Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors

Kai Zhang **Bernal Jiménez Gutiérrez** **Yu Su**
The Ohio State University
{zhang.13253, jimenezgutierrez.1, su.809}@osu.edu

Abstract

Recent work has shown that fine-tuning large language models (LLMs) on large-scale instruction-following datasets substantially improves their performance on a wide range of NLP tasks, especially in the zero-shot setting. However, even advanced instruction-tuned LLMs still fail to outperform small LMs on relation extraction (RE), a fundamental information extraction task. We hypothesize that instruction-tuning has been unable to elicit strong RE capabilities in LLMs due to RE's low incidence in instruction-tuning datasets, making up less than 1% of all tasks (Wang et al., 2022). To address this limitation, we propose QA4RE, a framework that aligns RE with question answering (QA), a predominant task in instruction-tuning datasets. Comprehensive zero-shot RE experiments over four datasets with two series of instruction-tuned LLMs (six LLMs in total) demonstrate that our QA4RE framework consistently improves LLM performance, strongly verifying our hypothesis and enabling LLMs to outperform strong zero-shot baselines by a large margin. Additionally, we provide thorough experiments and discussions to show the robustness, few-shot effectiveness, and strong transferability of our QA4RE framework. This work illustrates a promising way of adapting LLMs to challenging and underrepresented tasks by aligning these tasks with more common instruction-tuning tasks like QA.[1]

1 Introduction

Large language models (LLMs) (Brown et al., 2020; Chowdhery et al., 2022; Zhang et al., 2022) have been shown to achieve impressive performance on many NLP tasks. Using the in-context learning paradigm, without any parameter updating, LLMs are able to achieve comparable performance with small language models (LMs) fine-tuned on thousands of examples (Liu et al., 2022; Min et al., 2022a; Liang et al., 2022).[2] More recently, fine-tuning LLMs on datasets containing thousands of downstream tasks transformed into an instruction following format (i.e., *instruction-tuning*) has been shown to improve LLMs considerably across the board, especially in zero-shot setting (Iyer et al., 2022; Ouyang et al., 2022; Chung et al., 2022).

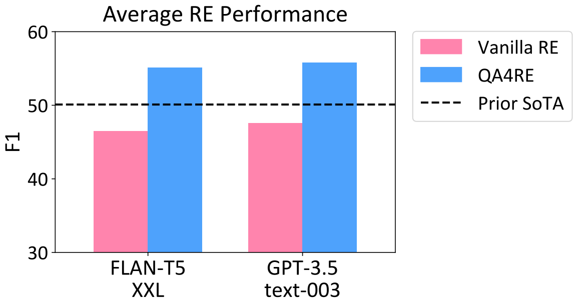


Figure 1: Main finding: Strong instruction-tuned LLMs underperform prior zero-shot RE methods using the standard (vanilla) RE formulation. Our QA4RE framework enables models in two sets of instruction-tuned LLMs (FLAN-T5 and GPT-3.5) to surpass the prior SoTA on 4 RE datasets by a large margin. Results are averaged over 4 RE datasets. We omit the word 'davinci' from the GPT-3.5 model displayed for brevity.

We examine the capability of LLMs in identifying the relationship between entities in a sentence, i.e., relation extraction (RE), a fundamental task in information extraction. Recent work (Jimenez Gutierrez et al., 2022) has found that LLMs underperform fine-tuned small LMs for RE in the biomedical domain. Our results on general domain RE in Fig. 1 reveal that even two of the most advanced instruction-tuned LLMs, FLAN-T5 XXL (Chung et al., 2022) and text-davinci-003 (Ouyang et al., 2022), fail to outperform the state-of-the-art (SoTA) zero-shot RE method based on small LMs (Sainz et al., 2021).

We hypothesize that the limited relation extraction capability of instruction-tuned LLMs could be

[1] Code and data are available at https://github.com/OSU-NLP-Group/QA4RE.

[2] We regard LMs with less than 1B params as small.

Findings of the Association for Computational Linguistics: ACL 2023, pages 794–812
July 9-14, 2023 ©2023 Association for Computational Linguistics

a byproduct of the low incidence of RE tasks in instruction-tuning datasets (Ouyang et al., 2022; Sanh et al., 2022; Chung et al., 2022; Wang et al., 2022).[3] To address the low incidence issue, we propose the QA4RE framework, which aligns RE with multiple-choice question answering (QA), a task that appears much more frequently in most instruction-tuning datasets—around 12-15% of all the tasks in both Wang et al. (2022) and Ouyang et al. (2022). Specifically, by casting the input sentence as a question and possible relation types as multiple-choice options (Fig. 2), LLMs are able to perform RE by selecting the option representing the correct relation type.

Thorough evaluations on four real-world relation extraction datasets and six instruction-tuned models from two different series (OpenAI GPT-3.5 and FLAN-T5 (Chung et al., 2022)) show that QA4RE brings significant gains over the standard RE formulation on, validating its effectiveness and our hypothesis concerning the low incidence of RE. More specifically, our framework enables text-davinci-003 and FLAN-T5-XXLarge to achieve an average of 8.2% and 8.6% absolute improvements in F1, respectively. For the first time, an LLM is able to outperform prior small-LM-based SoTA in the zero-shot setting by a large margin. In-depth analyses further demonstrate the robustness and few-shot effectiveness of QA4RE. More importantly, our framework has been proven to be effectively transferable on instruction-tuned models with various sizes, ranging from 80M to 175B. Our contributions are summarized as follows:

(1) We systematically investigate instruction-tuned LLMs on four real-world relation extraction datasets and note that their limited performance on RE might stem from the low incidence of RE tasks in instruction-tuning datasets.

(2) We reformulate RE as multiple-choice QA in an effort to appropriately leverage QA's much higher prevalence in instruction-tuning datasets and achieve significant improvements on six recent instruction-tuned LLMs, significantly outperforming previous SoTA zero-shot RE methods based on small LM for the first time.

(3) In addition, we demonstrate our QA4RE method's robustness to diverse prompt designs as well as its promising results in the few-shot setting.

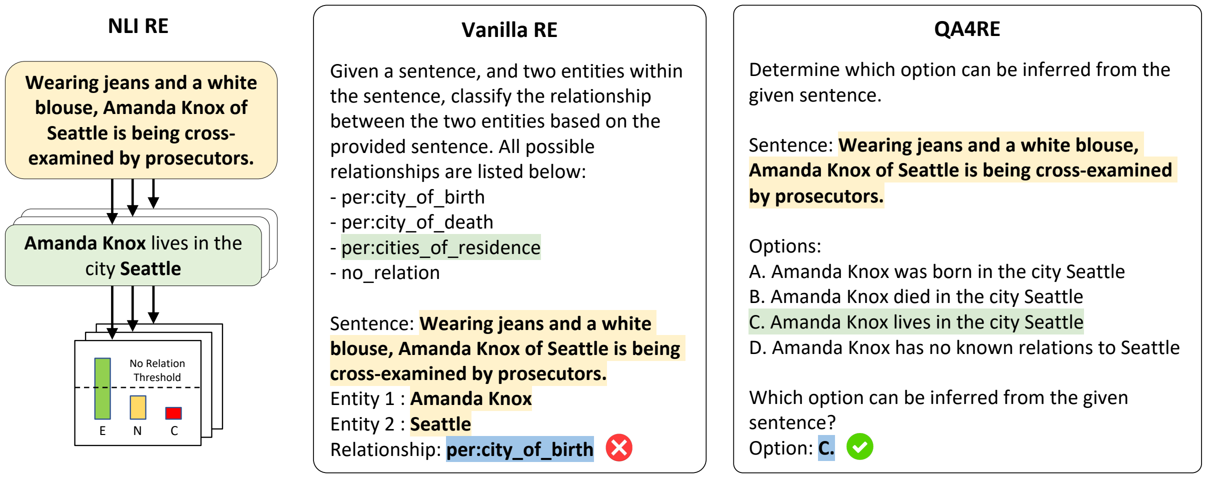
(4) Finally, we show the effectiveness of QA4RE framework is transferable and consistent on various instruction-tuned models with different sizes from 80M to 175B. Our study illustrates the potential of aligning infrequent and challenging tasks with frequent instruction-tuning tasks and can guide others in exploring this direction.

2 Related Work

Instruction Tuning. Large language models originally obtained impressive zero and few-shot performance by leveraging self-supervised next token prediction at massive scales. More recently, supervised fine-tuning on a large number of downstream tasks has been shown to improve LLM accuracy, robustness, fairness, and generalization to unseen tasks (Ouyang et al., 2022; Iyer et al., 2022; Wei et al., 2022a; Chung et al., 2022; Sanh et al., 2022). Several strategies have been developed to align LLMs to human instructions including Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) as well as the more standard language modeling objective, used to fine-tune LLMs on a wide range of tasks reformulated as instruction following tasks (Iyer et al., 2022; Wei et al., 2022a; Chung et al., 2022; Sanh et al., 2022).

Eliciting LLM Abilities. The high cost and increasingly private nature of LLM pre-training make it quite challenging to conclusively determine how different pre-training techniques bring about different LLM capabilities. Many factors involved in pre-training such as simple self-supervised scaling, code or multi-lingual text pre-training (Chowdhery et al., 2022; Chen et al., 2021; Chung et al., 2022) as well as the distinct versions of instruction-tuning mentioned above (Ouyang et al., 2022; Iyer et al., 2022; Wei et al., 2022a; Chung et al., 2022), can interact in a wide variety of ways to unleash the abilities LLMs display. Nonetheless, Fu and Khot (2022) hypothesize that the use of code during pre-training seems to improve an LM's reasoning ability, evidenced by the improved ability to leverage Chain-of-Thought prompting (Wei et al., 2022b) by models trained partially on code such as PaLM (Chowdhery et al., 2022), code-davinci-002 (Chen et al., 2021), and text-davinci-002/003 (Ouyang et al., 2022), compared to text-only models like text-davinci-001 and OPT-175B (Zhang et al., 2022). Additionally, instruction-tuning on a large set of tasks has been shown to improve generalization to unseen tasks, reduce the need for few-shot examples and improve accuracy and robustness

[3] RE-like tasks are <0.5% of the largest available instruction dataset (Wang et al., 2022); see Appendix A for details.


NLI RE
Wearing jeans and a white blouse, Amanda Knox of Seattle is being cross-examined by prosecutors.
Amanda Knox lives in the city Seattle
No Relation Threshold
E
N
C

Vanilla RE

Given a sentence, and two entities within the sentence, classify the relationship between the two entities based on the provided sentence. All possible relationships are listed below:
- per:city_of_birth
- per:city_of_death
- per:cities_of_residence
- no_relation

Sentence: **Wearing jeans and a white blouse, Amanda Knox of Seattle is being cross-examined by prosecutors.**
Entity 1 : **Amanda Knox**
Entity 2 : **Seattle**
Relationship: **per:city_of_birth** ✗

QA4RE

Determine which option can be inferred from the given sentence.

Sentence: **Wearing jeans and a white blouse, Amanda Knox of Seattle is being cross-examined by prosecutors.**

Options:
A. Amanda Knox was born in the city Seattle
B. Amanda Knox died in the city Seattle
C. Amanda Knox lives in the city Seattle
D. Amanda Knox has no known relations to Seattle

Which option can be inferred from the given sentence?
Option: **C.** ✓

Figure 2: This figure shows a schematic of the SoTA NLI zero-shot framework in which each sentence must be compared with each relation template (left), the vanilla formulation for prompting GPT-3 for RE as done in Jimenez Gutierrez et al. (2022) (center) and our multiple-choice QA setting, in which each relation is transformed into a template and GPT-3 is expected to predict only a single letter (right).

across many language tasks (Ouyang et al., 2022; Iyer et al., 2022; Chung et al., 2022).

Low-Resource Relation Extraction. Several reformulations of standard RE have enabled small LMs to achieve fairly strong performance in the zero and few-shot settings. Sainz et al. (2021) utilize small LMs fine-tuned on natural language inference (NLI) datasets to perform zero-shot RE by selecting the entity-filled relation template which is mostly entailed by the test sentence. Lu et al. (2022) frame RE as a summarization task and leverage generative models to summarize the relation between target entities in the sentence. Other low-resource RE methods augment prompt-tuning by using logical rules to create complex prompts from sub-prompts (Han et al., 2022) and injecting knowledge about entity types using learnable virtual tokens (Chen et al., 2022). Our current work uses several relation templates designed in these studies.

LLMs for Relation Extraction. In terms of exploring the RE capabilities of LLMs, most previous work has focused on investigating biomedical RE. Jimenez Gutierrez et al. (2022) report that LLMs underperform standard small LMs fine-tuning in the few-shot setting on a comprehensive set of biomedical RE datasets and show evidence that the poor handling of the none-of-the-above (NoTA) relation category is one of the major culprits. Furthermore, although a few RE-like tasks were included in Super Natural Instruction (Wang et al., 2022), these tasks constitute about 0.5% of the dataset and none of them were selected for model evaluation.

3 Methodology

In this section, we formally define the relation extraction problem and describe our multi-choice QA approach for the problem in detail.

3.1 Problem Statement

Relation extraction (RE) aims to extract the relationship between two given entities based on a specific sentence. More concretely, one relation example contains a sentence S as well as a head entity E_h and a tail entity E_t within S. Given a relation example (S, E_h, E_t), models are required to identify the relation between E_h and E_t expressed in the S from a set of pre-defined relation types.

3.2 Relation Templates

Recent low-resource RE approaches (Sainz et al., 2021; Lu et al., 2022; Han et al., 2022) utilize relation-entailed templates as label verbalization (e.g., "per:city_of_birth" -> "$\{E_h\}$ was born in the city $\{E_t\}$"). As illustrated in Fig. 2 (left), the current SoTA method for low-resource RE (Sainz et al., 2021) utilizes manually constructed relation templates to reformulate the RE task as a natural language inference (NLI) task.

To ensure a fair comparison, we utilize the same templates developed in previous studies (Sainz et al., 2021; Lu et al., 2022) to generate answer options within our QA4RE framework. Furthermore, in Sec. 6.2 we discuss the possibility of directly applying the NLI formulation for RE in LLMs.

3.3 QA4RE Framework

As shown in Fig. 2 (right), we reformulate the relation extraction task as a multi-choice QA problem. By integrating the given head and tail RE entities (E_h and E_t) into the relation templates and using them as multiple-choice options, LLMs are able to leverage extensive QA instruction fine-tuning which has dramatically improved recent models. Additionally, our method allows LLM to generate only an answer index instead of the verbalized relation as in previous work (Jimenez Gutierrez et al., 2022), also shown in Fig. 2 (center).

Type-Constrained Answer Construction. To transform RE into a multiple-choice question, for a given relation example (S, E_h, E_t), we utilize sentence S as the context in standard QA and create options composed of pre-defined templates filled with E_h and E_t entities. To fairly compare with previous work, we apply type constraints (when applicable) to eliminate options for relation types that are not compatible with the entity types of the head and tail entities. For instance, if the type of E_h is PERSON, the relation "org:country_of_headquarters" would be deemed invalid given that a person does not have headquarters.

4 Experiment Setup

4.1 Datasets

We evaluate our methods on four RE datasets: (1) TACRED (Zhang et al., 2017), (2) RETACRED (Stoica et al., 2021), (3) TACREV (Alt et al., 2020), and (4) SemEval 2010 Task 8 (SemEval for brevity) (Hendrickx et al., 2010). Following previous work (Sainz et al., 2021; Lu et al., 2022; Han et al., 2022; Chen et al., 2022), we report the micro averaged F1 with the none-of-the-above relation excluded. To keep OpenAI API costs under control, we randomly sample 1,000 examples from each dataset's test split as our test set.

4.2 Baselines

Zero-Shot. For small LM-based models, we evaluate two low-resource SoTA RE baselines: (1) As shown in Fig. 2 (left), NLI (Sainz et al., 2021) reformulates RE as a natural language inference task and leverages several LMs fine-tuned on the MNLI dataset (Williams et al., 2018): BART-Large (Lewis et al., 2020), RoBERTa-Large (Liu et al., 2019), and DeBERTa-XLarge (He et al., 2021). This method holds the SoTA performance on both zero and few-shot RE. (2) Besides, SuRE (Lu et al., 2022) frames RE as a summarization task and utilizes generative LMs such as BART-Large (Lewis et al., 2020) and PEGASUS-Large (Zhang et al., 2020), achieving competitive results in few-shot and fully-supervised settings.

For the NLI approach (Sainz et al., 2021), we report performance using their own templates on TACRED and TACREV. As this method does not have templates for RETACRED and SemEval, we use the templates from the follow-up work, SuRE (Lu et al., 2022), on these two datasets instead. All the zero-shot methods, including those on LLMs, apply entity type constraints to reduce the relation label space. Since SemEval does not provide entity types, the above methods use all possible relations in every instance as the label space.

Few-Shot. Though our main experiments focus on zero-shot RE, we further explore our method's capabilities by comparing their few-shot performance against several competitive small LM-based methods on the TACRED dataset.

The NLI baseline can be easily extended to the few-shot setting.[4] Furthermore, we add (1) standard Fine-Tuning (Jimenez Gutierrez et al., 2022), (2) PTR (Han et al., 2022) using prompt-tuning with logical rules, and (3) KnowPrompt (Chen et al., 2022) using entity type knowledge via learning virtual tokens, all of which are initialized with RoBERTa-Large (Liu et al., 2019). For hyperparameter details, please refer to Appendix B.1.

4.3 QA4RE Implementation Details

Our QA4RE framework utilizes the same templates and type constraints developed by prior work (Sainz et al., 2021; Lu et al., 2022). In particular, we use SuRE (Lu et al., 2022) templates for our QA4RE approach on all 4 datasets since NLI (Sainz et al., 2021) templates were only designed for TACRED. For prompt engineering, we explore prompt formats and task instructions for vanilla RE and QA4RE in pilot experiments, using text-davinci-002 on a 250-example subset of the TACRED dev set. We then use the same task instructions and prompt format for all four datasets and LLMs. Please refer to Appendix B.2 and B.3 for prompt format and relation verbalization template details, respectively.

[4] SuRE can also be extended to the few-shot setting but we were unable to replicate their results with the code provided.

Methods		TACRED			RETACRED			TACREV			SemEval			Avg.
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	F1
Baselines														
NLI$_{\text{BART}}$		42.6	65.0	51.4	59.5	34.9	44.0	44.0	74.6	55.3	21.6	23.7	22.6	43.3
NLI$_{\text{RoBERTa}}$		37.1	76.9	50.1	52.3	67.0	58.7	37.1	83.6	51.4	17.6	20.9	19.1	44.8
NLI$_{\text{DeBERTa}}$		42.9	76.9	55.1	71.7	58.3	64.3	43.3	84.6	57.2	22.0	25.7	23.7	50.1
SuRE$_{\text{BART}}$		13.1	45.7	20.4	17.9	34.6	23.6	14.1	52.3	22.2	0.0	0.0	0.0	16.5
SuRE$_{\text{PEGASUS}}$		13.8	51.7	21.8	16.6	34.6	22.4	13.5	54.1	21.6	0.0	0.0	0.0	16.4
GPT-3.5 Series														
ChatGPT	Vanilla	32.1	74.8	44.9	45.4	61.3	52.1	30.3	79.6	43.9	18.2	20.8	19.4	40.1
	QA4RE	32.8	68.0	44.2 (−0.7)	48.3	76.8	59.3 (+7.2)	34.7	79.1	48.2 (+4.3)	29.9	35.2	32.3 (+12.9)	46.0 (+5.9)
code-002	Vanilla	27.2	70.1	39.2	42.7	70.4	53.1	27.5	77.7	40.6	27.2	25.6	26.4	39.8
	QA4RE	37.7	65.4	47.8 (+8.6)	48.0	74.0	58.2 (+5.1)	31.7	65.5	42.7 (+2.1)	25.2	29.2	27.0 (+0.6)	43.9 (+4.1)
text-002	Vanilla	31.2	73.1	43.7	44.1	76.3	55.9	30.2	76.8	43.3	31.4	28.8	30.1	43.2
	QA4RE	35.6	68.4	46.8 (+3.1)	46.4	72.4	56.5 (+0.6)	35.7	76.8	48.8 (+5.4)	29.4	34.3	31.6 (+1.5)	45.9 (+2.7)
text-003	Vanilla	36.9	68.8	48.1	49.7	62.2	55.3	38.2	76.8	51.0	33.2	39.3	36.0	47.6
	QA4RE	47.7	78.6	**59.4** (+11.3)	56.2	67.2	61.2 (+5.9)	46.0	83.6	**59.4** (+8.4)	41.7	45.0	43.3 (+7.3)	**55.8** (+8.2)
FLAN-T5 Series														
XLarge	Vanilla	51.6	49.1	50.3	54.3	40.3	46.3	56.0	59.1	57.5	35.6	29.8	32.4	46.6
	QA4RE	40.0	78.2	53.0 (+2.7)	57.1	79.7	66.5 (+20.2)	40.7	85.9	55.3 (−2.2)	45.1	40.1	42.5 (+10.1)	54.3 (+7.7)
XXLarge	Vanilla	52.1	47.9	49.9	56.6	54.0	55.2	52.6	50.9	51.7	29.6	28.8	29.2	46.5
	QA4RE	40.6	82.9	54.5 (+4.6)	56.6	82.9	**67.3** (+12.1)	39.6	86.4	54.3 (+2.6)	41.0	47.8	**44.1** (+14.9)	55.1 (+8.6)

Table 1: Experimental results on four RE datasets (%). We omit the 'davinci' within the names of GPT-3.5 Series LLMs and ChatGPT refers to gpt-3.5-turbo-0301. We mark the best results in **bold**, the second-best underlined, and F1 improvement of our QA4RE over vanilla RE in green.

To systematically compare our QA4RE framework with the vanilla RE formulation, we evaluate them on two series of LLMs, resulting in seven models in total. In GPT-3.5 series LLMs, for LLMs accessible via Text Completion API (code-davinci-002, text-davinci-002, and text-davinci-003), we follow previous work (Jimenez Gutierrez et al., 2022) and use the logit bias option to constrain token generation to relation labels for vanilla RE and option indices for QA4RE. Due to the fewer available control options for LLMs in Chat Completion API (gpt-3.5-turbo-0301), we only set the temperature as 0 and use the default system prompt.

We also examine open-sourced FLAN-T5 series LLMs (Chung et al., 2022) that are trained on a mixture of tasks (Sanh et al., 2022; Wei et al., 2022a; Wang et al., 2022). The 1,836 tasks utilized in training include less than 0.5% of RE-similar tasks, making FLAN-T5 series LLMs the ideal models for verifying our hypothesis. Specifically, we use XLarge (3B) and XXLarge (11B) models and adopt the same prompts and greedy decoding strategy as GPT-3.5 series LLMs to ensure a fair comparison.

5 Results

5.1 Zero-Shot Results

Our main experimental results on four relation extraction datasets can be found in Tab. 1. We have the following observations from our results:

(1) By reformulating RE as QA, our framework improves upon the vanilla RE formulation on all the LLMs and most datasets, making them much stronger zero-shot relation extractors. In particular, text-davinci-003 and FLAN-T5 XL and XXL are able to outperform the prior SoTA, NLI$_{\text{DeBERTa}}$, by a large margin. One thing worth noting is that QA4RE brings the largest gain on the best LLM in each series (text-davinci-003 and FLAN-T5 XXL), showing that stronger LLMs may benefit more from our framework.

(2) The two FLAN-T5 LLMs in Tab. 1 benefit significantly from our QA4RE framework. Moreover, consistent and substantial improvements can also be observed in other FLAN-T5 models and the full test set, as discussed in Sec. 6.3 and Appendix C. Considering that relation extraction tasks account for less than 0.5% of the instruction tasks used to train FLAN-T5 models, these findings strongly support our hypothesis that aligning underrepresented tasks with more common instruction-tuning tasks, such as QA, unlocks LLMs' ability to solve low-frequency tasks.

(3) The SemEval dataset poses a significant challenge for all baselines given its lack of type-constraints, particularly for SuRE (Lu et al., 2022). With such a large search space, generative LMs without fine-tuning tend to summarize all examples into NoTA relation, resulting in its systematic failure. It should be noted that without type constraints, the RE problem becomes a 19-choice question answering task in our QA4RE framework. Despite this, our method still demonstrates substantial improvements for LLMs, particularly for text-davinci-003 and FLAN-T5 XXL.

5.2 Robustness to Verbalization Templates

For our experiments, we utilize manually written relation templates from previous work (Sainz et al., 2021; Lu et al., 2022). However, Lu et al. (2022) note that model performance may vary significantly with template design. Thus, to investigate the robustness of models to different templates, thorough experiments are conducted with four different templates, described in detail in Appendix B.3, across all zero-shot methods on the TACRED dataset. Tab. 2 shows results comparing these four templates on all methods used in our main experiments, including vanilla RE as a template-free reference.

<table>
<tr><th colspan="2">Methods</th><th>TEMP1</th><th>TEMP2</th><th>TEMP3</th><th>TEMP4</th></tr>
<tr><td colspan="2">NLI_{BART}</td><td>51.4</td><td>49.7</td><td>4.4</td><td>42.0</td></tr>
<tr><td colspan="2">$\text{NLI}_{\text{RoBERTa}}$</td><td>50.1</td><td>47.1</td><td>19.6</td><td>35.8</td></tr>
<tr><td colspan="2">$\text{NLI}_{\text{DeBERTa}}$</td><td>55.0</td><td>49.4</td><td>17.1</td><td>36.6</td></tr>
<tr><td colspan="2">$\text{SuRE}_{\text{BART}}$</td><td>19.9</td><td>20.4</td><td>2.1</td><td>10.1</td></tr>
<tr><td colspan="2">$\text{SuRE}_{\text{PEGASUS}}$</td><td>20.5</td><td>21.8</td><td>6.2</td><td>19.3</td></tr>
<tr><td rowspan="2">text-003</td><td>Vanilla</td><td colspan="4">48.1</td></tr>
<tr><td>QA4RE</td><td>**56.6**</td><td>**59.4**</td><td>**48.7**</td><td>**50.1**</td></tr>
</table>

Table 2: F1 score on TACRED with four templates (%). The best result using each template is marked in bold. text-003 refers to text-davinci-003.

From Tab. 2, we observe the following:
(1) Our method consistently outperforms small LM baselines and the vanilla RE framework, regardless of the template. It is worth noting that even with templates that are constructed with label name information only and no expert knowledge (TEMP3 and TEMP4), our QA framework still performs better than vanilla RE, indicating the effectiveness and consistency of our QA framework.
(2) NLI and SuRE performance is largely template dependent. When using carefully crafted high-quality templates (TEMP1 and TEMP2), several LM-based NLI methods outperform text-davinci-003 with vanilla RE. However, when equipped with templates created without expert knowledge (TEMP3 and TEMP4), the performance of both NLI and SuRE deteriorates dramatically. In contrast, QA4RE is more robust to variation in verbalization templates, reducing trial-and-error development efforts as well as making it more readily transferred to settings where obtaining quality templates may be limited due to the high cost of expert annotations, such as the biomedical or financial domains.

5.3 None-of-the-Above Relation Evaluation

The none-of-the-above (NoTA) relation (Gao et al., 2019; Sabo et al., 2021; Jimenez Gutierrez et al., 2022) is defined as the case where no relation of interest exists between the given entities. Jimenez Gutierrez et al. (2022) demonstrate that the earlier inferior performance of LLMs on RE tasks can be largely attributed to their inability to handle the NoTA relation. To evaluate the efficacy of zero-shot methods on NoTA relation, following previous work (Fei and Liu, 2016; Shu et al., 2017; Sainz et al., 2021), we apply NoTA-included macro F1 metric as well as micro and macro P vs. N (all positive relations vs. NoTA relation as binary classification) F1 metrics.

<table>
<tr><th colspan="2">Methods</th><th>Macro F1</th><th>Micro P vs. N</th><th>Macro P vs. N</th></tr>
<tr><td colspan="2">NLI_{BART}</td><td>49.8</td><td>75.9</td><td>71.1</td></tr>
<tr><td colspan="2">$\text{NLI}_{\text{RoBERTa}}$</td><td>43.7</td><td>68.5</td><td>65.8</td></tr>
<tr><td colspan="2">$\text{NLI}_{\text{DeBERTa}}$</td><td>55.0</td><td>75.6</td><td>72.3</td></tr>
<tr><td colspan="2">$\text{SuRE}_{\text{BART}}$</td><td>15.5</td><td>35.2</td><td>35.0</td></tr>
<tr><td colspan="2">$\text{SuRE}_{\text{PEGASUS}}$</td><td>14.9</td><td>32.4</td><td>31.5</td></tr>
<tr><td rowspan="2">text-003</td><td>Vanilla</td><td>45.3</td><td>72.8</td><td>69.5</td></tr>
<tr><td>QA4RE</td><td>**58.9**</td><td>**78.4**</td><td>**74.8**</td></tr>
</table>

Table 3: NoTA-included 42-class macro F1 as well as macro and micro P vs. N (all positive relations vs. NoTA) F1 on TACRED (%). The best result of each metric is bolded. text-003 refers to text-davinci-003. Ma and Mi are short for macro and micro, respectively.

From Tab. 3, we observe that, when enhanced by our QA framework, text-davinci-003 achieves significant improvement in NoTA-included metrics, outperforming the small LM-based NLI methods. This further demonstrates the effectiveness of our framework, even in handling the challenging NoTA relation. It is worth noting that these superior results are achieved by simply adding an entity-filled NoTA relation template as an answer option for QA, without the additional thresholding requirements of previous methods (Sainz et al., 2021; Lu et al., 2022). This eliminates the need for additional hyperparameter searching, which can be tricky for low-resource settings.

5.4 Few-Shot Results

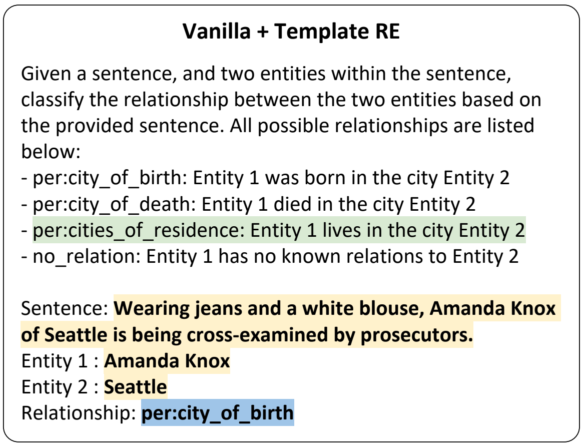
While zero-shot RE is our main focus, we also evaluate our method under the few-shot setting. Results are shown in Tab. 4. Due to budget limitations, we restrict our case study to the 4-shot scenario (i.e., 4 labeled examples per relation) with the best-performing LLM in the zero-shot setting (text-davinci-003). After determining the optimal number of in-context examples searched on the dev set, we randomly select the examples with the same entity type constraints from the given train set.

Interestingly, vanilla RE is unable to obtain any improvement from labeled examples, suggesting that it is also limited in the few-shot setting. The limited performance shown by vanilla RE indicates that few-shot demonstrations might bias the model towards incorrect relations in the context rather than helping it perform the task more accurately.

Methods	K=0	K=4	K=8	K=16	K=32
Fine-Tuning	-	9.0	21.2	29.3	33.9
PTR	-	26.8	30.0	32.9	36.8
KnowPrompt	-	30.2	33.7	34.9	35.0
$\text{NLI}_{\text{DeBERTa}}$-TEMP1	55.0	**64.2**	**64.7**	**58.7**	**65.7**
$\text{NLI}_{\text{DeBERTa}}$-TEMP2	49.4	**51.2**	47.3	**50.5**	48.1
Vanilla	48.1	46.2	-		
QA4RE	59.4	**62.0**	-		

Table 4: Few-shot F1 on TACRED (%). All results are averaged over 3 different training subsets for each K. We use text-davinci-003 for vanilla RE and QA4RE. For the best-performing baseline (NLI) as well as vanilla RE and QA4RE, we mark the results in **bold** when they are improved over their zero-shot alternatives.

Even employing our QA4RE framework, the few-shot text-davinci-003 does not outperform the DeBERTa-based NLI method (Sainz et al., 2021) when using their own templates (TEMP1). However, fine-tuning the NLI model on RE data can be brittle even with careful hyperparameter tuning, as evidenced by the unstable gains seen as more data is added for both TEMP1 and TEMP2. Furthermore, we find that few-shot NLI results when using TEMP2 drop substantially from TEMP1, suggesting that this approach also lacks robustness to templates in the few-shot setting. Thus, considering that our QA approach enables LLMs to obtain few-shot improvements over zero-shot results using random in-context learning example selection, obtains only around 2% lower performance than the best NLI model, and is robust to different template designs, our approach is competitive on few-shot RE and has the potential to achieve even stronger performance with more exploration. We leave further investigation on how to improve LLMs for few-shot RE to future work.



Vanilla + Template RE

Given a sentence, and two entities within the sentence, classify the relationship between the two entities based on the provided sentence. All possible relationships are listed below:
- per:city_of_birth: Entity 1 was born in the city Entity 2
- per:city_of_death: Entity 1 died in the city Entity 2
- per:cities_of_residence: Entity 1 lives in the city Entity 2
- no_relation: Entity 1 has no known relations to Entity 2

Sentence: **Wearing jeans and a white blouse, Amanda Knox of Seattle is being cross-examined by prosecutors.**
Entity 1 : **Amanda Knox**
Entity 2 : **Seattle**
Relationship: **per:city_of_birth**

Figure 3: The same example and templates as Fig. 2 but using templates for relation explanations.

6 Discussions

6.1 Are Relation Templates All LLMs Need?

We conduct an ablation study to better understand how relation templates contribute to the performance improvement obtained by QA4RE. As illustrated in Fig. 3, we fill the relation verbalization templates with markers *Entity 1* and *Entity 2* as relation explanations, thereby presenting the expert knowledge from the templates to the LLM. Using the same templates and type constraints, we compare this framework (termed Vanilla+TEMP) with vanilla RE and QA4RE on the TACRED dataset and GPT-3.5 series LLMs.

As shown in Tab. 5, introducing relation explanations using the same templates does not result in consistent or significant performance improvement. In fact, adding extra information to the task instruction might make it more challenging for the LLM to understand the task. In contrast, using our QA4RE framework, we do not need to separately specify the entities of interest or relation explanations; they are both naturally embedded in the answer options. These ablation results show that the gains from QA4RE mainly come from the QA reformulation, not simply from the relation explanations/templates.

6.2 QA4RE vs. NLI4RE

Given the strong performance obtained by small LMs using the NLI reformulation of RE, we leverage this same formulation (Sainz et al., 2021) for LLMs (termed NLI4RE).[5] More concretely, for each example, we use the LLM to predict whether

[5] We follow the NLI format from ANLI (Wang et al., 2022).

Methods		P	R	F1	ΔF1
code-002	Vanilla	27.2	70.1	39.2	-
	Vanilla + TEMP	27.5	71.8	39.7	+0.5
	QA4RE	37.7	65.4	47.8	+8.6
text-002	Vanilla	31.2	73.1	43.7	-
	Vanilla + TEMP	26.8	77.8	39.8	−3.9
	QA4RE	35.6	68.4	46.8	+3.1
text-003	Vanilla	36.9	68.8	48.1	-
	Vanilla + TEMP	36.9	76.5	49.8	+1.7
	QA4RE	**47.7**	**78.6**	**59.4**	+11.3

Table 5: Evaluation on TACRED regarding whether incorporating relation explanations based on the same templates into vanilla RE bridges its gap to QA4RE (%).

the given sentence (the premise) entails each answer option from the QA4RE formulation (the hypothesis). We allow the LLM to generate *entailment*, *neutral*, or *contradiction* for each sentence-relation pair. If the maximum probability of entailment among all possible positive relations is below the threshold of 0.5, the example will be classified as NoTA, as done by Sainz et al. (2021).

Formulation	RED	RERED	REV	Eval	Avg.
Vanilla	48.1	55.3	51.0	36.0	47.6
NLI4RE	41.7	36.8	39.2	22.4	35.0
QA4RE	**59.4**	**61.2**	**59.4**	**43.3**	**55.8**

Table 6: F1 of text-davinci-003 with different task formulations (%). RED, RERED, REV, and Eval are short for TACRED, RETACRED, TACREV, and SemEval datasets, respectively.

As shown in Tab. 6, when using the NLI formulation, text-davinci-003 surprisingly underperforms the vanilla RE formulation. The reason for its poor performance is two-fold: (1) The heuristically predefined threshold 0.5 is not ideal for LLMs and thus many positive predictions are classified as NoTA. However, it is also difficult to find a good threshold under the zero-shot setting. (2) Under NLI4RE, unlike vanilla RE or QA4RE, an LLM is not seeing the full relation space but assigning probabilities to each candidate hypothesis individually. The final prediction is thus more sensitive to the LLM's bias over different relations.

NLI4RE also requires multiple inference runs for each relation example to evaluate all the candidate relations, incurring a significantly higher cost.

6.3 QA4RE & Model Size

To verify the effectiveness and transferability of our QA4RE framework on smaller instruction-tuned models, we further evaluate the FLAN-T5 Small (80M), Base (250M), and Large (780M) on the full test set over four RE datasets. Tab. 7 shows our QA4RE framework can still bring considerable gains to instruction-tuned models with various sizes, even for the smallest one (80M). This demonstrates the effectiveness of QA4RE is transferable across various model sizes from 80M to 175B, considering the consistent improvements of QA4RE on several GPT-3.5 models.

LMs	Model Size	Avg. F1		
		Vanilla	QA4RE	Δ
GPT-3.5 Series				
text-001	175B	22.3	14.9	−7.4
code-002	175B	39.8	43.9	+4.1
text-002	175B	43.2	45.9	+2.7
text-003	175B	47.6	55.8	+8.2
FLAN-T5 Series				
Small	80M	19.5	25.0	+5.6
Base	250M	22.3	26.4	+4.2
Large	780M	34.8	41.8	+7.0
XLarge	3B	46.6	54.3	+7.7
XXLarge	11B	46.5	55.1	+8.6

Table 7: Effectiveness of QA4RE on both the GPT-3.5 series and FLAN-T5 with different sizes. The results are averaged over four RE datasets.

In the FLAN-T5 series, larger models benefit more from our framework. However, we note that this trend does not continue when scaling up to much larger GPT-3.5 models. In fact, all GPT-3.5 models except for text-davinci-003 benefit less from QA4RE than FLAN-T5 models. The smaller improvements of QA4RE on these models make their overall RE performance only comparable with models that are approximately 20 and 50 times smaller. This indicates that the wide variety of alignment strategies used by the GPT-3.5 series models discussed in Sec. 2 might not be universally more effective than standard instruction-tuning for improving model generalization on low-incidence tasks even when aligned to high incidence ones. Nevertheless, the strong improvement observed in the strongest models tested, text-davinci-003 and FLAN-T5-XXL, demonstrates the potential for QA4RE's effectiveness to continue as models become even more capable in the future.

7 Conclusions and Future Work

In this work, we first show that even the most recent instruction-tuned LLMs underperform fine-tuned small LMs on the relation extraction (RE) task. To address this limitation, we reformulate RE into multiple-choice question answering (QA) with the purpose of leveraging a task that is widely covered in instruction-tuning datasets like QA, instead of RE, which is barely present in these datasets. Comprehensive experiments demonstrate that our QA4RE framework unlocks the power of LLMs as zero-shot relation extractors, especially for two recent LLMs (text-davinci-003 and FLAN-T5 XXL). We also conduct thorough experiments to explore the robustness and few-shot effectiveness of our method as well as study in what LLM training scenarios it is most effective.

In future work, we hope to explore additional underrepresented tasks in instruction-tuning that might be challenging for LLMs and could be successfully aligned with more widely adopted instruction-tuning tasks like QA. Additionally, we plan to continue exploring this line of work by leveraging our QA4RE framework for other LLMs such as the OPT-series (Zhang et al., 2022; Iyer et al., 2022) and PaLM (Chowdhery et al., 2022), which are not included in this work due to the limited computational resources and/or access.

8 Limitations

Even though our method helps unleash the power of six recent strong LLMs as zero-shot relation extractors, earlier LLMs without strong instruction tuning such as text-davinci-001 saw no improvements from our framework. Additionally, although we carry out comprehensive experiments on the zero-shot RE setting, our few-shot exploration is more limited. It is still unclear from our investigation whether including even more training examples can improve LLM's RE performance and to what extent the same trends seen across GPT-3 models in the zero-shot setting hold steady in the few-shot setting. We leave answering these questions for future work.

9 Ethics Statement

In this work, we propose a method to improve LLM performance on the important and fundamental task of relation extraction. We do not anticipate any ethical issues regarding the topics of this research.

Acknowledgements

The authors would like to thank Renze Lou, colleagues from the OSU NLP group, and the anonymous reviewers for their valuable feedback. The authors would also like to thank Keming Lu for discussions and guidance on reproducing SuRE. This research was supported in part by NSF OAC 2112606, NIH R01LM014199, and Ohio Supercomputer Center (Center, 1987).

References

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ohio Supercomputer Center. 1987. Ohio supercomputer center.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2778–2788. ACM.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Geli Fei and Bing Liu. 2016. Breaking the closed world assumption in text classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 506–514. The Association for Computational Linguistics.

Hao Fu, Yao; Peng and Tushar Khot. 2022. How does gpt obtain its ability? tracing emergent abilities of language models to their sources. *Yao Fu's Notion*.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 6249–6254. Association for Computational Linguistics.

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

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2022. OPT-IML: scaling language model instruction meta learning through the lens of generalization. *CoRR*, abs/2212.12017.

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 EMNLP*, pages 4497–4512, Abu Dhabi, United Arab Emirates. 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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher R'e, Diana Acosta-Navas, Drew A. Hudson, E. Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan S. Kim, Neel Guha, Niladri S. Chatterji, O. Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas F. Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models. *ArXiv*, abs/2211.09110.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022, Dublin, Ireland and Online, May 27, 2022*, pages 100–114. 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. *CoRR*, abs/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 EMNLP*, pages 6575–6594, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022a. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5316–5330. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022b. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True Few-Shot Learning with Language Models.

Ofer Sabo, Yanai Elazar, Yoav Goldberg, and Ido Dagan. 2021. Revisiting few-shot relation classification: Evaluation data and classification schemes. *Trans. Assoc. Comput. Linguistics*, 9:691–706.

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 EMNLP*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Lei Shu, Hu Xu, and Bing Liu. 2017. DOC: deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2911–2916. Association for Computational Linguistics.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the TACRED dataset. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13843–13850. AAAI Press.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

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, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 35–45. Association for Computational Linguistics.

A Instruction Dataset Portion

	#Tasks	%RE	%QA
T0 (Sanh et al., 2022)	62	0	27.4
FLAN (Wei et al., 2022a)	62	0	21
MetaICL (Min et al., 2022b)	142	0	28.9
NaturalInstruct (Wang et al., 2022)	1731	<0.5	>12

Table 9: Popular instruction tuning datasets and proportion of RE and QA tasks in each.

As shown in Tab. 9, there is no RE task in T0 (Sanh et al., 2022), FLAN (Wei et al., 2022a), and MetaICL (Min et al., 2022b) instruction tuning datasets. Even in the largest available NaturalInstruct (Wang et al., 2022), RE tasks consist of only less than 0.5% of the total tasks. By contrast, QA is the most popular task format in all instruction tuning datasets. These observations indicate the low incidence of RE tasks and the dominance of QA tasks in datasets used for instruction tuning.

B Experimental Details

B.1 Hyperparameters for Few-Shot Methods

In the few-shot setting, for each K, we randomly sample 3 times to obtain different training subsets, each of which will be used as in-context demonstrations for LLMs or used to train the small language models in baselines. Report results are averaged over the three subsets. To avoid over-estimating few-shot performance with too many dev examples (Perez et al., 2021), we use 100 randomly selected examples of dev set for all the hyperparameter searching.

For LLMs, we use the dev set to search for the optimal number of in-context examples as a hyperparameter from $\{1, 2, 5\}$. Then we randomly select the same type-constrained in-context examples from the given train set.

For all small LM-based baselines, we use their publicly available code and hyper-parameters for training. According to the original papers of NLI (Sainz et al., 2021) and SuRE (Lu et al., 2022), we use the checkpoints available online and hyperparameters reported for model training. Unfortunately, we were unable to reproduce SuRE results with default hyperparameters. For standard Fine-Tuning (Jimenez Gutierrez et al., 2022), PTR (Han et al., 2022), and KnowPrompt (Chen et al., 2022), we perform a grid search over hyperparameters on dev with the range shown in Tab. 10.

We use 8 NVIDIA GeForce RTX 2080 Ti and 2 NVIDIA RTX A6000 to conduct all the experiments. The total GPU hours used and the cost for OpenAI API are listed in Tab. 11.

Hyperparameter	**Search Space**
Learning Rate 1:	$\{1e{-}5, 3e{-}5\}$
Weight Decay:	$\{0.01, 0.001\}$
Learning Rate 2:	$\{5e{-}5, 2e{-}4\}$

Table 10: Hyperparameters used for grid search of few-shot methods. Learning Rate 2 is used for training new tokens in PTR (Han et al., 2022) and virtual tokens in KnowPrompt (Chen et al., 2022).

	Num of Params (Millions)	**Total GPU Hours**	**Total Cost**
RoBERTa-Large	354	284	-
DeBERTa-XLarge	900	14	-
BART-Large	406	2	-
Pegasus-Large	568	50	-
FLAN-T5 S	80	<1	-
FLAN-T5 M	250	<1	-
FLAN-T5 L	780	1	-
FLAN-T5 XL	3,000	2	-
FLAN-T5 XXL	11,000	4	-
OpenAI Text API	175,000	-	$835
OpenAI Chat API	?	-	$4

Table 11: Total GPU Hours for open sources LMs and cost for using OpenAI API (all version included).

B.2 Prompts for LLMs

As shown in Tab. 12, we list all templates used in this paper including vanilla + TEMP in Tab. 5, NLI4RE in Tab. 6, and vanilla as well as QA4RE in all experiments.

B.3 Relation Verbalization Templates

In the relation verbalization template robustness experiment shown in Tab. 2, the differences between four templates are described below using the *org:top_members/employees* relation from TACRED benchmark as an example:

1. Concrete Examples: *$\{E_h\}$ is a chairman/ president/director of $\{E_t\}$*

2. Semantic Relationship: *$\{E_h\}$ is a high level member of $\{E_t\}$*

3. Straightforward: *The relation between $\{E_h\}$ and $\{E_t\}$ is top members or employees*

4. Word Translation: *$\{E_h\}$ organization top members or employees $\{E_t\}$*

Methods		TACRED			RETACRED			TACREV			SemEval			Avg.
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	F1
Small	Vanilla	9.5	40.9	15.4	22.8	50.2	31.3	9.1	41.9	15.0	10.0	11.8	10.8	18.1
	QA4RE	13.8	52.2	21.8 (+6.4)	33.5	66.2	44.5 (+13.2)	13.7	55.2	22.0 (+7.0)	5.9	7.1	6.4 (−4.4)	23.7 (+5.6)
Base	Vanilla	14.1	31.1	19.4	21.1	26.8	23.6	14.1	33.3	19.8	14.9	17.9	16.2	19.8
	QA4RE	17.1	54.7	26.0 (+6.6)	33.0	65.2	43.8 (+20.2)	17.2	58.5	26.6 (+6.8)	6.7	8.0	7.3 (−8.9)	25.9 (+6.2)
Large	Vanilla	22.8	58.6	32.8	37.5	60.8	46.4	22.6	61.9	33.1	23.7	19.7	21.5	33.5
	QA4RE	30.3	78.5	43.7 (+10.9)	44.5	72.6	55.2 (+8.8)	29.9	82.4	43.9 (+10.8)	24.8	15.8	19.3 (−2.2)	40.5 (+7.1)
XLarge	Vanilla	48.8	49.0	48.9	55.8	39.8	46.4	52.0	55.7	**53.8**	34.9	29.6	32.0	45.3
	QA4RE	37.6	78.6	50.9 (+2.0)	56.2	79.9	66.0 (+19.6)	38.2	84.7	52.7 (−1.1)	44.4	39.9	42.1 (+10.1)	52.9 (+7.7)
XXLarge	Vanilla	48.2	45.3	46.7	56.1	53.7	54.9	50.6	50.6	50.6	29.2	28.1	28.6	45.2
	QA4RE	38.1	82.9	**52.2** (+5.5)	55.9	82.0	**66.5** (+11.6)	38.3	88.1	53.4 (+2.8)	40.2	47.5	**43.5** (+14.9)	**53.9** (+8.7)

Table 8: FLAN-T5 results on full test set of four RE datasets (%). We mark the best results in **bold**, the second-best underlined, and F1 improvement of our QA4RE over vanilla RE in green.

The first set of templates was written by Sainz et al. (2021), while the remaining three were explored by Lu et al. (2022). We use the templates from their official GitHub repositories.[6] In addition, we further list relation verbalization templates used by all LLMs in our paper in Tab. 13, Tab. 14, and Tab. 15.

C Full Test Results on FLAN-T5

We present the full test set results of all four RE datasets in Tab. 8. Our observations align with the findings from experiments on 1,000 test examples:
(1) Our QA4RE framework can bring consistent and significant improvements over all FLAN-T5 series models on the averaged results. Additionally, larger models benefit more from our framework. These two signals strongly demonstrate the effectiveness of QA4RE.
(2) We notice that our QA4RE does not improve smaller versions of FLAN-T5 on SemEval, a 19-choice QA task. This may be due that these models have difficulties in understanding long input fed by QA4RE.

[6] Templates for Robustness Experiments:
TEMP1: https://github.com/osainz59/Ask2Transformers/blob/master/resources/predefined_configs/tacred.relation.config.json
TEMP3: https://github.com/luka-group/SuRE/blob/main/data templates/tacred/rel2temp_forward.json
TEMP4: https://github.com/luka-group/SuRE/blob/main/data /templates/tacred/rel2temp_raw_relation.json

Formulations	Prompts
Vanilla RE	Given a sentence, and two entities within the sentence, classify the relationship between the two entities based on the provided sentence. All possible Relationships are listed below: - [Possible Relation 1] - [Possible Relation 2] - [NoTA Relation] Sentence: [Sentence S] Entity 1: [Head Entity E_h] Entity 2: [Tail Entity E_t] Relationship:
Vanilla + TEMP	Given a sentence, and two entities within the sentence, classify the relationship between the two entities based on the provided sentence. All possible Relationships are listed below with explanations: - [Possible Relation 1]: [Relation 1 Template] - [Possible Relation 2]: [Relation 2 Template] - [NoTA Relation]: [NoTA Relation Template] Sentence: [Sentence S] Entity 1: [Head Entity E_h] Entity 2: [Tail Entity E_t] Relationship:
NLI4RE	In this task, you will be presented with a premise and a hypothesis sentence. Determine whether the hypothesis sentence entails (implies), contradicts (opposes), or is neutral with respect to the given premise sentence. Please answer with "Contradiction", "Neutral", or "Entailment". Premise: [Sentence S] Hypothesis: [Entities in Relation 1 Template] Category:
QA4RE	Determine which option can be inferred from the given Sentence. Sentence: [Sentence S] Options: A. [Entities in Relation 1 Template] B. [Entities in Relation 2 Template] C. [Entities in NoTA Relation Template] Which option can be inferred from the given Sentence? Option:

Table 12: Prompt Formats of frameworks for LLMs in this paper. We only demonstrate NLI4RE with 1 template for simplicity.

Relation	**Template**
no_relation	$\{E_h\}$ has no known relations to $\{E_t\}$
per:stateorprovince_of_death	$\{E_h\}$ died in the state or province $\{E_t\}$
per:title	$\{E_h\}$ is a $\{E_t\}$
org:member_of	$\{E_h\}$ is the member of $\{E_t\}$
per:other_family	$\{E_h\}$ is the other family member of $\{E_t\}$
org:country_of_headquarters	$\{E_h\}$ has a headquarter in the country $\{E_t\}$
org:parents	$\{E_h\}$ has the parent company $\{E_t\}$
per:stateorprovince_of_birth	$\{E_h\}$ was born in the state or province $\{E_t\}$
per:spouse	$\{E_h\}$ is the spouse of $\{E_t\}$
per:origin	$\{E_h\}$ has the nationality $\{E_t\}$
per:date_of_birth	$\{E_h\}$ has birthday on $\{E_t\}$
per:schools_attended	$\{E_h\}$ studied in $\{E_t\}$
org:members	$\{E_h\}$ has the member $\{E_t\}$
org:founded	$\{E_h\}$ was founded in $\{E_t\}$
per:stateorprovinces_of_residence	$\{E_h\}$ lives in the state or province $\{E_t\}$
per:date_of_death	$\{E_h\}$ died in the date $\{E_t\}$
org:shareholders	$\{E_h\}$ has shares hold in $\{E_t\}$
org:website	$\{E_h\}$ has the website $\{E_t\}$
org:subsidiaries	$\{E_h\}$ owns $\{E_t\}$
per:charges	$\{E_h\}$ is convicted of $\{E_t\}$
org:dissolved	$\{E_h\}$ dissolved in $\{E_t\}$
org:stateorprovince_of_headquarters	$\{E_h\}$ has a headquarter in the state or province $\{E_t\}$
per:country_of_birth	$\{E_h\}$ was born in the country $\{E_t\}$
per:siblings	$\{E_h\}$ is the siblings of $\{E_t\}$
org:top_members/employees	$\{E_h\}$ has the high level member $\{E_t\}$
per:cause_of_death	$\{E_h\}$ died because of $\{E_t\}$
per:alternate_names	$\{E_h\}$ has the alternate name $\{E_t\}$
org:number_of_employees/members	$\{E_h\}$ has the number of employees $\{E_t\}$
per:cities_of_residence	$\{E_h\}$ lives in the city $\{E_t\}$
org:city_of_headquarters	$\{E_h\}$ has a headquarter in the city $\{E_t\}$
per:children	$\{E_h\}$ is the parent of $\{E_t\}$
per:employee_of	$\{E_h\}$ is the employee of $\{E_t\}$
org:political/religious_affiliation	$\{E_h\}$ has political affiliation with $\{E_t\}$
per:parents	$\{E_h\}$ has the parent $\{E_t\}$
per:city_of_birth	$\{E_h\}$ was born in the city $\{E_t\}$
per:age	$\{E_h\}$ has the age $\{E_t\}$
per:countries_of_residence	$\{E_h\}$ lives in the country $\{E_t\}$
org:alternate_names	$\{E_h\}$ is also known as $\{E_t\}$
per:religion	$\{E_h\}$ has the religion $\{E_t\}$
per:city_of_death	$\{E_h\}$ died in the city $\{E_t\}$
per:country_of_death	$\{E_h\}$ died in the country $\{E_t\}$
org:founded_by	$\{E_h\}$ was founded by $\{E_t\}$

Table 13: Templates for TACRED and TACREV datasets.

Relation	**Template**
no_relation	$\{E_h\}$ has no known relations to $\{E_t\}$
per:religion	$\{E_h\}$ has the religion $\{E_t\}$
org:country_of_branch	$\{E_h\}$ has a branch in the country $\{E_t\}$
org:stateorprovince_of_branch	$\{E_h\}$ has a branch in the state or province $\{E_t\}$
org:city_of_branch	$\{E_h\}$ has a branch in the city $\{E_t\}$
org:shareholders	$\{E_h\}$ has shares hold in $\{E_t\}$
org:top_members/employees	$\{E_h\}$ has the high level member $\{E_t\}$
org:members	$\{E_h\}$ has the member $\{E_t\}$
org:website	$\{E_h\}$ has the website $\{E_t\}$
per:parents	$\{E_h\}$ has the parent $\{E_t\}$
org:number_of_employees/members	$\{E_h\}$ has the number of employees $\{E_t\}$
org:political/religious_affiliation	$\{E_h\}$ has political affiliation with $\{E_t\}$
per:age	$\{E_h\}$ has the age $\{E_t\}$
per:origin	$\{E_h\}$ has the nationality $\{E_t\}$
org:alternate_names	$\{E_h\}$ is also known as $\{E_t\}$
per:other_family	$\{E_h\}$ is the other family member of $\{E_t\}$
per:identity	$\{E_h\}$ is the identity/pronoun of $\{E_t\}$
per:identity	$\{E_h\}$ and $\{E_t\}$ are the same person
per:siblings	$\{E_h\}$ is the siblings of $\{E_t\}$
org:member_of	$\{E_h\}$ is the member of $\{E_t\}$
per:children	$\{E_h\}$ is the parent of $\{E_t\}$
per:employee_of	$\{E_h\}$ is the employee of $\{E_t\}$
per:spouse	$\{E_h\}$ is the spouse of $\{E_t\}$
org:dissolved	$\{E_h\}$ dissolved in $\{E_t\}$
per:schools_attended	$\{E_h\}$ studied in $\{E_t\}$
per:country_of_death	$\{E_h\}$ died in the country $\{E_t\}$
per:stateorprovince_of_death	$\{E_h\}$ died in the state or province $\{E_t\}$
per:city_of_death	$\{E_h\}$ died in the city $\{E_t\}$
per:date_of_death	$\{E_h\}$ died in the date $\{E_t\}$
per:cause_of_death	$\{E_h\}$ died because of $\{E_t\}$
org:founded	$\{E_h\}$ was founded in $\{E_t\}$
org:founded_by	$\{E_h\}$ was founded by $\{E_t\}$
per:countries_of_residence	$\{E_h\}$ lives in the country $\{E_t\}$
per:stateorprovinces_of_residence	$\{E_h\}$ lives in the state or province $\{E_t\}$
per:cities_of_residence	$\{E_h\}$ lives in the city $\{E_t\}$
per:country_of_birth	$\{E_h\}$ was born in the country $\{E_t\}$
per:stateorprovince_of_birth	$\{E_h\}$ was born in the state or province $\{E_t\}$
per:city_of_birth	$\{E_h\}$ was born in the city $\{E_t\}$
per:date_of_birth	$\{E_h\}$ has birthday on $\{E_t\}$
per:charges	$\{E_h\}$ is convicted of $\{E_t\}$
per:title	$\{E_h\}$ is a $\{E_t\}$

Table 14: Templates for RETACRED datasets.

Relation	**Template**
Other	{subj} has no known relations to {obj}
Component-Whole(e1,e2)	{subj} is the component of {obj}
Component-Whole(e2,e1)	{obj} is the component of {subj}
Instrument-Agency(e1,e2)	{subj} is the instrument of {obj}
Instrument-Agency(e2,e1)	{obj} is the instrument of {subj}
Member-Collection(e1,e2)	{subj} is the member of {obj}
Member-Collection(e2,e1)	{obj} is the member of {subj}
Cause-Effect(e1,e2)	{subj} has the effect {obj}
Cause-Effect(e2,e1)	{obj} has the effect {subj}
Entity-Destination(e1,e2)	{obj} is the destination of {subj}
Entity-Destination(e2,e1)	{subj} is the destination of {obj}
Content-Container(e1,e2)	{obj} contains {subj}
Content-Container(e2,e1)	{subj} contains {obj}
Message-Topic(e1,e2)	{obj} is the topic of {subj}
Message-Topic(e2,e1)	{subj} is the topic of {obj}
Product-Producer(e1,e2)	{obj} produces {subj}
Product-Producer(e2,e1)	{subj} produces {obj}
Entity-Origin(e1,e2)	{subj} origins from {obj}
Entity-Origin(e2,e1)	{obj} origins from {subj}

Table 15: Templates for SemEval datasets.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8.

□ A2. Did you discuss any potential risks of your work?
Not applicable. Our work helps LLM solve the relation extraction tasks, we don't anticipate any risks.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Before Section 1 (abstract) and Section 1 (introduction).

☑ A4. Have you used AI writing assistants when working on this paper?
Grammarly. Grammar check for sections 1-8.

B ☑ Did you use or create scientific artifacts?

Section 3, Section 4.

☑ B1. Did you cite the creators of artifacts you used?
Section 4.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
For artifacts we used in the paper, they have licenses in the public GitHub repos.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
For artifacts we used in the paper, they have licenses in the public GitHub repos. We are following the standard use of these artifacts. Our code will be released under the same license.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
These datasets are widely used as relation extraction benchmarks in the research field and as far as we know, no previous work has reported offensive or sensitive content in these datasets.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Our code focuses on using OpenAI API for a specific task, relation extraction, and we only have tested our code on standard English benchmarks for the relation extraction task.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Sections 5 and 6.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B.1

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.2 and Appendix B.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4.2 Experimental Setup and Section 5.4. Averaged results over multiple runs are reported.

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No, we use the standard data format without extra processing and we use official GitHub repos for baseline comparison.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.