summary, prefix "he" with tag <ENT0> and suffix it with tag </ENT0>. Also, prefix "Dunne" with tag <ENT1> and suffix it with </ENT1>. Use no more than 10 words for the summary.

Sentence: In September 2008, against the orders of his doctor and the wishes of his family, he flew to Las Vegas to attend the kidnap-robbery trial of O.J. Simpson, a postscript to his coverage of Simpson's 1995 murder trial that spiked Dunne's considerable fame.

Figure 3: Top: Targeted summarization helps relation extraction. ChatGPT predicts the incorrect relation "other family" (a family relation other than immediate family) when the original text is used, but the gold label is "no relation" (top left). The presence of some cues in the sentence such as the word "family" may have confused the model. The prediction of the model on the summarized version, however, is correct (top right). Bottom: Our prompt for summarizing the text supporting the relation between the entities.

In the task of relation extraction, you are given a sentence and a pair of entities in the sentence. The goal is to select the relation between the two entities from a predefined set of candidate relations.

The predefined relations are "Other", "Cause-Effect", "Instrument-Agency", "Product-Producer", ...

The definition of each relation is as follows. Note that in relation examples or relation instances, X and Y are replaced with actual words.

Cause-Effect: ...
Instrument-Agency: ...
Product-Producer: ...
...

If none of the above relations holds between the two entities, we output "Other".

Now, based on the above definitions and restrictions, which of the following relations may hold between "master" and "stick" or between "stick" and "master" in the following sentence? You must select at least 3 of the relations that might hold. Be as permissive as possible when selecting the relations that might hold.

Sentence: The school **master** teaches the lesson with a **stick**.

Options:
A. Other
B. Cause-Effect
C. Instrument-Agency
D. Product-Producer
E. Content-Container
F. Entity-Origin
G. Entity-Destination
H. Component-Whole
I. Member-Collection
J. Message-Topic

List at least "3" of the options that might hold between "master" and "stick" or between "stick" and "master" in the sentence. Be as permissive as possible when selecting the relations that might hold. Format your response as a python list of single letters without any additional things including your notes and explanations.

Definition: Instrument-Agency(X,Y) relation is true of a sentence S that mentions entities X and Y if and only if the situation described in S entails the fact that X is the instrument (tool) of Y or, equivalently, that Y uses X.

Definition – Restrictions:
(a) X is an entity and Y implies an activity or an explicit actor. That is to say, there exists an activity even if the close context for X and Y includes no verb. Examples: "laser/X printer/Y" means "the printer uses laser (for printing)"; "axe/X murderer/Y" means "murderer uses axe for killing".

(b) Both X and Y can be a physical object, an abstract object or an organization.

(c) Y cannot use their (body) parts as instruments. The restriction is meant to prevent overlaps with Component-Whole. If a method, principle, technique exists on its own, independently of Y (vacuum/X cleaner/Y or microwave/X oven/Y), than X is an Instrument used by Y instead of an integral and functional part of Y.

(d) People are not usually classified as Instruments, unless they are clearly non-agentive in the situation.

(e) Properties, capabilities, aptitutes, skills, attitutes etc. are not acceptable as Instruments.

(f) Location can be Instrument but only when the use is the emphasis of the sentence ("People used the trail to
reach California", but not "People travelled on the trail from Kansas to Dallas").

(g) Animals can be used as Instruments.

(h) Means of transport can be Instruments

(i) Raw materials, materials, ingredients, pieces and all the other things that are used to build, assemble, prepare, are acceptable Instruments. So are power sources and external resources used by machine, device, etc. in operating.

(j) Wearing, putting on is accepted as a way of using, on the basis that the wearer Y is generally wearing an item X for some reason (to keep warm, carry things in, protect him/herself, etc).

(k) Selling, buying is accepted as a way of using (for a living, or whatever).

Figure 4: Left: Our prompt that reduces the number of candidate relation types for SemEval from 10 to 3. Parts of the prompt omitted for brevity. Right: The part of annotation guidelines that we selected to use for Instrument-Agency relation in the prompt.

Please give 10 examples of the relation "founded by" between a organization (called head entity) and a person (called tail entity) using a single subject-verb-object structure containing the head entity and the tail entity.

Note that the tail entity is the person, organization, or geopolitical entity that founded the assigned organization.

Prefix the head entity with tag <ENT0> and suffix it with tag </ENT0>. Also, prefix the tail entity with tag <ENT1> and suffix it with tag </ENT1>.
Produce your response as a list of strings in a json list object.

Figure 5: Our prompt for synthesizing example sentences for relations. In this example, the relation is “founded by”.

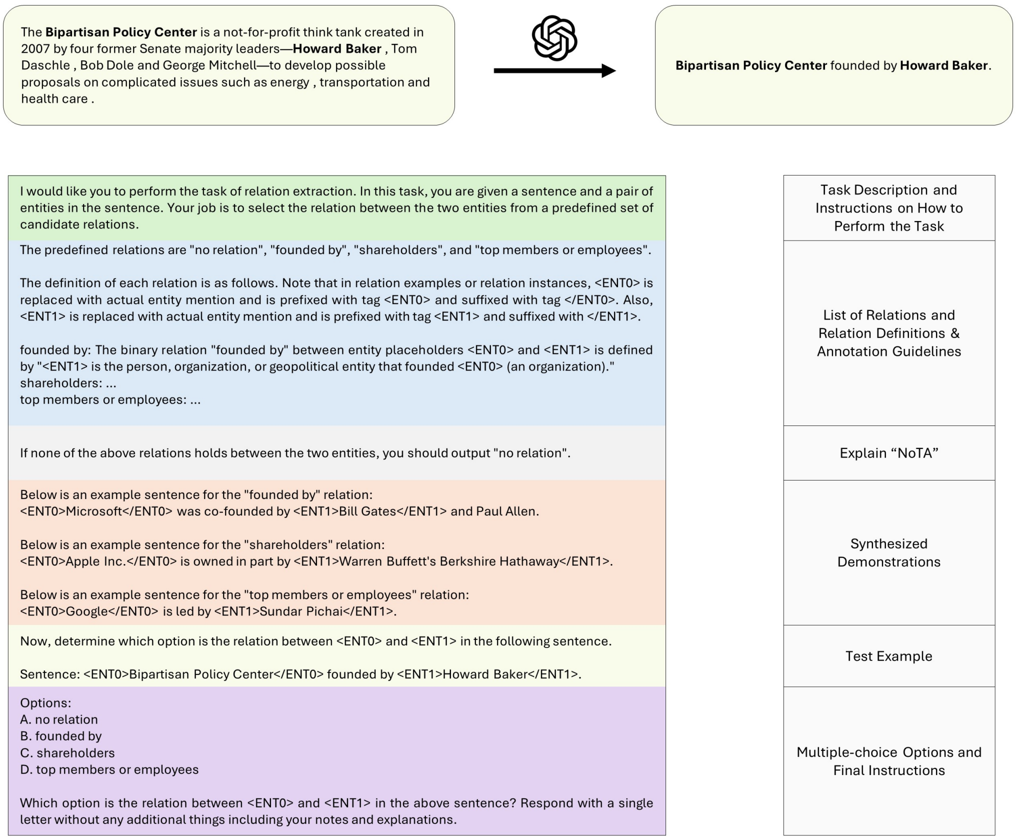


Figure 6: Top: The relation between the head and tail entities in a test example is summarized by an LLM. Bottom: The structure of our final prompt. The prompt uses the summarized example. The example is chosen from the TACRED dataset.