ERA-CoT: Improving Chain-of-Thought through Entity Relationship Analysis

Anonymous ACL submission

Abstract

Large language models (LLMs) have achieved commendable accomplishments in various natural language processing tasks. However, LLMs still encounter significant challenges when dealing with complex scenarios involving multiple entities. These challenges arise from the presence of implicit relationships that demand multi-step reasoning. In this paper, we propose a novel approach ERA-CoT, which aids LLMs in understanding context by capturing relationships between entities and supports the reasoning of diverse tasks through Chain-of-Thoughts (CoT). Experimental results show that ERA-CoT demonstrates the superior performance of our proposed method compared to current CoT prompting methods, achieving a significant improvement of an average of 5.1% on GPT3.5 compared to previous SOTA baselines. Our analysis indicates that ERA-CoT increases the LLM's understanding of entity relationships, significantly improves the accuracy of question answering, and enhances the reasoning ability of LLMs.

1 Introduction

Large language models (LLMs) (Hoffmann et al., 2022; Chowdhery et al., 2023; Touvron et al., 2023) have shown remarkable in-context learning capabilities in various natural language processing (NLP) tasks, including machine translation (Vilar et al., 2022; Moslem et al., 2023), question answering (Robinson et al., 2022; Li et al., 2022; Lazaridou et al., 2022), and named entity extraction (Chowdhery et al., 2023; Brown et al., 2020), etc. Recently, prompting strategies like Chain-of-Thought (CoT) (Wei et al., 2022) have garnered attention due to their capacity to significantly enhance LLMs reasoning capabilities. Considering the ability of CoT to guide LLMs in breaking down complex reasoning processes into simple steps, it stands out compared to standard zero-shot and few-shot methods.

However, due to the presence of numerous entities such as characters, locations, etc., and the multitude of implicit relationships among them in certain scenarios, CoT still faces significant challenges in handling these situations. Named Entity Recognition (NER) has typically been employed when addressing these tasks. NER is a sequence labeling task in nature, where the model needs to assign an entity-type label to each token within a sentence (Wang et al., 2023b). Relation extraction is a category of methods for handling entity relationships within text passages. Various studies (Li et al., 2023; Zhang et al., 2023a) have also investigated the performance of LLMs in zero-shot relation extraction. However, without additional prompts, LLMs have limited entity and relation extraction capabilities. Considering the importance of contextual content in answering questions, addressing knowledge-intensive tasks also requires a comprehensive analysis of entity relationships.

In this paper, we propose Entity Relationship Analysis with Chain-of-Thought (ERA-CoT), a novel framework to better address reasoning tasks in complex entity scenarios. First, we extract all the entities involved in the text; second, we extract the directly mentioned explicit relationships between entities based on the text; then, we infer the indirect implicit relationships between entities based on these explicit relationships and the hidden information in the text; after that, we let the model score the implicit relationships based on the reliability of the relationships, set a threshold for judging the reliability of the relationships, and eliminate the implicit relationships that are lower than the threshold; finally, answer the questions based on the previously extracted entities and the obtained implicit and explicit relationships.

We conducted experiments on six widely adopted datasets and compared with four baseline methods. The results show that ERA-CoT outperforms baselines on nearly all benchmarks, achiev-

ing a significant improvement of about 5.1% on average. From the performance results, our method outperforms on all three types of reasoning problems: commonsense reasoning, mathematical reasoning, and logical reasoning. This indicates that enhancing the model's understanding of entity relationships can significantly boost the reasoning abilities and accuracy in answering questions of LLMs. Our main contributions can be summarized as follows.

- We introduce ERA-CoT, a novel framework designed to conduct relationship analysis among multiple entities within complex scenarios during the zero-shot problem-solving process, which significantly strengthens the reasoning and comprehension abilities of LLMs.

- Our method extends entity relationship analysis and relation extraction to CoT. It is capable of both further complex relationship inference after entity extraction in NER, and step-by-step accurate logical analysis for any complex scenario on a zero-shot setting.

- Compared to baselines, we achieved an accuracy improvement of approximately 7.3%. Our approach excels not only on GPT-3.5 but also demonstrates significant improvements on the open-source large model Llama-2. This indicates the versatility of our method for problem reasoning across various models and scenarios.

2 Related Work

2.1 Chain of thought

To utilize LLMs to solve more complex and logical reasoning tasks, Wei et al. (2022) extended in-context learning by introducing the concept of Chain of Thought (CoT) through a step-by-step reasoning process. Kojima et al. (2022) found that simply adding a leading sentence "Let's think step by step" to a cue allowed LLMs to perform zero-shot logical reasoning without any additional human prompts (Chu et al., 2023). Subsequently, CoT-SC (Wang et al., 2023c) introduces a self-consistency strategy to replace the greedy decoding strategy. Auto-CoT(Zhang et al., 2023b) automatically constructs CoT based on questions, eliminating the instability of manual prompts. Complex-CoT(Fu et al., 2023) employs multi-step reasoning estimation on CoT based on complexity. RE2(Xu et al., 2023a) utilizes a question rephrasing strategy to enhance the model's understanding of questions with zero prompts. Wang et al. (2023a) breaks down the problem into planning and solving steps to generate and answer the Chain-of-Thought. These studies highlight the importance of CoT in enhancing LLMs' reasoning and planning abilities in complex situations. However, further refinement of CoT is needed in scenarios involving complex relationships with multiple entities.

2.2 Named Entity Recognition

Named Entity Recognition (NER) is the task of identifying mentions of entities from unstructured text and categorizing them properly (Moscato et al., 2023). NER not only serves as a standalone tool for Information Extraction (IE) but also plays a crucial role in various NLP applications, such as text understanding (Zhang et al.; Cheng and Erk, 2020), information retrieval (Luo et al., 2020b; Taillé et al., 2020), question answering (Luo et al., 2020a), machine translation (Malmasi et al., 2022), knowledge base construction (Tabassum et al., 2020), etc.

Recently, with the popularity of LLMs, NER has seen more profound development. Malmasi et al. (2022) presents a large multilingual dataset to represent contemporary challenges including low-context scenarios, and syntactically complex entities in NER. Wang et al. (2023b) proposes GPT-NER to resolve the gap by transforming the sequence labeling task to a generation task that can be easily adapted by LLMs. Das et al. (2022) and Ashok and Lipton (2023) study few-shot NER and demonstrate the great performance of few-shot NER in prompt engineering and cross-domain NER. This paper will go further by analyzing the relationships between entities.

2.3 Relation Extraction

Relation extraction involves extracting relationships based on relevant context, extracting the relationship between two given entities. It plays a crucial role in information extraction and knowledge construction. Traditional relation extraction involves fine-tuning pre-trained language models to learn relationship features, thereby providing solutions (Han et al., 2019; Zhou and Chen, 2022; Lyu and Chen, 2021; Xu et al., 2023b) for relation extraction in the few-shot setting. Some studies (Jimenez Gutierrez et al., 2022; Wei et al., 2023) have explored the performance of relation extraction on large-scale models and discussed the drawbacks of large models in relation extraction. Zhang et al. (2023a) aligns relation extraction with ques-

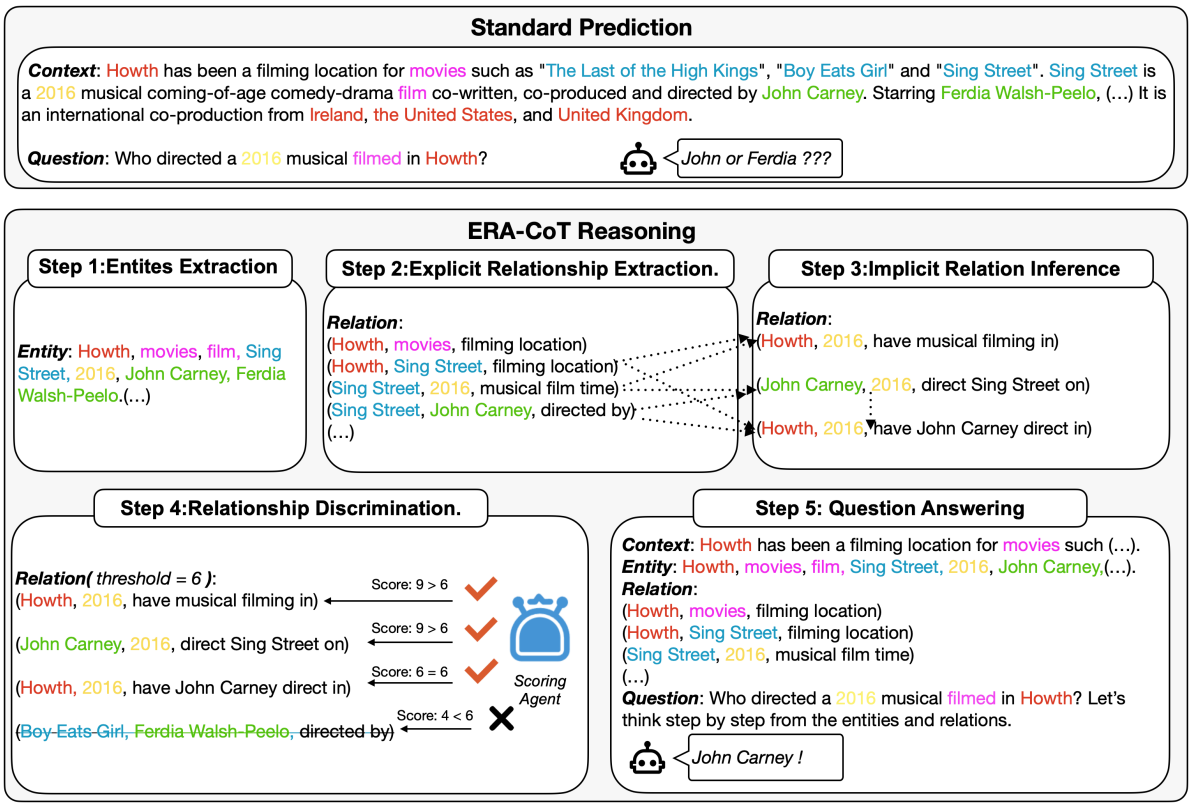

Figure 1: The top of the figure represents the standard prediction process. The bottom of the figure shows the five-step inference process of ERA-CoT, which relies on the extraction of entities and the inference and analysis of relationships between entities to obtain the results.

tion answering, obtaining corresponding relationships through question and answer based on relation templates. Li et al. (2023) transforms relation extraction into a multi-stage question-answering process, offering a novel approach to address zero-shot relation extraction. Wan et al. (2023) give a solution to relation extraction under a few-shot prompting setting. Our approach combines Chain-of-Thought with relation extraction by optimizing the relation extraction process. Through stage-wise relation extraction, it enhances the ability to extract relationships while strengthening problem-solving capabilities.

3 Methodology

As shown in Figure 1, we introduce ERA-CoT, a novel framework to enhance the model's understanding and reasoning of entity relationships in various NLP tasks. It consists of five stages, each involving different degrees of enhanced understanding of entity relations. Simultaneously, leveraging the model's capabilities, it can jointly learn explicit and implicit relationships in the context throughout this progressive process, filtering out relevant knowledge to enhance the capability and performance of question reasoning.

Problem Formulation. Given an input sequence x, the CoT method is to predict the answer via:

$$y = \arg\max_{y_i} P(y_i|\mathcal{T}, x), \tag{1}$$

where $\mathcal{T}$ is the task instruction and y_i indicate all possible results of y. Our framework is to optimize the process based on the following steps.

Step 1: Entities Extraction. Leveraging the information extraction capability of LLMs (Ashok and Lipton, 2023), we present the sentence to the model and request it to extract all n entities $\mathcal{E} = \{(s_i, t_i)\}_{i=1}^{n}$ from the provided text, expressing them as a pair relationship $e_i = (s_i, t_i)$. These relationships encompass specific information about the entities span s_i and their corresponding entity type t_i. Formally, we provide an input sentence to the large model, utilizing its NER capability to predict the corresponding entity spans and classifications. The option of entity types is related to a predefined entity set S.

Additionally, we adopt Self-Consistency (SC) (Narang et al., 2023) to evaluate the consistency of NER, as an entity is extracted from the query x, it

is measured by LLM to verify n times to determine whether it is an entity. If the upvote is higher than $\lceil \frac{n}{2} \rceil$, the entity is deemed as a valid entity. Such a design could help to remove false NER extraction.

Step 2: Explicit Relationship Extraction. We aim to explore the relevant relationships between different entities in the zero-shot setting. While entities may have direct relationships in a sentence, we generate multiple pairs of relationships for each entity. Specifically, leveraging the contextual capabilities of the LLM, we extract all relation entity pairs that are directly stated in the context. The relation could be simplified as triplets (e_i, e_j, r) where entity e_i and entity e_j are sampled from texts and r is the relations of these two entities. SC method is utilized to evaluate the consistency of explicit relation as previous step. The explicit relation set $\mathcal{R}_e$ could formulated as:

$$\mathcal{R}_e = \bigcup_{i,j \in \mathcal{E}} \{(e_i, e_j, r)\}. \quad (2)$$

Through this process, the explicit relations are deconstructed into triplets formulation. As explicit relationships are relatively clear in the text, we use LLMs to extract explicit relationships from the text. Then, we can use these relations to help the following steps to find implicit entity relations.

Step 3: Implicit Relationships Inference. We aim to perform entity relationship inference based on the preceding steps. While implicit relationships require multi-step inference, they are more challenging to discover compared to explicit relationships that can be directly extracted from the context. Therefore, we need to infer implicit relationships based on previously found explicit relationships. Specifically, we provide the original context x along with all generated relationships $\mathcal{R}_0$ in the preceding steps, request LLMs to make reasoning, and generate k most relevant relation. Assume that the intermediate relations are $T_i = (e_i, e_{i+1}, r_i)$. It can help to generate reasonable relation triplets $T_{1 \to n} = (e_1, e_n, r_k)$ from a relation chain T_1, T_2, ..., T_{n-1}, where n is length of chain.

The procedure can be formalized using the following equation while r_k indicates the k different relations from LLMs reasoning:

$$\mathcal{R}'_i = \bigcup_{i,j \in \mathcal{E}} \{(e_i, e_j, r_k)\}. \quad (3)$$

Step 4: Relationship Discrimination. Relying on the Self-Correction (Ganguli et al., 2023) capability of the LLM, we set an LLM as a scoring agent. We provide the origin context and all triplets relations generated in Step 3 to the agent. Then, each relation gets a score from the agent. For triplets that are more likely to express the correct relationship, we assign a higher score to this relationship. A score threshold v_{th} is established to assess whether the model correctly infers a relationship. Formally, for each triplet, the scoring agent model gives a value $\mathcal{V}(i, j, k)$ to assess the confidence level of the relation triplet (e_i, e_j, r_k).

For scores below our pre-defined threshold v_{th}, we consider the relationships below the threshold scores as irrelevant. Relationships deemed irrelevant or incorrect are eliminated, and those remaining with higher scores are considered correct in the relationship discrimination process. This approach helps eliminate some erroneous relationships caused by model reasoning during relationship discrimination, thereby improving its consistency and accuracy. In this case, the implicit relation set could be stated as:

$$\mathcal{R}_i = \bigcup_{i,j \in \mathcal{E}, \mathcal{V}(i,j,k). \geq v_{th}} \{(e_i, e_j, r_k)\}. \quad (4)$$

Step 5: Question Answering. Building upon all the relationships described above, we formalize their expressions and incorporate them with the original context into prompts. We utilize all entities, all relation triplets, and the context to predict the questions in this step. The ERA-CoT method could finally formulate as:

$$\text{y} = \arg\max_{y_i} P(y_i | \mathcal{R} = [\mathcal{R}_e, \mathcal{R}_i], \mathcal{E}, \mathcal{T}, S, x). \quad (5)$$

4 Experimental Setup

4.1 Datasets and Models

We consider three reasoning scenarios, i.e., commonsense reasoning, logical reasoning, and mathematical reasoning. Specifically, for commonsense reasoning, we use StrategyQA (Geva et al., 2021) and CSQA (Talmor et al., 2019); for logical reasoning, we use LogiQA (Liu et al., 2021), HotpotQA (Yang et al., 2018), and 2WikiMultiHopQA (Ho et al., 2020); for mathematical reasoning, we use GSM8K (Cobbe et al., 2021). For models, we use GPT3.5 (with 175 billion parameters) (OpenAI, 2023) and Llama2$_{13B}$ (Touvron et al., 2023).

Model	Methods	StrategyQA	CSQA	LogiQA	HotpotQA	2WikiMHQA	GSM8K
GPT3.5	Vanilla LM	65.4	72.1	28.2	49.7	55.7	52.5
	CoT	63.2	77.2	36.4	52.4	61.2	70.4
	CoT-SC@5	65.1	78.3	38.2	52.8	65.6	74.8
	Auto-CoT	64.6	77.6	38.6	53.1	64.3	77.1
	Complex-CoT	64.2	76.2	38.6	52.5	65.3	**80.1**
	PS	65.7	77.5	37.8	53.2	63.8	76.2
	PS+	66.2	77.1	38.9	53.7	64.5	75.8
	RE2	67.1	79.3	39.5	53.3	65.4	76.5
	ERA-CoT	**71.4**	**83.2**	**45.2**	**58.4**	**70.2**	79.5
$\text{Llama2}_{\text{13B}}$	Vanilla LM	57.2	58.3	24.5	34.2	28.2	17.8
	CoT	55.1	64.2	30.2	37.1	32.4	18.9
	CoT-SC@5	57.2	66.8	32.4	36.8	34.6	21.2
	Auto-CoT	56.8	66.5	31.9	37.5	35.2	20.1
	Complex-CoT	54.8	65.2	32.1	37.1	35.1	23.8
	PS	56.8	66.2	31.6	36.9	34.2	22.4
	PS+	57.6	66.9	32.4	36.7	34.8	23.1
	RE2	58.4	67.5	33.1	37.5	36.1	22.9
	ERA-CoT	**61.5**	**72.6**	**35.5**	**39.2**	**38.9**	**24.5**

Table 1: Main experimental results. The best results are highlighted in bold. We use accuracy as the evaluation metric. CoT-SC@5 represents retrieving five CoT reasoning chains to make majority votes.

4.2 Baselines

To evaluate our method holistically, we compare ERA-CoT with the leading CoT methods baselines:

Vanilla LM, directly presents tasks and the corresponding questions, predicting the outcomes of the questions through in-context learning.

Chain-of-Thought (CoT) (Wei et al., 2022), predicts the answers by generating explanations and steps.

CoT-SC (Wang et al., 2023c), generates multiple paths of Chain-of-Thought and votes to select the highest-voted result as the final result.

Auto-CoT (Zhang et al., 2023b), is a baseline that automatically generates multi-step reasoning in natural language.

Complex-CoT (Fu et al., 2023) utilizes a complexity-based strategy, sampling multiple chain-of-thoughts, and selecting answers that consistently align across complex inference chains through a majority vote.

PS and PS+ (Wang et al., 2023a), is a zero-shot CoT that breaks down the problem into planning and solving steps to generate the answers of Chain-of-Thought. PS+ extracts more details information like variables to help the inference process.

RE2 (Xu et al., 2023a), a plug-and-play approach that entails re-reading the question before engaging in the reasoning process.

4.3 Implementation

We access the GPT models through the OpenAI API, using `gpt-3.5-turbo-0301`. Additionally, for $\text{Llama2}_{\text{13b}}$, we utilize the model parameters provided in the original code. We set the generation temperature to 0.3. Unless otherwise specified, we set the number of generations k as 3. We use the `gpt-3.5-turbo-0301` to serve as our relation discriminator scoring agents to value. To ensure the reliability of the results, we conduct five rounds of experiments for each dataset, taking their average scores as the evaluation results. The prompts for CoT and PS are from Wei et al. (2022) and Wang et al. (2023a) as a comparison to our proposed framework. For evaluation metrics, we utilize exact match (EM) in our experiments.

5 Experiments

5.1 Main Results

ERA-CoT outperforms all baselines in all benchmarks on $\text{Llama2}_{\text{13B}}$ and 5 out of 6 benchmarks on GPT3.5. From Table 1, we can observe that our method shows great capability in three categories of reasoning problems. For instance, ERA-

Model	Datasets	Only EE	EE+ERE	EE+ERI	ERA-CoT
GPT3.5	StrategyQA	65.2	67.9	67.3	69.4
	CSQA	77.9	80.5	81.1	83.2
	LogiQA	37.2	41.5	42.1	45.2
	HotpotQA	53.5	55.8	55.9	58.4
	2WikiMHQA	64.2	68.1	67.2	70.2
	GSM8K	77.5	78.6	77.8	78.2
Llama_{13B}	StrategyQA	57.1	57.7	58.2	60.5
	CSQA	65.7	68.9	68.1	72.6
	LogiQA	31.9	32.4	33.7	35.5
	HotpotQA	35.8	36.6	36.9	39.2
	2WikiMHQA	34.4	34.9	35.2	38.9
	GSM8K	22.9	24.1	24.7	24.5

Table 2: Performance comparisons upon different combinations and settings of entities relation steps.

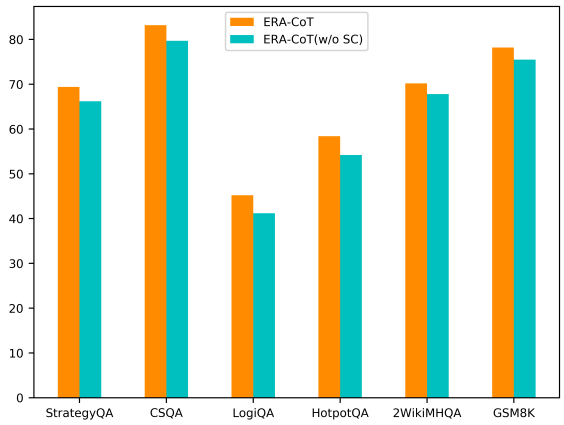

Figure 2: Comparison on the use of SC for the first two steps. Evaluation on the final performance.

CoT achieves an average improvement of 3.8% on GPT3.5. This indicates that through entity relation knowledge, the LLMs could make better predictions and enhance their performance.

Commonsense reasoning. For the StrategyQA and CommonsenseQA datasets, ERA-CoT shows an average improvement of approximately **+6.1%** compared to CoT. Although CoT-SC@5 could improve the performance of CoT, it still has a disparity of average scores with ERA-CoT on GPT3.5 (71.7% vs. 77.3%). It is worth noting that StrategyQA gets a performance drop compared to Vanilla LM, which may be caused by irrelevant text or incorrect reasoning chains. ERA-CoT mitigates the impact of irrelevant text by controlling the reasoning process based on entity relations. It performs a similar outcome on the Llama_{13B} when we apply entity relation as a prompt to instruct the model. On the StrategyQA dataset, the result achieved an improvement of **+3.1%** compared to RE2. Similar improvements are observed on the CommonsenseQA dataset, which indicates the ERA-CoT method could potentially result in very strong performances.

Logical Reasoning. Logical Reasoning contains more implicit relations. According to Table 1, ERA-CoT evaluates three logical datasets: LogiQA, HotpotQA, and 2WikiMHQA. The results on these three datasets demonstrate an average **+5.1%** improvement on GPT3.5, highlighting the effectiveness of this method compared to other baselines. Meanwhile, **ERA-CoT exhibits a better performance increment in logical reasoning tasks** that surpasses commonsense reasoning tasks and mathematical reasoning tasks, suggesting that our approach may be more suitable for tasks involving relational reasoning. On a small-scale LLM, ERA-CoT Llama_{13B} achieves similar performance, indicating that ERA-CoT may have better capabilities in long-text comprehension and entity logical reasoning. Compared to the recent competitive work RE2, our method shows an average relative improvement of **5.2%** in logical reasoning.

Mathematical Reasoning. The mathematical reasoning ability of ERA-CoT is evaluated on the GSM8K dataset. Compared to other methods, ERA-CoT outperforms most baselines on GSM8K, falling slightly behind Complex-CoT. Our approach primarily relies on the analysis of contextual entity relationships to assist the model in understanding the problems. As GSM8K involves natural language-formulated questions, this process potentially addresses some errors in the analysis of relational chains in CoT.

5.2 Ablation Studies

ERA-CoT includes multiple processes for handling entities and relationships. We combine various steps to assess the impact of entity extraction and relationship inference on model performance. We categorize the process into the following situations:

- *Only EE*: This variant represents only entity extraction, involving the use of large models for named entity recognition.

- *EE+ERE*: This setting indicates simultaneous extraction of entities and explicit relationship extraction. After completing the first two steps of ERA-CoT, this process directly proceeds to question prediction using the answers from the output of the initial two steps.

Model	Dateset	validation				w/o validation			
		k@1	k@3	k@5	k@10	k@1	k@3	k@5	k@10
GPT3.5	StrategyQA	66.4	71.4	72.8	73.1	65.2	65.9	70.8	71.9
	CSQA	79.5	83.2	84.8	85.6	78.5	80.2	83.4	84.8
	LogiQA	41.0	45.2	46.5	47.5	39.8	43.3	44.4	45.1
	HotpotQA	54.4	58.4	61.2	63.5	53.2	53.9	57.1	58.5
	2WikiMHQA	67.3	70.2	71.1	72.2	66.5	68.7	67.3	70.9
	GSM8K	75.8	79.5	80.2	80.1	74.9	76.0	77.1	77.3
$\text{Llama2}_{\text{13B}}$	StrategyQA	58.2	61.5	65.9	64.9	57.5	57.9	61.3	60.5
	CSQA	68.4	72.6	71.3	70.1	68.1	69.0	70.3	71.6
	LogiQA	33.6	35.5	37.8	40.5	31.7	34.1	35.7	38.4
	HotpotQA	37.6	39.2	40.9	42.0	36.9	38.1	41.2	41.0
	2WikiMHQA	36.0	38.9	39.5	41.1	35.5	37.2	38.1	38.9
	GSM8K	22.1	24.5	26.9	27.3	21.8	24.0	25.6	26.9

Table 3: The impact of different numbers and relationship discrimination step of implicit relations on model accuracy, k indicates the number of implicit relations reasoning by model.

- *EE+ERI*: It implies that after entity extraction, we directly infer relationships between entities and make answer predictions based on the results of entity extraction and relationship inference.

Entity extraction, relationship extraction and inference are effective for answering questions. Table 2 demonstrates the positive impact of entity relationships on task inference. Performance of *Only EE* declines on multiple datasets compared to ERA-CoT when there is no relationship extraction and inference. After entity extraction based on our prompts, task predictions show significant improvement. On GPT3.5, the average performance for each task increased by **+2.0%**, and on $\text{Llama}_{\text{13B}}$ it increased by **+1.1%**. Additionally, direct relationship inference enhances the model's answer prediction performance. However, as it does not undergo relationship extraction, the inferred relationships lack based relations of entities, resulting in less competitive evaluation performance compared to ERA-CoT. Additionally, for mathematical reasoning, the performance of *EE+ERE* and *EE+ERI* are close to ERA-CoT. This may be attributed to the simplicity of relationships in mathematical tasks, making them easy to extract or identify.

Effectiveness of Self-Consistency. From Figure 2, we can observe that removing SC resulted in an average performance decrease by an average of **-3.2%** for the ERA-CoT method on GPT3.5. This result highlights the necessity of integrating self-consistency in the first two steps of this process.

5.3 Analysis on Implicit Relationship

In the process of deducing implicit relationships, we need to identify k relationships between each pair of entities. Subsequently, these k relationships are scored based on the contextual meaning, where higher scores indicate more reliable relationships, and lower scores suggest a lower likelihood of the relationship. By setting a reasonable threshold, implicit relationships with scores surpassing this threshold are retained for the final question prediction. We investigated the impact of the value of k on model performance. Specifically, we conducted experiments on the ERA-CoT dataset using the GPT and $\text{Llama2}_{\text{13B}}$ models, evaluating their performance on different datasets.

A reasonable number of implicit relationships contributes to a better understanding of the context. The experimental results are shown in Table 3. As the number of implicit relationships increases, the model's accuracy tends to improve. This indicates that discovering more implicit relationships enhances the model's reasoning ability, ultimately increasing the likelihood of correct answer predictions. Specifically, when k is relatively small, the model's accuracy significantly improves. However, an excessively large number may lead to hallucinations. As k continues to increase, the improvement in accuracy becomes smaller, and there may even be a slight decline in accuracy. This is because in-

ferring too many implicit relationships may lead to illusion effects, affecting the model's final reasoning judgment. Therefore, considering the balance between model effectiveness and complexity, we found that setting k to 3 results in better accuracy for the model. Additionally, in complex relationship scenarios, this choice does not significantly increase the complexity of the approach.

Relation discrimination steps help eliminate incorrect relationship pairs. We investigated whether performing relation discrimination steps during model accuracy evaluation plays an important role. In the discrimination step, the model scores all implicit relationships and filters out those with scores below v_{th}. The results show that models with discrimination steps exhibit higher accuracy compared to models without discrimination steps. This process helps enhance the correctness and robustness of the model's reasoning, explaining the overall higher accuracy of models with discrimination steps.

Small LLMs still possess the ability to identify implicit relationships. Our experiments on $\text{Llama2}_{\text{13B}}$ indicate that as k increases from 1 to 5, the model shows an average improvement of **+4.4%**. However, small models have a weaker understanding of the text, and as the number of inferred implicit relationships continues to increase, small models are more prone to generating incorrect relationships, leading to unstable performance.

5.4 Error analysis

We manually analyzed 100 errors in each dataset and categorized them into following four categories: (i) **Entity Extraction Errors *(En.)*** – failure to recognize all entities in the text; (ii) **Explicit Relationship Extraction Errors *(Ex.)*** – entities extracted correctly, but failing to extract all explicit relationships in the text or extracting non-existing explicit relationships; (iii) **Implicit Relationship Inference Errors *(Im.)*** – correct extraction of entities and explicit relationships, but inferring non-existing implicit relationships; and (iv) **Answer Errors *(An.)*** – correct inference of relationships by *ERI*, but providing an incorrect answer.

Implicit relationship inference is prone to errors. Table 4 shows the error category results for each dataset. Considering the error categories, the probability of *Im.* is consistently the highest, while *En.* has a relatively lower probability. This suggests that the inference of implicit relationships has the most significant impact on the model's accuracy.

Task	Dateset	*En.*	*Ex.*	*Im.*	*An.*
Commonsense Reasoning	StrategyQA	7%	21%	23%	15%
	CSQA	6%	16%	16%	10%
Logical Reasoning	LogiQA	10%	21%	32%	3%
	HotpotQA	7%	16%	38%	10%
	2WikiMHQA	8%	22%	35%	13%
Mathematical Reasoning	GSM8K	2%	10%	21%	12%

Table 4: Error categories per dataset. Multiple categories are allowed for each example.

Error rates are also influenced by dataset characteristics. We observe that for common-sense reasoning datasets, error rates for *Ex.* and *Im.* are close. This is attributed to the model potentially misjudging relationships between entities due to an incomplete understanding of common-sense knowledge.

For logical reasoning datasets, *Ex.* is higher compared to other datasets. This is because these datasets typically involve longer texts with more relationships, requiring the model to have a higher level of relationship extraction. However, compared to other errors, the error rate of *An.* is relatively small. This implies that if the relationships between entities are correctly inferred, the model is likely to answer questions correctly. This indicates that accurate relationship inference is highly beneficial for correctly answering questions in this type of dataset.

For mathematical reasoning datasets, implicit inference contributes to improving dataset accuracy, but *An.* should not be ignored. This is because even if relationship inference is correct, errors in calculations may still lead to incorrect answers.

6 Conclusions

In this paper, we introduce the method of ERA-CoT to address open-domain question answering and knowledge reasoning tasks. By leveraging inference and detection of entity relationships, this method exhibits remarkable performance in tasks with lengthy texts or those involving numerous and complex entity relationships. Extensive experiments demonstrate the superiority of the proposed method in various reasoning categories. We hope that our method can be applied to enhance the performance of LLMs in diverse domains.

Limitation

Our work has limitations in certain scenarios. First, ERA-CoT relies on context analysis through the extraction of relationships and entities, so its performance improvement is not significant in tasks with fewer entity relationships, such as symbolic reasoning tasks. Additionally, due to the various relationships between entities, even after relationship extraction and inference, there may still be some relationships that the model fails to correctly infer or extract. This could result in missing relationships during result predictions, affecting the accuracy of predictions. In future work, we aim to address these two issues. We plan to explore whether the model can effectively understand the internal structure of entities and conduct a more in-depth analysis of entity relationships to enhance the effectiveness of LLM prompting.

Ethics Statement

The datasets we used are sourced from the current public datasets. The prompts we employed do not collect or utilize personal information or information from other individuals. Furthermore, they do not contain any sensitive words or oppose any individual or group. Our work only extracts the entity and relation from these datasets to predict the answer to tasks. Our method strictly adheres to the license and policies of released LLMs and publicly available datasets, and our work could be further integrated with other methods.

References

Dhananjay Ashok and Zachary C. Lipton. 2023. Promptner: Prompting for named entity recognition.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Pengxiang Cheng and Katrin Erk. 2020. Attending to entities for better text understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7554–7561.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca J Passonneau, and Rui Zhang. 2022. Container: Few-shot named entity recognition via contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6338–6353.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.

D Ganguli, A Askell, N Schiefer, T Liao, K Lukošiūtė, A Chen, et al. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An open and extensible toolkit for neural relation extraction. 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): System Demonstrations*, pages 169–174, Hong Kong, China. Association for Computational Linguistics.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

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.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*.

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

Junlong Li, Zhuosheng Zhang, and Hai Zhao. 2022. Self-prompting large language models for open-domain qa. *arXiv preprint arXiv:2212.08635*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3622–3628.

Da Luo, Jindian Su, and Shanshan Yu. 2020a. A bert-based approach with relation-aware attention for knowledge base question answering. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Ying Luo, Fengshun Xiao, and Hai Zhao. 2020b. Hierarchical contextualized representation for named entity recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8441–8448.

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.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. Multiconer: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809.

Vincenzo Moscato, Marco Postiglione, and Giancarlo Sperlí. 2023. Few-shot named entity recognition: definition, taxonomy and research directions. *ACM Transactions on Intelligent Systems and Technology*, 14(5):1–46.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*.

Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the Eleventh International Conference on Learning Representations*.

OpenAI. 2023. Gpt-4 technical report.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2022. Leveraging large language models for multiple choice question answering. *arXiv preprint arXiv:2210.12353*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Jeniya Tabassum, Mounica Maddela, Wei Xu, and Alan Ritter. 2020. Code and named entity recognition in stackoverflow. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4913–4926.

Bruno Taillé, Vincent Guigue, and Patrick Gallinari. 2020. Contextualized embeddings in named-entity recognition: An empirical study on generalization. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 383–391. Springer.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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 4149–4158.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*.

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. *arXiv preprint arXiv:2305.02105*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023b. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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. 2023. Zero-shot information extraction via chatting with chatgpt.

Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, and Jian guang Lou. 2023a. Re-reading improves reasoning in language models.

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.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

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.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

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

A Dataset Statistics

Table 5 provides detailed information about the data included in the experiment, where the sampled data are randomly selected from datasets, with a minimum of 500 samples taken.

Dataset	Num.	Length	Domain
StrategyQA	1390	12.3	Commonsense Reasoning
CSQA	3675	18.5	Commonsense Reasoning
LogiQA	1735	103.8	Logical Reasoning
HotpotQA	3000	114.2	Logical Reasoning
2WikiMHQA	1539	95.3	Logical Reasoning
GMS8K	1319	46.9	Mathematical Reasoning

Table 5: Dataset statistics, where "Num." represents the number of sampled datasets, and "Length" is the number of average tokens in the sampled dataset.

B Prompting Template

The following is the prompt statement used by ERA-CoT, guiding different steps that need to be processed with Chain-of-Thought. These prompts are proposed under a zero-shot setting, intended for resolution based on the initially provided context and query of the question.

B.1 Prompt for Entities Extraction

Entities Extraction

Given a sentence, possible entities may include:[*individuals, organizations, locations, ..., percentages*]. Find all entities based on the provided sentence.
Sentence: [Sentence S]
Entities:

B.2 Prompt for Entities Relation Extraction

Entities Relation Extraction

Given a sentence, and all entities within the sentence. Extract all relationships between entities which directly stated in the sentence. Every relationship stated as a triple: (E_A, E_B, $Relation$)
Sentence: [Sentence S]
Entities: [Entities List $\{E_i\}$]
Relationships:

B.3 Prompt for Entities Relation Inference

Entities Relation Inference

Given a sentence, all entities, and all explicit relationships within the sentence. Infer all possible implicit relationships between entities. For each pair of entities, infer up to [k] implicit relationships.
Every relationship stated as a triple: (E_A, E_B, $Relation$)
Sentence: [Sentence S]
Entities: [Entities List $\{E_i\}$]
Explicit Relationships:[Explicit Relationship List $\{R_e\}$]
Implicit Relationships:

B.4 Prompt for Relationship Discrimination

Relationship Discrimination

Given a sentence, and all uncertain relationships within the sentence. Score the confidence level of each relationship.
The confidence score ranges from 0 to 10, where a higher score indicates a higher likelihood of the relationship being correct.
Every relationship stated as a triple: (E_A, E_B, $Relation$)
Sentence: [Sentence S]
Entities: [Entities List $\{E_i\}$]
Uncertain Relationships:[Implicit Relationship List $\{R_i\}$]
Scores:

B.5 Prompt for Question Answering

Question Answering

Given a sentence, all entities and all relationships within the sentence. Answering the question.
Every relationship stated as a triple: (E_A, E_B, $Relation$)
Sentence: [Sentence S]
Entities: [Entities List $\{E_i\}$]
Relationships:[Relationship List $\{R\}$]
Question: [Question Q]
Answer:

C Experiment Cost

We conduct experiments using the GPT3.5 API, utilizing the `gpt-3.5-turbo-0301` model, at a cost of \$0.002 per 1K tokens. In total, we spend \$529.

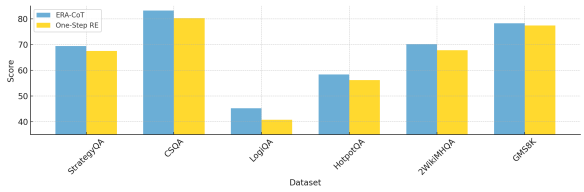

Figure 3: Comparison of performance between Single-Step and Multi-Step Relation Extraction. The score is evaluated on accuracy.

D Evaluation Metrics

We use accuracy and exact match as the evaluation metric for different datasets. Specifically, for datasets like StrategyQA, CSQA, and LogiQA that contain options, we utilize the accuracy based on whether the options match the standard answers. For problems like GSM8K, where the output is a number, we use regular expressions for exact match judgment of the answers. For datasets like 2WikiMQA that do not contain question options, we compare the output with answer alternatives and also use the exact match method for accuracy estimation. The same processing approach is adopted for different methods across these datasets.

E Why Not One-Step Relation Extraction?

In the experiment of relation extraction, we complete it in three steps. The first step is explicit relation extraction. The second step is implicit relation inference based on the results of explicit relation extraction. The third step is to score the implicit relationships and remove the low-scoring situations that might lead to errors. Explicit relation extraction is defined as the corresponding relationships that LLMs or fine-tuned proprietary models can directly obtain from the context, and these relations must involve the context.

For one-step relation extraction, GPT-RE (Wan et al., 2023) provides a solution by extracting examples of relation extraction from questions that are similar to the inquiry, and then using few-shot prompting to assist LLMs in generating results for relation extraction. However, in the zero-shot situation, Vanilla RE instructs LLMs to search for all the relations of entities only based on the context and the entity list. This type of query is likely to cause errors, leading to a negative impact on answering

questions.

To better assess the necessity of dividing relation extraction into three steps, we conduct ablation experiments comparing situations with only explicit relation extraction, only implicit relation reasoning, and our method. The results highlight the importance of performing explicit relation extraction and implicit relation reasoning separately. Here, we will further compare single-step relation extraction with three-step relation extraction.

Entity Relation Extraction(One Step)

Given a sentence, and all entities within the sentence. Extract all relationships between entities in the sentence.
Every relationship stated as a triple: (E_A, E_B, $Relation$)
Sentence: [Sentence S]
Entities: [Entities List $\{E_i\}$]
Relationships:

Figure 3 shows the performance gap between single-step and multi-step extraction. In logical reasoning questions like those in LogiQA and HotpotQA, there is a significant performance gap, illustrating that multi-step relation extraction helps in accurately extracting relations and preventing erroneous information from affecting the model's answer, especially when complex relationships are involved. Especially in large language models, irrelevant information can easily influence the outcome (Shi et al., 2023), and incorrect relations are more likely to lead to erroneous answers.

F Example

Context:A Creative Commons license allows for the free distribution of an otherwise copyrighted piece of work, Works that are deemed in the public domain do not require a copyright or permission to use, Any work created before 1923 is in the public domain, Sandro Boticelli's The Birth of Venus painting was from 1486.
Question: Was a person sold a Creative Commons License for Boticelli's The Birth of Venus ripped off?
Answer: False.

Entities Extraction:
Given a sentence, possible entities may include:[*individuals, organizations, locations, ..., percentages*]. Find all entities based on the provided sentence.
Sentence: A Creative Commons license allows for...
Entities: [Creative Commons license, copyrighted piece of work,public domain, copyright, permission,work, 1923, public domain,Sandro Boticelli, The Birth of Venus painting, 1486]

Explicit Relation Extraction:
Given a sentence, and all entities within the sentence. Extract all relationships between entities which directly stated in the sentence. Every relationship stated as a triple: (E_A, E_B, $Relation$)
Sentence: A Creative Commons license allows for...
Entities: [Creative Commons license, copyrighted piece of work, ...
Relationships: (Creative Commons license, copyrighted piece of work, allows for) (copyrighted piece of work, permission, require) (Works, public domain, deemed) (work, created before, 1923) (work, public domain, is in) (Sandro Boticelli's The Birth of Venus painting, was from, 1486)

Implicit Relation Inference: Given a sentence, all entities, and all explicit relationships within the sentence. Infer all possible implicit relationships between entities. For each pair of entities, infer up to 3 implicit relationships.
Every relationship stated as a triple: (E_A, E_B, $Relation$)
Sentence: A Creative Commons license allows for ...
Entities: [Creative Commons license, ...
Explicit Relationships:(Creative Commons license, copyrighted piece of work, allows for) (copyrighted piece of work, permission, require)...
Implicit Relationships:(Creative Commons license, copyrighted piece of work, is) (Creative Commons license, permission, allows for) (copyrighted piece of work, public domain, may be) (public domain, require, not) (copyrighted piece of work, created before, may be) (public domain, created before, is) (work, copyrighted piece of work, is) (work, permission, requires) (The Birth of Venus painting, copyright, may have) (The Birth of Venus painting, permission, may require) (Sandro Boticelli, The Birth of Venus painting, created) (Sandro Boticelli, The Birth of Venus painting, is in) (Sandro Boticelli, The Birth of Venus painting, is copyrighted)

Relationship Discrimination:
Given a sentence, and all uncertain relationships within the sentence. Score the confidence level of each relationship.
The confidence score ranges from 0 to 10, where a higher score indicates a higher likelihood of the relationship being correct.
Every relationship stated as a triple: (E_A, E_B, $Relation$)
Sentence: A Creative Commons license allows for ...
Entities: [Creative Commons license, ...
Uncertain Relationships:(Creative Commons license, copyrighted piece of work, is) (Creative Commons license, permission, allows for) (copyrighted piece of work, public domain, may be)...
Scores:[8,5,7,6,9,8,7,6,5,9,8,4]

Question Answering:
Given a sentence, all entities and all relationships within the sentence. Answering the question.
Every relationship stated as a triple: (E_A, E_B, $Relation$)
Sentence: A Creative Commons license allows for ...
Entities: [Creative Commons license, ...
Relationships:(Creative Commons license, copyrighted piece of work, allows for) (copyrighted piece of work, permission, require)...
Question: Was a person sold a Creative Commons License for Boticelli's The Birth of Venus ripped off?
Answer: No, a person who obtained a Creative Commons License for Boticelli's The Birth of Venus was not ripped off, because a Creative Commons license allows for the free distribution of copyrighted works, and Sandro Boticelli's The Birth of Venus painting was created in 1486, which is well before 1923, thus making it in the public domain and not requiring copyright or permission to use.

Table 6: Case on StrategyQA.