The Answer Lies Within: Self-Derived Rewards Enable Explainable Relation Extraction

Xinyu Guo[1], **Zhengliang Shi**[2], **Minglai Yang**[1], **Mahdi Rahimi**[1], **Mihai Surdeanu**[1]
[1]University of Arizona, Tucson, AZ, United States
[2]Shandong University, Qingdao, China.
xinyuguo@arizona.edu, msurdeanu@arizona.edu

Abstract

Despite their strong general reasoning, large language models still struggle with one-shot relation extraction when no candidate relation labels are given. We identify two failure modes: models often attend to irrelevant tokens rather than to the phrases that actually carry the relation, and they predict relations at a different level of abstraction from the one human annotators expect. We address both with two components: (1) CogRE, *a cognitively inspired reasoning framework* that structures RE into a series of processes mimicking human text-processing; and (2) Hit@Dict, *a reinforcement learning intermediate reward strategy* that rewards relation-bearing phrases in the reasoning trace, aligning the reasoning with the target label. The reward is computed against a credit dictionary, a set of high-value phrases extracted automatically from the model's own correct predictions. Our experiments show that our framework improves both accuracy and explanation quality by addressing these two challenges. For example, CogRE with Qwen2.5-14B-Instruct on One-shot NYT29 achieves 24.65% F1, surpassing prior reasoning-based designs. Optimizing this approach with RL using Hit@Dict further improves performance by +23.46% points. Finally, human evaluation shows that our best model generates relational phrases closely aligned with gold labels, increasing human explanation quality ratings by 54% (relative). Code

1 Introduction

Relation extraction (RE) aims to identify relations between entity mentions in text (Zelenko et al., 2003; Bunescu and Mooney, 2005) and serves as a fundamental task in high-stakes domains (Adadi and Berrada, 2018; Goodman and Flaxman, 2017). However, in such domains explainability is critical; traditional RE methods based on feature-based models, neural network, or small language models (Kambhatla, 2004; Zeng et al., 2014; Soares et al., 2019; Vacareanu et al., 2024) suffer from limited explainability. While recent works prompt large language models (LLMs) to perform structured reasoning before extracting relations, they typically rely on handcrafted relation labels and descriptions that are expensive to annotate (Li et al., 2023). To this end, we study a variant of the realistic one-shot RE setting (Alam et al., 2024) that demands both generalization and explainability. *Formally, given a support sentence for each relation and no explicit relation labels, models are required to both extract relations from test sentences and generate explanations for their predictions.*

Label	of support sentence: *per:alternate_name* of test sentence: *no_relation*
Support Sentence **Test Sentence**	**Gadahn** is also known as **Azzam al-Amriki**. **Arcandor** was known as Karstadt in **2000**.
LLMs Prediction	These two sentences convey the same relation: an entity is also known by another name. Test sentence: *per:alternate_name* ✗.

Table 1: An example of attention misalignment for the task of predicting whether the test sentence has the same relation label as a provided support sentence. Bold text marks the target entity pair. The model predicts the same relation based on superficial lexical overlap (e.g., "known as") despite an incompatible target entity pair.

However, our pilot experiments show that, despite the strong language understanding and reasoning capabilities of LLMs (Gao et al., 2023; Luo et al., 2024; Ahn et al., 2024), they still perform poorly in realistic one-shot RE with two common failure patterns: (1) *Attention Misalignment*: models fail to focus on the semantics that truly convey the relation. As in the example shown in Table 1, models are often distracted by irrelevant tokens in test sentences when attempting to mimic the relation expressed in the support sentences. (2) *Granularity Mismatch*: models fail to align with the granularity of human annotation schemas. Without predefined relation labels, they struggle to distinguish fine-grained relations (e.g., city-level *vs.*

country-level headquarters relations).

To address these problems, we propose a structured reasoning framework inspired by human cognition (CogRE) and use reinforcement learning (RL) with a Hit@Dict reward to bring LLM reasoning in line with human annotation schemas.

Specifically, CogRE mimics how people process text: comprehension does not come from storing input word by word (Miller, 1956), but from a construction–integration process that yields a coherent logical chain (Kintsch, 1988). Inspired by this, we decompose RE into three steps (shown in Part 2 of Figure 1): (1) chunking from text into logical propositions; (2) anchoring certain phrases as cues; and finally (3) integrating these cues through a verbalized explanation. Further, we introduce Hit@Dict, an intermediate reward strategy for RL that approximates reasoning supervision by counting occurrences of relation-relevant phrases against a credit dictionary. To build this credit dictionary, we extract relational cues from explanations of true-positive samples generated by vanilla LLMs. We show that these cues provide a more stable reward signal, most likely because they are derived from the model's own training distribution rather than human annotations.

By providing fine-grained supervision for the reasoning process, this intermediate strategy complements reinforcement learning with verifiable rewards (RLVR) (Guo et al., 2025), where rewards based only on the final answer or reasoning format offer limited supervision for the reasoning content without using LLM-as-a-Judge.

In summary, our contributions are fourfold:

1. We propose a novel RE framework loosely inspired by structured cognition. This bottom-up design mitigates LLM attention misalignment when handling complex sentences.

2. We design Hit@Dict, an RL intermediate reward strategy that provides fine-grained supervision for LLM reasoning by matching reasoning cues against a self-generated phrase dictionary.

3. We introduce a dual evaluation protocol that evaluates both RE performance as well as the quality of the accompanying explanations. CogRE outperforms strong RE baselines by up to +47.1% F1 (relative); RL with Hit@Dict further boosts performance by up to +187.1% (relative) over the baseline on realistic one-shot TACRED and NYT29, while improving human-rated explanation quality by 24.72% and 54.24% (relative).

4. Our pilot experiment identifies a main failure of LLMs on RE as a mismatch between the inferred relations and RE annotations granularity (e.g., failing to distinguish geographic scales in ORG:CITY_OF_HEADQUARTERS). Human analysis shows that adding Hit@Dict reward improves explanation alignment with RE labels in 37.5% of cases. Models trained with Hit@Dict tend to include relational cues aligned with gold labels (e.g., *enroll* and *attend* for PER:SCHOOLS_ATTENDED), while models without it generate vague terms like *associated*.

2 Related Work

Explainable Relation Extraction. Relation extraction is applied in high-stakes domains (Adadi and Berrada, 2018; Goodman and Flaxman, 2017), where explainability is critical. Traditional RE models, including feature-based methods, neural networks, and small language models, provide explainability through attention weights (Zhou et al., 2016), feature importance (Kambhatla, 2004), post-hoc analysis (Shahbazi et al., 2020), or rules (Vacareanu et al., 2024; Tang and Surdeanu, 2023), but offer limited explainability due to the lack of language-based explanations.

LLM Reasoning. Explicit reasoning in LLMs enhances explainability through human-readable reasoning traces (Wei et al., 2022; Chu et al., 2025). For RE, recent work leverages LLMs via few-shot prompting (Wan et al., 2023; Ma et al., 2023; Wadhwa et al., 2024; Xu et al., 2023), structured multi-step reasoning (Li et al., 2023), and instruction tuning (Ouyang et al., 2022; Qi et al., 2024; Jiao et al., 2023), but typically relies on handcrafted relation labels and descriptions. Recently, reinforcement learning with verifiable rewards has improved both accuracy and explainability (He et al., 2025). However, explanation-oriented rewards remain unexplored. Existing methods rely either on simple format-based signals (Dai et al., 2025; Wen et al., 2025; Xin et al., 2025) or costly LLM-as-a-judge approaches (Saha et al., 2025; Huang et al., 2025; Zhu et al., 2026; Yang et al., 2026). In this work, we design a cognitively structured reasoning framework for RE without such annotations, and propose an RL self-rewarding strategy that provides low-cost yet fine-grained supervision.

Hints from Cognitive Psychology. Existing work shows that cognitive psychology provides insights for LLMs and evidences their cognitive capabili-

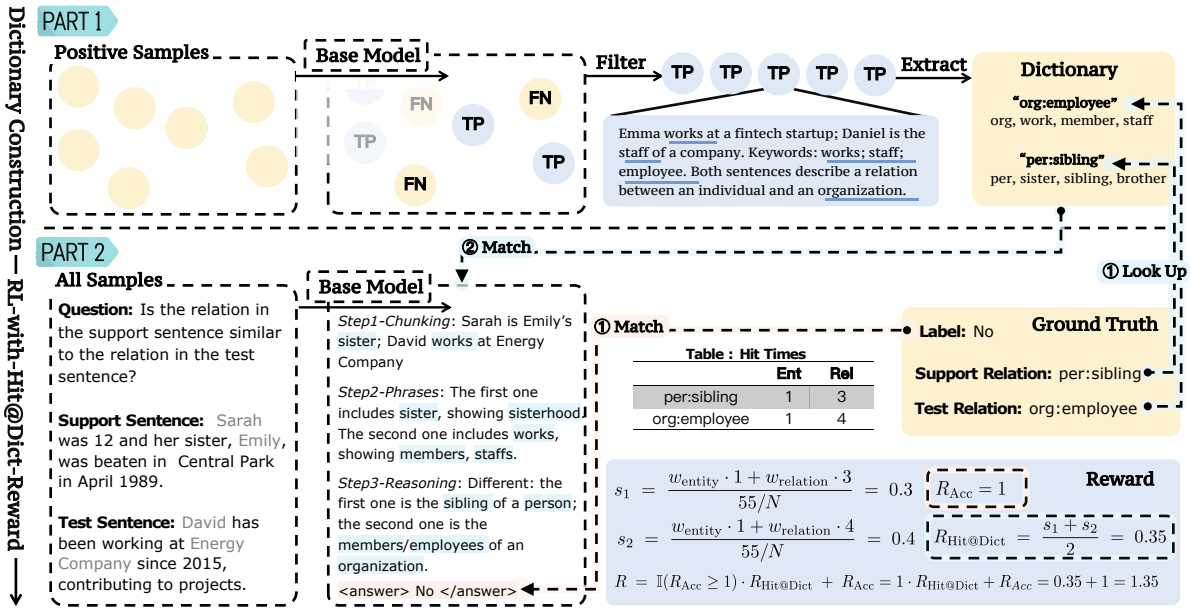


Figure 1: Framework overview. *Part 1 - Dictionary Construction*: We apply a base model to positive samples, retain true positive (TP) instances, then use GPT-4o or Qwen2.5-14B-Instruct to extract relational phrases from their explanations to construct a relational phrase dictionary. *Part 2 - Reinforcement Learning with* HIT@DICT: During RL training, the model's stepwise CogRE outputs on all training samples are scored by accuracy (answers) and HIT@DICT (explanations). To compute HIT@DICT, we first look up relational phrases from the dictionary for both the support and test relations. These phrases are then matched against the LLM outputs to obtain hit statistics (Hit Times Table), and HIT@DICT is computed as a normalized hit rate (Section 3.2).

ties (Yax et al., 2024; Niu et al., 2024). Cognitive psychology has extensively studied how humans process information. The Construction-Integration model describes comprehension in four steps: forming concepts, elaborating, inferring new propositions, and integrating them into a representation (Kintsch, 1988). Several studies show: *chunking* reduces cognitive load (Miller, 1956), *phrase anchors* guide attention (Kintsch and Van Dijk, 1978), and cognitive monitoring improves strategy adjustment (Flavell, 1979). Inspired by these, we frame RE as a three-step framework, including *chunking*, *anchoring*, and *integrative reasoning*.

3 Method

Our pilot experiments (§ 5.5) reveal that LLMs always conduct token-level matching between two sentences and overlook the semantics that truly convey the relation. To address this gap, we design a framework loosely inspired by cognitive science.

3.1 Cognitive-Structured RE

As shown in Figure 1 (Part 2), our Cognitive-Structured RE (COGRE) framework formulates RE reasoning into three steps. First, *Proposition Chunking*, where the LLM summarizes each sentence into a relational proposition. This step ensures that the LLMs' reasoning starts with compressed propositions rather than long token sequences. Next, *Phrase Anchoring*, where the LLMs anchor relational phrases in the input sentences and propositions, which grounds the LLMs' relation-matching reasoning in the original sentence and the extracted propositions. The final step is *Integrative Reasoning*. The LLMs are prompted to integrate propositions and relational phrases into a coherent reasoning chain. Formally, let the LLM $\mathcal{M}$ be parameterized by θ. Given an input pair $x = (s_1, s_2)$, where s_1 is the support sentence and s_2 is the test sentence, the model generates an explanation $\hat{z}$ followed by a final label $\hat{y} \in \{\texttt{Yes}, \texttt{No}\}$. The label indicates whether the two sentences express the same relation. The generation process can be formulated as $(\hat{z}, \hat{y}) \sim \mathcal{M}_\theta(\cdot \mid s_1, s_2)$.

3.2 Reinforcement Learning with HIT@DICT

In RL training, improving the quality of reasoning without a fine-grained reward is challenging, while format-based rewards that ignore reasoning content provide only weak guidance. To overcome these limitations, we propose an efficient and effective intermediate reward strategy, namely HIT@DICT, and integrate it with the accuracy reward to incentivize LLMs' reasoning in RE tasks.

Reward Function Design. We design a reward function with two complementary components. The first part is the HIT@DICT reward, which evaluates the occurrences of relational phrases in the LLM's explanation based on the predefined credit dictionary. The second part is the accuracy reward, which evaluates the correctness of the predicted results. The final reward is formulated as:

$$\mathcal{R} = \mathbb{I}(\mathcal{R}_{\text{Acc}} \geq 1) \cdot \mathcal{R}_{\text{Hit@Dict}} + \mathcal{R}_{\text{Acc}} \quad (1)$$

Here, $\mathcal{R}_{\text{Acc}}$ denotes the accuracy reward, while $\mathcal{R}_{\text{Hit@Dict}}$ judges the explanation quality. The indicator function $\mathbb{I}(\cdot)$ is used to gate the explanation reward, i.e., $\mathcal{R}_{\text{Hit@Dict}}$ is triggered only for correct predictions. This ensures that the LLM is incentivized to generate correct predictions and explanations aligned well with relational knowledge.

HIT@DICT Reward. As shown in Part 2 of Figure 1, the HIT@DICT reward measures how many relational phrases in model-generated explanation $\hat{z}$ match a pre-built relational phrases dictionary. This dictionary is a core component, which collects relation labels appearing in the training dataset, together with their associated relational phrases.

How can the HIT@DICT *reward be applied?* Given an explanation $\hat{z}$ and a relation label r, we compute the HIT@DICT score as follows. Let $\text{Ent}(r)$ denotes the set of entity-related phrases and $\text{Rel}(r)$ the set of relational phrases associated with r. We define the weighted hit counts as:

$$\begin{aligned} \mathcal{H}_{\text{ent}}(\hat{z}, r) &= \sum_{p \in \text{Ent}(r)} \mathbf{1}[p \in \hat{z}] \\ \mathcal{H}_{\text{rel}}(\hat{z}, r) &= \sum_{p \in \text{Rel}(r)} \mathbf{1}[p \in \hat{z}] \end{aligned} \quad (2)$$

where $\mathbf{1}[\cdot]$ is an indicator function that equals 1 if phrases p appears in $\hat{z}$, and 0 otherwise. For the special case $r = \text{NO_RELATION}$, we set $\mathcal{H}_{\text{ent}} = \mathcal{H}_{\text{rel}}$. The total weighted hits are given by:

$$\mathcal{H}(\hat{z}, r) = w_{\text{ent}} \cdot \mathcal{H}_{\text{ent}}(\hat{z}, r) + w_{\text{rel}} \cdot \mathcal{H}_{\text{rel}}(\hat{z}, r) \quad (3)$$

with two hyperparameters w_{ent} and w_{rel}. Let $|\hat{z}|$ denote the number of words in $\hat{z}$, normalized by a hyperparameter of N. Hyperparameter studies are in Section 5.6. The final score is defined as:

$$\mathcal{S}(\hat{z}, r) = \frac{\mathcal{H}(\hat{z}, r)}{|\hat{z}|/N} \quad (4)$$

Finally, the overall HIT@DICT reward aggregates the contributions from both sentences s_1 and s_2, calculated as $\mathcal{R}_{\text{HIT@DICT}} = (\mathcal{S}(\hat{z}, r_1) + \mathcal{S}(\hat{z}, r_2))/2$. Here r_1 and r_2 denote the ground truth relation labels for s_1 and s_2. HIT@DICT scoring example in Figure 1.

How to construct a dictionary of relational phrases? Unlike human-crafted phrases or rules, these phrases are automatically derived from explanations included in the outputs of vanilla LLMs. In particular, we start by sampling all positive items for which the support and test sentences share the same relation label. Then, the vanilla model answers all the positive items, and we filter the true positive items with the final answer `Yes`. For each label that appears in the training dataset, we randomly sample one to five LLM-generated explanations. These relation labels, combined with their associated explanation cases, are input into a LLM,[1] which is prompted to extract the representative relational phrases from these cases. Additionally, each relation label is decomposed into phrases as part of the relational phrases. After text post-processing, these relation labels and their associated phrases are added to the relation phrase dictionary. See Alg. 1 and Figure 1 for details and an example.

Accuracy Reward. We introduce the accuracy reward $\mathcal{R}_{\text{Acc}}(\hat{y}, y)$ to evaluate the correctness of the predicted label. In one-shot settings, each test sentence is matched with K supports, with at most one positive and the rest negative, leading to a 1:K imbalance. Following (Lin et al., 2019), to counter this, we weigh rewards for YES predictions more heavily, encouraging the model to align with the task's inherent imbalance. The specific reward values are chosen based on these intuitive observations and are fixed throughout training:

$$\mathcal{R}_{\text{Acc}}(\hat{y}, y) = \begin{cases} 3.0, & \text{if } \hat{y} = \text{Yes} \wedge y = \text{Yes} \\ 1.0, & \text{if } \hat{y} = \text{No} \wedge y = \text{No} \\ -3.0, & \text{if } \hat{y} = \text{Yes} \wedge y = \text{No} \\ -1.0, & \text{if } \hat{y} = \text{No} \wedge y = \text{Yes} \\ 0.0, & \text{otherwise} \end{cases} \quad (5)$$

Training Process. We optimize COGRE with Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Formally, given a group of m explanation and label pairs $\mathcal{O} = \{(\hat{z}_i, \hat{y}_i) \mid i \in [m]\}$ sampled from COGRE for the same input (s_1, s_2), we assign each pair a reward $R(\hat{z}_i, \hat{y}_i)$ using our designed function $\mathcal{R}$. GRPO encourages relative

[1] We test both Qwen2.5-14B-Instruct and GPT-4o for phrase extraction, obtaining identical results (Appendix A.4).

Algorithm 1 Building the Relational Phrase Dictionary

Require: Training set $\mathcal{D}_{\text{train}} = \{(s_1, s_2, r_1, r_2, y)\}$; vanilla LLM $\mathcal{M}$; GPT-4o API; sample size per label $K \in \{1, \ldots, 5\}$
Ensure: Phrase dictionary Dict : relation_label $\mapsto$ phrase list

$\mathcal{C} \leftarrow \{(s_1, s_2, r_1, r_2, y) \in \mathcal{D}_{\text{train}} \mid r_1 = r_2 \ \wedge \ y = \text{“Yes”}\}$ ▷ Pairs with identical relation labels
$\mathcal{G} \leftarrow \emptyset$ ▷ Good cases predicted correctly by the vanilla LLM
for all $(s_1, s_2, r_1, r_2, y) \in \mathcal{C}$ **do**
  $(\hat{z}, \hat{y}) \leftarrow \mathcal{M}(s_1, s_2)$ ▷ Vanilla LLM inference: explanation & label
  if $\hat{y} =$ “Yes” **then**
    $\mathcal{G} \leftarrow \mathcal{G} \cup \{(s_1, s_2, r_1, \hat{z})\}$ ▷ Keep good cases
  end if
end for
Group $\mathcal{G}$ by relation label: for each $r \in \mathcal{R}$, let $\mathcal{G}_r = \{(s_1, s_2, r, \hat{z}) \in \mathcal{G}\}$
Dict $\leftarrow \emptyset$
for all $r \in \mathcal{R}$ **do**
  $\mathcal{S}_r \leftarrow \text{SampleUpTo}(\mathcal{G}_r, K)$ ▷ Randomly sample $1 \sim 5$ good cases per label
  $\text{prompt}_r \leftarrow \text{BuildPrompt}(r, \ \mathcal{S}_r)$ ▷ Prompt includes the label and several examples
  $\text{phrases}_r \leftarrow \text{GPT-4o}(\text{prompt}_r)$ ▷ Generate relational phrases list
  $\text{labelPhrases}_r \leftarrow \text{Tokenize}(\text{label}_r)$ ▷ Decompose the relation label into phrases
  $\text{Dict}[r] \leftarrow \text{PostProcess}(\text{phrases}_r \cup \text{labelPhrases}_r)$ ▷ Lowercasing, dedup, stemming/lemmatization, stopword removal
end for
return Dict

improvements within a group by normalizing each reward against the group mean. Specifically, the *group-relative advantage* of the i-th explanation–label pair is defined as:

$$\mathcal{A}_i = \frac{\mathcal{R}(\hat{z}_i, \hat{y}_i) - \frac{1}{m}\sum_{j=1}^{m} \mathcal{R}(\hat{z}_j, \hat{y}_j)}{\text{std}(\{\mathcal{R}(\hat{z}_j, \hat{y}_j) \mid j \in [m]\})}, \quad (6)$$

where std$(\cdot)$ is the standard deviation of group rewards. The overall GRPO objective is optimized to maximize a clipped function with a KL penalty:

$$\mathcal{L}(\theta) = \mathbb{E}_{(\hat{z}_i, \hat{y}_i) \sim \mathcal{O}} \Big[\min \Big(\rho_i \, \mathcal{A}_i, \\ \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon) \, \mathcal{A}_i \Big) - \beta \, \text{KL}(\theta \,\|\, \theta_{\text{ref}}) \Big]. \quad (7)$$

ρ_i is the importance ratio between updated and old policy probabilities, ϵ controls the clipping range, and β weights the penalty for diverging from a reference model θ_{ref}.

4 Experimental Setup

Benchmark. Our work focuses on realistic Few-Shot Relation Extraction (FSRE) (Alam et al., 2024), which predicts relation labels for test sentences from several support sentences without predefined labels or descriptions. This setting is more challenging than standard FSRE due to minimal supervision (see Appendix A.1). We follow this realistic FSRE and conduct experiments on *one-shot TACRED, NYT29, and WIKIDATA*. The relation labels in the training and testing partitions are out-of-distribution. Besides, since traditional RE methods rely on small classifiers, RE benchmarks are built to be extremely large. Following previous work (Li et al., 2023), we randomly sampled 1,000 episodes per partition according to the original label distribution (Statistics in Appendix A.6).

Evaluation. We adopt a dual evaluation protocol containing both automatic and human evaluation. We use the F1 score as the automatic evaluation metric. For human evaluation, we rated explanations on a 3-point Likert scale: 2 points for the correctness and conciseness of the summaries, plus 1 point if it aligns with the RE labeling. The detailed evaluation rubric is provided in Appendix A.2. Two annotators with NLP backgrounds independently rated the sampled explanations. The Cohen's kappa score is 0.693, indicating substantial agreement and that our evaluation rubric is well-defined.

Baselines. We follow the evaluation method in Realistic FSRE and compare our method with two categories of baselines. ***Conventional RE Models***: Semantic Rule Matcher (Vacareanu et al., 2024), which combines a neural classifier with rules, achieving state-of-the-art results on realistic Few-Shot TACRED and NYT29. ***Prompting RE baselines***: (i) SUMASK (Li et al., 2023) reformulates relation extraction as a question answering task. We implement SUMASK under realistic FSRE constraints (no label) ; see Appendix A.9.1. (ii) Naive prompting: we include *direct-matching* (i.e., only producing `Yes`/`No`) and *simple-reasoning* (i.e., reasoning before `Yes`/`No`); see Appendix A.9.

Implementation Details. We sample 20,000 items from the training partition, preserving the distribution of relation labels and maintaining an approximate 1:7 ratio between positive and negative

Method	One-shot TACRED			One-shot NYT29			Avg.		
	Prec%	Recall%	F1%	Prec%	Recall%	F1%	Prec%	Recall%	F1%
Semantic Rule Matcher	32.45	19.72	24.52	22.23	13.45	16.76	27.34	16.59	20.64
SUMASK *(Phi-4)*	4.44	31.71	7.78	10.96	26.13	15.44	7.70	28.92	11.61
Phi-4									
Direct Matching	5.43	26.83	9.03	9.04	23.15	13.00	7.24	24.99	11.02
Simple Reasoning	5.69	58.54	10.38	8.71	30.68	13.57	7.20	44.61	11.98
CogRE (*ours*)	22.53	50.00	31.06	12.03	30.68	17.28	17.28	40.34	24.17
CogRE After RL with Acc	26.90	47.56	34.36	45.50	37.36	41.02	36.20	42.46	37.69
CogRE After RL with Hit@Dict + Acc	26.88	60.98	37.31	45.14	44.89	45.01	36.01	52.94	41.16
Qwen2.5-14B-Instruct									
Direct Matching	13.33	2.44	4.12	48.73	13.63	21.31	31.03	8.04	12.72
Simple Reasoning	5.67	34.15	9.72	11.85	34.23	17.61	8.76	34.19	13.67
CogRE (*ours*)	29.49	28.05	28.75	20.18	31.67	24.65	24.84	29.86	26.70
CogRE After RL with Acc	26.83	40.24	32.20	26.17	29.40	27.69	26.50	34.82	29.95
CogRE After RL with Hit@Dict + Acc	22.08	62.20	32.58	63.34	38.78	48.11	42.71	50.49	40.34

Table 2: Precision (P), recall (R), and F1 on the one-shot TACRED and one-shot NYT29 datasets. We split the table into three blocks: baseline methods; prompting methods without RL, including vanilla prompting and our CogRE framework; and models trained with RL on top of the same CogRE prompting, with the accuracy reward alone or combined with Hit@Dict. We use fixed $N = 5$, $w_{\text{ent}} = 0.4$, $w_{\text{rel}} = 1.0$; hyperparameter experiments in Table 5. The blue color highlights F1 scores, with darker (from to) shades indicating larger values.

instances (statistics in Appendix A.5). We implement our method with Qwen2.5-14B-Instruct and Phi-4 using fixed reward hyperparameters, $N = 5$, $w_{\text{ent}} = 0.4$, and $w_{\text{rel}} = 1.0$ in the main experiments reported in Table 2. To assess robustness, we further evaluate our method under different reward hyperparameter settings, with $N \in \{3, 5, 7\}$, $w_{\text{ent}} \in \{0.2, 0.4, 0.6\}$, and $w_{\text{rel}} \in \{0.8, 1.0, 1.2\}$, and report the results in Table 5. We optimize the model using Verl (Sheng et al., 2025) with an actor learning rate of 1×10^{-6}, KL regularization (coefficient 0.01). Training is conducted on 4 NVIDIA H100 80 GB GPUs. A complete run on the 14B–15B model takes 20 GPU-hours.

5 Experimental Results

5.1 Main Results

Table 2 reports the main results under different reasoning designs and RL settings, using either accuracy reward alone or combined with Hit@Dict.

CogRE improves accuracy with balanced precision and recall. Our CogRE consistently outperforms all baselines with a better precision–recall trade-off. In contrast, the Semantic Rule Matcher (the previous SOTA), based on rules and a small language model, yields high precision but lower recall. Prompting-based LLM baselines rely solely on LLMs' generalization ability, leading to high recall but lower precision. CogRE combines both perspectives: it anchors reasoning with phrases while leveraging LLMs' generalization through summarization and integrative reasoning.

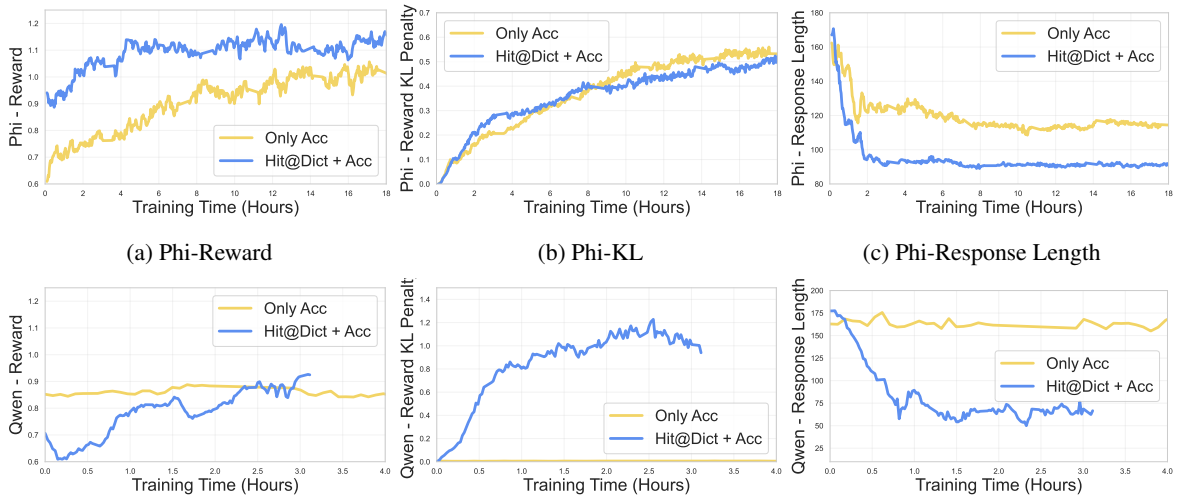
Hit@Dict reward improves task accuracy. Table 2 shows that Qwen2.5-14B-Instruct, trained only with an accuracy reward, surpasses its non-trained version by +3.45% and +3.04%, while the trained Phi-4 leads to +3.30% and +23.74% improvements. Hit@Dict + Acc reward further boosts accuracy across models, which outperforms accuracy-only training and achieves the best results. In particular, Qwen2.5-14B-Instruct achieves F1=48.11% with Hit@Dict + Acc, a 73.74% relative gain over the accuracy-only version. We also evaluate the quality of the explanations (see § 5.5).

5.2 Behavior of Hit@Dict Reward

We analyze the impact of the Hit@Dict and accuracy rewards during the RL training process. Figure 2 shows the RL training process of both Phi-4 and Qwen2.5-14B-Instruct on the one-shot NYT29.

Hit@Dict accelerates the convergence of training. In Figure 2(a), Hit@Dict +*Acc* curve rapidly increases to above 1.1 within 5 hours and stabilizes between 1.1–1.2 after 10 hours; in contrast, the *Only Acc* curve climbs slowly from 0.6 and stabilizes around 1.0 after 13 hours.

Hit@Dict provides more stable training. As shown in Figure 2 (d) and (e), Qwen2.5-14B-Instruct, trained only with the accuracy reward, exhibits stagnant reward values and an extremely small reward–KL penalty, indicating the policy re-



(a) Phi-Reward (b) Phi-KL (c) Phi-Response Length

(d) Qwen-Reward (e) Qwen-KL (f) Qwen-Response Length

Figure 2: Training dynamics on the one-shot NYT29 dataset for Phi-4 and Qwen2.5-14B-Instruct. The X-axes show training time(Hours). The Y-axes show reward, KL penalty, and response length. We compare reinforcement learning with accuracy reward *Only Acc* and with the combined HIT@DICT reward HIT@DICT *+Acc*. We use fixed hyperparameters with $N = 5$, $w_{\text{ent}} = 0.4$, and $w_{\text{rel}} = 1.0$.

Method	One-shot TACRED			One-shot NYT29		
	Prec%	Recall%	F1%	Prec%	Recall%	F1%
COGRE	22.53	50.00	31.06	12.03	30.68	17.28
- *w/o chunking*	7.05	57.68	12.56	8.20	27.41	12.63
- *w/o phrases*	16.10	52.44	24.64	10.39	31.39	15.62
- *w/o reasoning*	5.94	28.05	9.81	9.91	24.57	14.12

Table 3: Ablation studies on one-shot TACRED and NYT29 with Phi-4. Darker shades indicate higher F1.

mains almost unchanged from its initial state. In contrast, Qwen2.5-14B-Instruct, trained with the HIT@DICT reward, shows steady growth.

HIT@DICT encourages more concise explanations. As Figure 2 (c) and (f) show, with the HIT@DICT reward, the response length compresses to 75–90 tokens. Combined with our analysis of the explanations, we found that in most cases, the models produce more concise explanations. It enhances reasoning efficiency with concise outputs.

5.3 Ablation Study

We analyze the effectiveness of each component in COGRE by removing one step at a time:(1) *w/o chunking*: removes chunking; (2) *w/o phrases*: removes phrase anchoring; (3) *w/o reasoning*: removes reasoning. Results are shown in Table 3, from which we draw three observations. *First*, all three steps contribute to the final performance. Removing any step from our framework causes a clear performance drop, ranging from –1.66% to –21.25%. *Second*, chunking and phrase anchoring mainly improve precision. Removing either lowers

Method	One-shot WIKIDATA		
	Prec%	Recall%	F1%
Semantic Rule Matcher	6.54	6.08	6.31
SUMASK *(Phi-4)*	1.48	27.45	2.80
Direct Matching	1.13	49.02	2.22
Simple Reasoning	1.99	56.86	3.85
COGRE *(ours)*	5.08	52.94	9.28
COGRE After RL with Acc	14.29	31.37	19.63
COGRE After RL with HIT@DICT + Acc	37.84	27.45	31.81

Table 4: Out-of-domain evaluation for Phi-4 trained on the one-shot NYT29 (same as Table 2) on the one-shot WIKIDATA. Darker shades indicate higher F1.

precision on both datasets (–15.48% and –3.83% for chunking; –6.43% and –1.64% for phrase anchoring), while recall is far less affected and even increases on TACRED (+7.68% and +2.44%). This indicates that both steps suppress spurious matches rather than retrieve additional ones. *Third*, reasoning improves both precision and recall, with a larger impact on recall. Removing it substantially reduces recall (–21.95% on TACRED; –6.11% on NYT29) alongside precision (–16.59% and –2.12%).

5.4 Out-of-Distribution Generalization

These Phi-4 checkpoints trained on the one-shot NYT29 are further evaluated on the one-shot WIKIDATA. Results (Table 4) show that COGRE consistently outperforms baseline approaches in F1 score. RL training with the accuracy-only reward improves F1 to 19.63% on Phi-4. Further optimization with our HIT@DICT + Acc yields substantial additional gains, improving F1 by +12.18% on Phi-4. In contrast, when transferring from NYT29 to

WIKIDATA, model performance with other methods drops substantially. For example, Phi-4 exhibits F1 drops of 12.64% with SUMASK, 10.78% with direct matching, and 9.72% with simple reasoning. We also conduct a qualitative analysis of explanation quality on WIKIDATA, showing that COGRE after RL with HIT@DICT + Acc generates more concise explanations that align better even with unseen RE relation labels. These findings demonstrate that our method generalizes robustly to datasets with unseen relation labels.

5.5 Error Analysis of LLM on RE-reasoning

Simple reasoning strategy. We analyze the explanations of vanilla LLMs using a random reasoning strategy. For this stage, we select Qwen2.5-14B-Instruct and GPT-4o. From their explanations, we identify two common failure patterns. *First*, failing to focus on semantics that truly convey the relation. When matching two sentences, LLMs frequently focus on irrelevant tokens in the second sentence, aiming to align with the relation conveyed in the first. We provide a simplified example in Table 1. In this case, LLMs incorrectly focus on two names in the second sentence in order to mimic the relation of PER:ALTERNATE_NAME in the first sentence. *Second*, failing to align with the granularity of the RE human-annotation schema. Without the human-crafted descriptions of relation labels, LLMs struggle to distinguish between similar human-defined relations, e.g., ORG:COUNTRY_OF_HEADQUARTERS and ORG:CITY_OF_HEADQUARTERS.

COGRE. We evaluate the quality of LLM-generated explanations at three stages: (1) vanilla LLMs with the COGRE, and after RL training, (2) with only an accuracy reward, and (3) with both HIT@DICT and accuracy reward. We validate Qwen2.5-14B-Instruct and Phi-4 on two datasets. For each LLM–dataset–stage combination, we sample 40 explanations, with 10 explanations per category (TP, TN, FP, FN). Results (Appendix A.3) show that human evaluation scores improve by 54.24% (relative). Compared with vanilla and accuracy-only, HIT@DICT combined with accuracy reward enables more concise summaries and better alignment with human annotations. For example, in the Phi–TACRED setting, among the 40 analyzed cases, the model trained with the HIT@DICT + Acc produced more concise summaries in 8 cases and exhibited better alignment with human labeling in 15 cases. Specifically, the trained model tends to include relational phrases closely aligned with gold labels in their explanation (e.g., *enroll*, *attend*, and *university* for the relation PER:SCHOOLS_ATTENDED), while the untrained model generates vague terms such as *associated*. See some case comparisons in Appendix A.12.

No.	Hyperparameters			One-shot NYT29		
	N	w_{ent}	w_{rel}	Prec%	Recall%	F1%
1	5	0.4	1.0	63.34	38.78	48.11
2	3	0.4	1.0	43.75	56.68	49.38
3	7	0.4	1.0	44.06	52.13	47.76
4	5	0.2	1.0	45.73	51.70	48.53
5	5	0.6	1.0	45.61	48.72	47.12
6	5	0.4	0.8	47.34	48.01	47.67
7	5	0.4	1.2	35.05	72.58	47.27

Table 5: Hyperparameters sensitivity on one-shot NYT29 with Qwen2.5-14B-Instruct. We test seven configurations by varying w_{ent}, w_{rel}, N individually.

5.6 Hyperparameter Sensitivity Analysis

We analyze the sensitivity of our method to its hyperparameters, including the entity weight w_{ent}, the relation weight w_{rel}, and the normalization factor N. We vary $N \in \{3, 5, 7\}$, $w_{\text{ent}} \in \{0.2, 0.4, 0.6\}$, and $w_{\text{rel}} \in \{0.8, 1.0, 1.2\}$. To ensure a fair comparison, all hyperparameter experiments are conducted using Qwen2.5-14B-Instruct on the same NYT29 dataset. Results under all seven configurations are reported in Table 5. We find that the model's performance remains stable across these seven hyperparameter configurations. The F1 score remains largely consistent at 47.98 ± 0.78. This small variance indicates that our method is robust to changes in hyperparameter configurations.

6 Conclusion

In this work, we introduced COGRE, a RE reasoning framework loosely inspired by cognitive science, which decomposes RE into three human-like text-processing steps. To incentivize LLM reasoning to align with RE labeling schemas via RL, we proposed HIT@DICT, a lightweight reward strategy that derives reasoning supervision from the model's own correct predictions. Experiments and human evaluations on one-shot TACRED and NYT29 demonstrate that our framework achieves enhanced accuracy and explanation quality. Human analysis confirms that our reward design encourages models to generate more concise and label-aligned reasoning. Future work will extend HIT@DICT to more reasoning-intensive tasks.

Limitations

This work does not investigate the internal causal mechanisms of LLMs, but instead focuses on improving their textual reasoning process. In particular, we do not analyze internal model representations or design methods that directly improve reasoning through representation-level interventions. Mechanistic interpretability approaches, such as sparse autoencoders, may provide useful tools for such analysis. Our work instead focuses on improving explainability through textual reasoning, leaving mechanistic interpretability for future work.

Ethical Considerations

This work does not involve human subjects research beyond standard human evaluations. All datasets, and pretrained models used in this work are publicly available and used in accordance with their intended research use and applicable licenses. We use public relation extraction benchmarks and do not collect new personal data. We provide dataset descriptions, evaluation settings, and data statistics in Section 4 and Appendix. AI-assistants were only used to polish the writing, such as improving grammar, readability. All technical ideas, experiments, analysis, and final verification were conducted by the authors.

Acknowledgments

Mihai Surdeanu declares a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

References

Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE accError Analysisess*, 6:52138–52160.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *Preprint*, arXiv:2402.00157.

Fahmida Alam, Md Asiful Islam, Robert Vacareanu, and Mihai Surdeanu. 2024. Towards realistic few-shot relation extraction: A new meta dataset and evaluation. *Preprint*, arXiv:2404.04445.

Razvan Bunescu and Raymond Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 724–731.

Xu Chu, Zhijie Tan, Hanlin Xue, Guanyu Wang, Tong Mo, and Weiping Li. 2025. Domaino1s: Guiding llm reasoning for explainable answers in high-stakes domains. *Preprint*, arXiv:2501.14431.

Runpeng Dai, Tong Zheng, Run Yang, Kaixian Yu, and Hongtu Zhu. 2025. R1-re: Cross-domain relation extraction with rlvr. *Preprint*, arXiv:2507.04642.

John H Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34:906–911.

Shen Gao, Zhengliang Shi, Minghang Zhu, Bowen Fang, Xin Xin, Pengjie Ren, Zhumin Chen, Jun Ma, and Zhaochun Ren. 2023. Confucius: Iterative tool learning from introspection feedback by easy-to-difficult curriculum. *Preprint*, arXiv:2308.14034.

Bryce Goodman and Seth Flaxman. 2017. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Haoyang He, Zihua Rong, Kun Ji, Chenyang Li, Qing Huang, Chong Xia, Lan Yang, and Honggang Zhang. 2025. Rethinking reasoning quality in large language models through enhanced chain-of-thought via rl. *Preprint*, arXiv:2509.06024.

Hui Huang, Yancheng He, Hongli Zhou, Rui Zhang, Wei Liu, Weixun Wang, Wenbo Su, Bo Zheng, and Jiaheng Liu. 2025. Think-j: Learning to think for generative llm-as-a-judge. *Preprint*, arXiv:2505.14268.

Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji, and Jiawei Han. 2023. Instruct and extract: Instruction tuning for on-demand information extraction. *Preprint*, arXiv:2310.16040.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *Proceedings of the ACL interactive poster and demonstration sessions*, pages 178–181.

Walter Kintsch. 1988. The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review*, 95:163.

Walter Kintsch and Teun A Van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review*, 85:363–394.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023. Revisiting large language models as zero-shot relation extractors. *Preprint*, arXiv:2310.05028.

Enlu Lin, Qiong Chen, and Xiaoming Qi. 2019. Deep reinforcement learning for imbalanced classification. *Preprint*, arXiv:1901.01379.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. 2024. Improve mathematical reasoning in language models by automated process supervision. *Preprint*, arXiv:2406.06592.

Xilai Ma, Jing Li, and Min Zhang. 2023. Chain of thought with explicit evidence reasoning for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2334–2352.

George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.

Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, Keyu Chen, Ming Li, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, and 1 others. 2024. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. *arXiv preprint arXiv:2409.02387*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2024. Adelie: Aligning large language models on information extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387.

Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, and Tianlu Wang. 2025. Learning to plan & reason for evaluation with thinking-llm-as-a-judge. *Preprint*, arXiv:2501.18099.

Hamed Shahbazi, Xiaoli Z. Fern, Reza Ghaeini, and Prasad Tadepalli. 2020. Relation extraction with explanation. *Preprint*, arXiv:2005.14271.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, page 1279–1297. ACM.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.

Zheng Tang and Mihai Surdeanu. 2023. It takes two flints to make a fire: Multitask learning of neural relation and explanation classifiers. *Computational Linguistics*, 49:117–156.

Robert Vacareanu, Fahmida Alam, Md Asiful Islam, Haris Riaz, and Mihai Surdeanu. 2024. Best of both worlds: A pliable and generalizable neuro-symbolic approach for relation classification. *Preprint*, arXiv:2403.03305.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2024. Revisiting relation extraction in the era of large language models. *Preprint*, arXiv:2305.05003.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. Gpt-re: In-context learning for relation extraction using large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 3534–3547.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, Haosheng Zou, Yongchao Deng, Shousheng Jia, and Xiangzheng Zhang. 2025. Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond. *Preprint*, arXiv:2503.10460.

Rihui Xin, Han Liu, Zecheng Wang, Yupeng Zhang, Dianbo Sui, Xiaolin Hu, and Bingning Wang. 2025. Surrogate signals from format and length: Reinforcement learning for solving mathematical problems without ground truth answers. *Preprint*, arXiv:2505.19439.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. How to unleash the power of large language models for few-shot relation extraction? *Preprint*, arXiv:2305.01555.

Minglai Yang, Xinyu Guo, Utkarsh Tyagi, Mian Zhang, Razvan Dumitru, Sunjie Hou, Yunzhong He, Daniel Yue Zhang, and Ying Liu. 2026. Rubric dropout: A simple way to mitigate reward hacking in rubric-as-reward rl. *Preprint*, arXiv:2608.11669.

Nicolas Yax, Hernán Anlló, and Stefano Palminteri. 2024. Studying and improving reasoning in humans and machines. *Communications Psychology*, 2(1):51.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of machine learning research*, 3:1083–1106.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pages 2335–2344.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.

Minghang Zhu, Chuyang Wei, Junhao Xu, Yilin Cheng, Zhumin Chen, and Jiyan He. 2026. Deeprubric: Evidence-tree rubric supervision for efficient reinforcement learning of deep research agents. *arXiv preprint arXiv:2606.17029*.

A Appendix

Content of Appendix

A.1 Realistic Few-Shot Relation Extraction
A.2 Human Evaluation Rubric
A.3 Human Evaluation Results
A.4 Dictionary Robustness Analysis
A.5 Statistics of the Sampled Training Set
A.6 Statistics of the Sampled Testing Set
A.7 Prompt for CogRE framework
A.8 Prompt for Relation Phrases Extraction
A.9 Prompt for Baselines
A.9.1 Prompts for SUMASK
A.9.2 Direct-Answer Prompt
A.9.3 Simple-Reasoning Prompt
A.10 Prompt for Ablation experiments
A.10.1 COGRE- w/o chunking Prompt
A.10.2 COGRE- w/o reasoning Prompt
A.10.3 COGRE- w/o phrases Prompt
A.11 Inference Results across Model Families and Sizes on realistic One-shot RE task
A.12 Cases Comparison of Phi-4 on TACRED
A.13 Comparison of Extracted Dictionaries by Open-source and Closed-source LLMs

A.1 Realistic Few-Shot Relation Extraction

The Relation Extraction (RE) task studied in this paper is the realistic Few-Shot Relation Extraction (FSRE) task (Alam et al., 2024). This task is fundamentally different from the standard FSRE setting; it is harder and more realistic. In realistic FSRE, the model infers a relation by comparing a test sentence against several support sentences *without predefined labels or label descriptions*. This reflects real-world RE where many relations lack manual descriptions, sentences are unlabeled, and response time is critical.

A.1.1 Framework of realistic few-shot RE setting

As shown in Figure 3 for each relation, we provide one support sentence as an example of that relation. Given a test sentence, the model compares it against each support sentence individually and outputs *Yes* or *No*. If all outputs are *No*, the predicted label is *no_relation*. If at least one output is *Yes*, the corresponding relation is selected. The final prediction is then compared with the gold label to compute the F1 score.

A.1.2 Characteristics of realistic few-shot RE setting

This task setting has the following key characteristics: **(1) it is a multi-class Relation Extraction task**, **(2) it follows a realistic few-shot RE setting**, and **(3) it is substantially more challenging than traditional multi-class RE**. We provide a detailed explanation below.

(1) Multi-class Relation Extraction. The RE task considered in this paper is a multi-class relation classification problem. Although our method adopts a pairwise matching formulation as an intermediate step, the overall task remains multi-class RE, and evaluation is based on whether the predicted relation matches the gold label.

This pairwise formulation is common in few-shot relation extraction. For example, SUMASK (Li et al., 2023) compares a test sentence with each relation triplet by generating relation-specific questions and making one-by-one decisions. Similarly, Semantic Rule Matcher (Vacareanu et al., 2024) performs pairwise matching between support and test sentences before producing a final multi-class prediction.

(2) A harder setting than traditional multi-class RE. We argue that this support-based few-shot RE setting is more challenging than traditional multi-class relation classification for several reasons.

Absence of explicit label semantics. In traditional multi-class RE, relation labels and their semantic descriptions are explicitly provided. In contrast, in our setting the relation semantics are not given directly and must be inferred from the support sentences. As a result, the model must first abstract a relational concept from the support example and then compare it with the concept inferred from the test sentence. This leads to two common failure modes discussed in Section 5.5 (Error Analysis): (i) failure to focus on relation-specific semantic cues, and (ii) interdependent errors during relation abstraction across sentences.

Local rather than global relation semantics. In traditional multi-class RE, the model has access to the entire label set, which facilitates fine-grained discrimination between relations. In contrast, our setting exposes the model only to a single support example at a time, resulting in local and incomplete relation semantics. For example, consider the relations *org:country_of_headquarters*

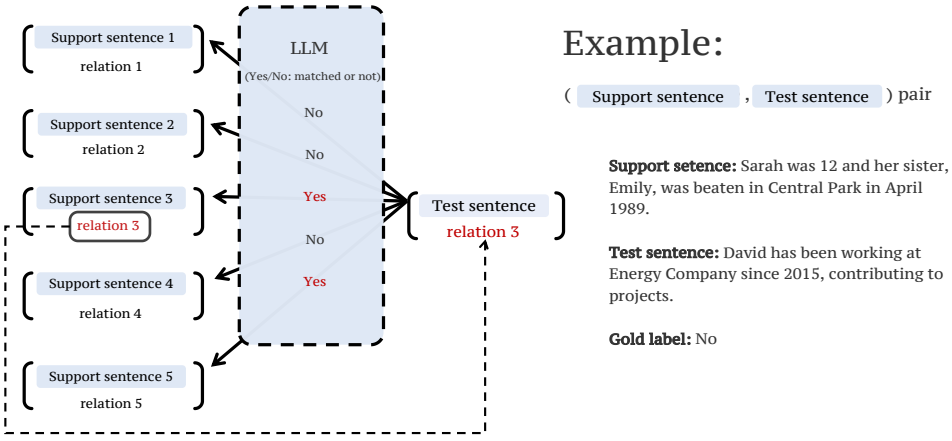


Figure 3: Illustration of the realistic few-shot relation extraction setting. Pairwise support sentence and test sentence matching results are aggregated to produce a multi-class prediction.

and *org:city_of_headquarters*. In a traditional multi-class setup, both labels are present, enabling direct comparison. However, in our setting the model may only observe a support sentence expressing *org:country_of_headquarters*. When the test sentence expresses *org:city_of_headquarters*, the model lacks explicit label definitions or global context, and thus tends to abstract a coarse concept such as *place_of_headquarters*, leading to confusion between the two.

Empirical evidence of increased difficulty. This support-based matching formulation has been shown to be challenging in prior work, where overall performance is reported to be low (Alam et al., 2024). As shown in Appendix A.11 and Table 21, we evaluate a wide range of models under this setting, including 3B–7B models (Qwen2.5-3B/7B, Phi-4-mini-Instruct, Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.2) and large mainstream LLMs (DeepSeek-V3, GPT-4o, GPT-3.5-Turbo). We observe that these models perform poorly under our setting when using either *Random Reasoning* or *Direct Matching*. Moreover, reasoning-based generation often increases over-confidence, leading to more false positives. These results demonstrate that realistic few-shot RE remains challenging for current LLMs.

A.2 Human Evaluation Rubric

Main Rubric. To assess the quality of model-generated explanations, we adopt a human evaluation rubric that emphasizes both concise summarization and alignment with RE labeling. The rubric assigns scores based on three major criteria: (1) the correctness and conciseness of the summarization for the support sentence, ensuring that the key relational information is accurately captured without including irrelevant details; (2) the correctness and conciseness of the summarization for the test sentence, evaluated under the same principles; and (3) the alignment of the explanation with the labeling of whether the relations expressed in the two sentences match or not, with any abstraction error or illogical reasoning resulting in a deduction. Each explanation is scored on a 3-point scale, with details provided as follows.

Human evaluation rubric for explanation quality (max score: 3).

```
[1 point]
    A correct and concise summarization
    of the support sentence is awarded 1
     point.

[1 point]
    A correct and concise summarization
    of the test sentence is awarded 1
    point.

[1 point]
    A reasonable explanation of whether
    relation_1 and relation_2 match or
    do not match is awarded 1 point. Any
     abstraction error results in the
    loss of this point.

 Special handling:
    - If summarization is incorrect but
    the explanation is logically
    reasonable, the third point can
    still be awarded.
    - Points are deducted when the model
     confuses similar relations,
      e.g., "city" vs. "country", or "
      reference" vs. "alternate_name".
```

Common types of abstraction errors. In addition to the main rubric, we present several common types of abstraction errors to help graders develop a clearer understanding of how such errors should be identified and penalized. These error types serve as practical references to ensure consistent and fair scoring. The detailed description and illustrative examples are provided for each case as follows.

Common Abstraction Errors for Human Rating

```
Abstraction Error:
    If two sentences express the same
    relation, they must be abstracted in
     the same direction and at the same
    level.
    Example:
    ``He, 12-years-old, got a good offer
    .'' and ``Jam is 12.''
    - Correct: per:age; per:age  [0
    points]
    - Incorrect: per:age; per:number
    [-1 point]

several common types:
    - Lack of Higher-Level Deductive
    Abstraction:
        Description: When a higher-level
         deductive abstraction is
        required to align the relations,
         failing to apply it leads to
        error.
        Example:
            ``STX Finland is part of the
             international STX Europe
            Group'' and
            ``Merck will acquire all of
            Millipore.''
            - Correct: org:parents; org:
            parents  [0 points]
            - Incorrect: org:parents;
            org:transaction  [-1 point]

    - Over- or Under-Focusing on Details
    :
        Description: The abstraction
        direction and level are correct,
         but the model misjudges due to
        being overly detailed or overly
        general.
        Example:
            ``The arrangement of
            financing for Millipore Corp
             in the US'' and
            ``Burlington Northern Santa
            Fe Corp is the biggest bet
            yet on a US economic
            recovery.''
            - Correct: org:
            country_of_headquarters; org
            :country_of_headquarters  [0
             points]
            - Incorrect: First = org:
            financial_transaction,
            Second = org:economic_event
             [-1 point]
```

Rubric Reliability Verification via Cohen's Kappa. Furthermore, to verify the reliability of our rubric, we engaged two independent annotators. Both annotators had NLP research backgrounds. They first evaluated the explanations in the setting of Phi-4 on the one-shot TACRED dataset. Each annotator followed the rubric, scoring explanations based on the three criteria and abstraction error types. This independent evaluation enabled us to measure the consistency between annotators. Specifically, for the Phi–TACRED setting, we computed the Cohen's kappa coefficient to quantify inter-annotator agreement. The resulting kappa value was 0.693, indicating substantial agreement. This shows that our rubric is well-defined and practical. Therefore, we consistently adopted this rubric across all subsequent evaluations, including different stages and experimental settings.

A.3 Human Evaluation Results

To further evaluate explanation quality, we conducted human evaluation across three training stages: (i) vanilla LLMs with the CogRE framework, (ii) after RL training with the accuracy reward, and (iii) after RL training with the Hit@Dict reward.

We selected Qwen2.5-14B-Instruct and Phi-4 as base models and evaluated them on two datasets: one-shot TACRED and NYT29. For each LLM–dataset–stage combination, we sampled 40 explanations, with 10 explanations drawn from each of the four categories: true positives, true negatives, false positives, and false negatives. We consistently adopted the rubric we defined in the Appendix A.2. Therefore, for each category, the maximum score is 30 points.

We provide the results of human evaluation in Table 6. The results show that the models trained with the Hit@Dict reward consistently outperform both untrained models and accuracy-reward–trained models. For example, in the Phi–TACRED setting, the number of correct explanations in the no_yes and yes_no categories increases substantially after applying the Hit@Dict reward (from 24 to 29 and from 16 to 26, respectively). Similar improvements are observed in the Qwen–TACRED setting, particularly in the yes_no category (from 18 to 26).

Setting / Category	Stages		
	_untrained	_rl_ACC._reward	_rl_HIT@DICT _reward
Phi – TACRED			
no_yes	24	26	29
yes_no	16	18	26
yes_yes	27	27	29
no_no	22	22	27
Phi – NYT29			
no_yes	18	25	24
yes_no	11	16	27
yes_yes	19	21	22
no_no	21	25	26
Qwen – TACRED			
no_yes	22	22	24
yes_no	18	18	26
yes_yes	26	26	29
no_no	19	21	25
Qwen – NYT29			
no_yes	19	22	20
yes_no	8	5	22
yes_yes	19	14	21
no_no	13	17	28

Table 6: Human evaluation results across four settings (Phi/Qwen × TACRED/NYT29). Each cell reports the score of a category in one combination under human evaluation. The categories are defined as follows: *no_yes*: the ground truth is No but the model prediction is Yes; *yes_no*: the ground truth is Yes but the model prediction is No; *yes_yes*: both the ground truth and the model prediction are Yes; *no_no*: both the ground truth and the model prediction is No;

A.4 Dictionary Robustness Analysis

Building this dictionary is straightforward: it is small (each relation includes only about 10 phrases) and can be created either manually from a few true-positive explanations or automatically using small or large LLMs. We performed a drop-in replacement analysis by using Qwen2.5-14B instead of GPT-4o to extract phrases from explanations on the NYT29 dataset. We evaluate the effect from two aspects:

(1) Keyword Overlap. We compare the overlap rate between keyword sets extracted by GPT-4o and Qwen2.5-14B. The two models produce highly consistent keyword sets, indicating that the extraction process is not sensitive to the choice of LLM.

To measure the consistency between the keyword sets generated by GPT-4o and Qwen2.5-14B, we use the Overlap Coefficient. Given two keyword sets K_{GPT4o} and K_{Qwen}, the overlap is defined as:

$$\text{Overlap} = \frac{|K_{\text{GPT4o}} \cap K_{\text{Qwen}}|}{\min\left(|K_{\text{GPT4o}}|, |K_{\text{Qwen}}|\right)}. \quad (8)$$

Table 7 shows the phrases for the label */location/location/contains* separately by Qwen2.5-14B and GPT4o. We list all the phrases for all the labels separately by Qwen2.5-14B and GPT4o in Appendix A.13. In this case shown in the table, we have:

$$|K_{\text{GPT4o}} \cap K_{\text{Qwen}}| = 5,\ \min(|K_{\text{GPT4o}}|, |K_{\text{Qwen}}|) = 6, \quad (9)$$

.

Across all 15 relations in the NYT29 training set, the two models produce highly consistent keyword sets. In total, we observe:

$$|K_{\text{GPT4o}} \cap K_{\text{Qwen}}| = 80, \quad (10)$$

$$\min(|K_{\text{GPT4o}}|, |K_{\text{Qwen}}|) = 87, \quad (11)$$

which yields an overall overlap of:

$$\text{Overlap} = \frac{80}{87} = 0.91954. \quad (12)$$

(2) Performance of Our Method with a New Dictionary. We rebuild the entire dictionary using the phrases extracted by Qwen2.5-14B and re-run the RL training with Acc+HIT@DICT. The resulting F1 score remains very close to the GPT-4o-based setup, showing that our method is robust under drop-in replacement.

Qwen2.5-14B	**GPT4o**
/location/location/conta	*/location/location/contains*
"city"	"city"
"location"	"location"
"province"	"provincial"
"capital"	"capital"
"located"	"located"
	"contains"
"part"	

Table 7: The phrases for the label */location/location/contains* separately by Qwen2.5-14B and GPT4o. The yellow means the same phrases generated by both open-source and close-source models. We list all the phrases for all the labels separately by Qwen2.5-14B and GPT4o in Appendix A.13.

Dictionary Source	**P**	**R**	**F1**
GPT-4o	63.34	38.78	48.11
Qwen2.5-14B	45.35	51.28	48.13

Table 8: Performance of our method with dictionaries built from different LLMs. Comparable F1 scores indicate robustness to dictionary replacement.

A.5 Statistics of the Sampled Training Set

To ensure fairness and representativeness in one-shot relation extraction evaluation, we construct sampled training sets for both one-shot NYT29 and TACRED that preserve the distributional properties of the original datasets. The algorithm of sampling training data is shown in Algorithm 2.

Tables 9 and 10 report the ratio of positives to negatives in the training partition of one-shot NYT29 and TACRED, respectively. In these tables, we compare the ratio of positives to negatives, and the proportion of one of relation labels-NO_RELATION before and after sampling. For NYT29, it can be seen that the ratio of positive items (r, r), negative items with NO_RELATION, $(r, \text{NO_RELATION})$, and negative items without NO_RELATION, (r, r') in the sampled set aligns well with the original dataset. A similar pattern holds for TACRED, where the sampled data maintains the same balance across positive and negative categories.

In addition, Tables 11 and 12 further analyze the distribution of all the relation labels in the original dataset and the sampled dataset, respectively. Since we sampled the training data strictly according to the original distribution of relation labels, the sampled training datasets have a similar label distribution to the original ones. These statistics show that our sampled training datasets faithfully reflect the statistical properties of the original datasets, thereby avoiding biases introduced by over- or under-sampling specific relations.

	Original	Sampled
Positive items (r, r)	71376	2670
Negative items with no_relation $(r, \text{no_relation})$	571560	9660
Negative items without no_relation (r, r')	285504	7670

Table 9: Ratio of positives and negatives on the original training partition of one-shot NYT29 and our sampled version. *Positive items* (r, r): pairs where both sentences express the same relation r. *Negative items with no_relation* $(r, \text{no_relation})$: pairs where one sentence expresses a relation r and the other is labeled as no_relation. *Negative items without no_relation* (r, r'): pairs where the two sentences express different relations r and r', neither being no_relation.

	Original	Sampled
Positive items (r, r)	7170	2583
Negative items with no_relation $(r, \textit{no_relation})$	732075	14834
Negative items without no_relation (r, r')	28680	2583

Table 10: Ratio of positives and negatives on the original training partition of one-shot TACRED and our sampled version. *Positive items* (r, r): pairs where both sentences express the same relation r. *Negative items with no_relation* $(r, \text{no_relation})$: pairs where one sentence expresses a relation r and the other is labeled as no_relation. *Negative items without no_relation* (r, r'): pairs where the two sentences express different relations r and r', neither being no_relation.

Relation	**Original**	**Sampled**
/business/company/founders	52,201	1,634
/business/company/place_founded	51,408	1,571
/business/person/company	66,753	2,226
/film/film_location/featured_in_films	50,425	1,590
/location/country/capital	70,537	2,351
/location/location/contains	135,244	4,934
/location/us_county/county_seat	50,394	1,572
/location/us_state/capital	51,938	1,681
/people/deceased_person/place_of_burial	49,984	1,545
/people/ethnicity/geographic_distribution	51,557	1,608
/people/person/children	51,332	1,553
/people/person/ethnicity	50,625	1,566
/people/person/nationality	74,607	2,511
/people/person/place_lived	71,195	2,437
/people/place_of_interment/interred_here	50,240	1,561
no_relation	571,560	9,660

Table 11: Distribution of relation labels in the one-shot NYT29 training partition (original vs. sampled). Each row corresponds to a relation type, where the *Original* column reports the number of instances in the full dataset and the *Sampled* column reports the number of instances included in our one-shot sampled version. The relation NO_RELATION indicates sentence pairs that do not express any annotated relation.

Relation	Original	Sampled
org:alternate_names	31,532	1,156
org:city_of_headquarters	31,016	997
org:dissolved	30,011	912
org:members	30,038	968
org:number_of_employees/members	30,236	929
org:political/religious_affiliation	30,633	938
org:shareholders	30,288	945
org:stateorprovince_of_headquarters	30,702	1,011
org:subsidiaries	30,279	988
org:website	30,307	946
per:cause_of_death	30,388	947
per:charges	29,975	930
per:cities_of_residence	30,833	1,003
per:city_of_birth	30,118	912
per:countries_of_residence	30,870	1,035
per:country_of_birth	29,958	907
per:country_of_death	29,833	903
per:date_of_death	30,644	940
per:employee_of	32,941	1,383
per:other_family	30,106	976
per:parents	30,290	959
per:religion	30,145	935
per:spouse	30,576	967
per:stateorprovince_of_birth	30,293	908
per:title	35,913	1,671
no_relation	732,075	14,834

Table 12: Distribution of relation labels in the one-shot TACRED training partition (original vs. sampled). Each row corresponds to a relation type, where the *Original* column reports the number of instances in the full dataset and the *Sampled* column reports the number of instances included in our one-shot sampled version. The relation NO_RELATION indicates sentence pairs that do not express any annotated relation.

Algorithm 2 Sampling Procedure for Training Data

Require: Original dataset $\mathcal{D}$; sampling quotas Q; maximum positives per label K
Ensure: Sampled dataset $\mathcal{D}'$

1: Split $\mathcal{D}$ into subsets:
   - $\mathcal{D}_{r,r}$: $ss_relation = ts_relation$ ▷ Positive pairs
   - $\mathcal{D}_{r,no}$: $ts_relation =$ no_relation ▷ One relation + no_relation
   - $\mathcal{D}_{r,r'}$: $ss_relation \neq ts_relation$, $ts_relation \neq$ no_relation ▷ Different relations
2: $\mathcal{D}' \leftarrow \emptyset$
3: From $\mathcal{D}_{r,r}$: if a relation has more than K pairs, randomly down-sample to K; otherwise keep all. Add to $\mathcal{D}'$.
4: From $\mathcal{D}_{r,no}$: group by $ss_relation$. For each relation r, sample $Q[r]$ pairs (or all if fewer available). Add to $\mathcal{D}'$.
5: From $\mathcal{D}_{r,r'}$: randomly sample $2,583$ pairs, approximately preserving label distribution. Add to $\mathcal{D}'$.
6: Shuffle $\mathcal{D}'$.
7: Report statistics: $|\mathcal{D}'|$, #positives, #negatives, and ratio.
8: **return** $\mathcal{D}'$

A.6 Statistics of the Sampled Testing Set

For the testing partition, we adopt a different sampling strategy from the training data, using a random sampling approach. The resulting sampled testing set preserves the same distributional properties as the original dataset. Specifically, (1) the ratio of positive to negative instances remains consistent, (2) the distribution of relation labels is well aligned, and (3) the proportion of NO_RELATION instances is maintained. These results confirm that our random sampling strategy produces a representative testing partition that faithfully reflects the characteristics of the original testing dataset.

Category	Original	Sampled
Positive items (r, r)	7,055	704
Negative items with no_relation $(r, \textit{no_relation})$	114,725	11,480
Negative items without no_relation (r, r')	28,220	2,816

Table 13: Ratio of positives and negatives on the original testing partition of one-shot NYT29 and our sampled version. *Positive items* (r, r): pairs where both sentences express the same relation r. *Negative items with no_relation* $(r, \text{no_relation})$: pairs where one sentence expresses a relation r and the other is labeled as no_relation. *Negative items without no_relation* (r, r'): pairs where the two sentences express different relations r and r', neither being no_relation.

Category	Original	Sampled
Positive items (r, r)	772	82
Negative items with no_relation $(r, \textit{no_relation})$	146,140	14,590
Negative items without no_relation (r, r')	3,088	328

Table 14: Ratio of positives and negatives on the original testing partition of one-shot TACRED and our sampled version. *Positive items* (r, r): pairs where both sentences express the same relation r. *Negative items with no_relation* $(r, \text{no_relation})$: pairs where one sentence expresses a relation r and the other is labeled as no_relation. *Negative items without no_relation* (r, r'): pairs where the two sentences express different relations r and r', neither being no_relation.

Category	Original	Sampled
Positive items (r, r)	526	51
Negative items with no_relation $(r, \textit{no_relation})$	147,370	14,745
Negative items without no_relation (r, r')	2,104	204

Table 15: Ratio of positives and negatives on the original testing partition of one-shot WIKIDATA and our sampled version. *Positive items* (r, r): pairs where both sentences express the same relation r. *Negative items with no_relation* $(r, \text{no_relation})$: pairs where one sentence expresses a relation r and the other is labeled as no_relation. *Negative items without no_relation* (r, r'): pairs where the two sentences express different relations r and r', neither being no_relation.

Relation	Original	Sampled
/business/company/major_shareholders	30,510	3,055
/location/administrative_division/country	40,690	4,130
/location/country/administrative_divisions	41,160	4,040
/location/neighborhood/neighborhood_of	39,520	3,960
/people/deceased_person/place_of_death	33,395	3,335
no_relation	114,725	11,480

Table 16: Distribution of relation labels in the one-shot NYT29 testing partition (original vs. sampled). Each row corresponds to a relation type, where the *Original* column reports the number of instances in the full dataset and the *Sampled* column reports the number of instances included in our one-shot sampled version. The relation NO_RELATION indicates sentence pairs that do not express any relation.

Relation	Original	Sampled
no_relation	146,140	14,590
org:founded_by	15,442	1,611
org:member_of	15,246	1,557
org:top_members/employees	16,431	1,648
per:children	14,977	1,413
per:city_of_death	14,944	1,535
per:date_of_birth	15,127	1,528
per:origin	15,513	1,582
per:schools_attended	15,148	1,449
per:siblings	15,333	1,519
per:stateorprovinces_of_residence	15,699	1,568

Table 17: Distribution of relation labels in the one-shot TACRED testing partition (original vs. sampled). Each row corresponds to a relation type, where the *Original* column reports the number of instances in the full dataset and the *Sampled* column reports the number of instances included in our one-shot sampled version. The relation NO_RELATION indicates sentence pairs that do not express any annotated relation.

Relation	Original	Sampled
Fach	1743	171
IUCN conservation status	1653	165
Solid solution series with	1758	141
affiliation	1656	165
airline alliance	1617	150
airline hub	1790	174
ancestral home	1686	132
antiparticle	1629	192
architectural style	1653	198
arterial supply	1731	126
based on	1626	138

Table 18: Distribution of relation labels in the one-shot WIKIDATA testing partition (original vs. sampled). Each row corresponds to a relation type, where the *Original* column reports the number of instances in the full dataset and the *Sampled* column reports the number of instances included in our one-shot sampled version. The relation NO_RELATION indicates sentence pairs that do not express any annotated relation. **The number of relation labels in WIKIDATA is extremely large, so we split it into three parts. This is the first part.**

Relation	Original	Sampled
brother	1685	183
canonization status	1575	162
capital	1717	185
carries	1698	189
cast member	1802	170
composer	1708	156
conferred by	1688	183
connecting line	1742	216
contributor	1617	162
convicted of	1698	150
country	2481	298
country for sport	1679	162
country of origin	1833	204
creator	1804	192
culture	1584	144
depicts	1644	153
describes the fictional universe	1692	132
director	1761	149
director of photography	1695	186
drafted by	1616	183
editor	1626	144
electoral district	1665	123
employer	1614	150
enclave within	1719	120
fabrication method	1707	171
father	1735	135
field of work	1682	159
foundational text	1755	195
founder	1641	177
from fictional universe	1675	156
genre	1873	224
geography of topic	1731	183
given name	1675	150
has cause	1590	165
has quality	1573	141
head of state	1682	168
highest judicial authority	1755	162
influenced by	1583	141
instance of	2186	193
instrumentation	1743	177
is a list of	1761	180

Table 19: Distribution of relation labels in the one-shot WIKIDATA testing partition (original vs. sampled), part 2 of 3.

Relation	Original	Sampled
item operated	1776	171
killed by	1821	201
license	1700	147
list of episodes	1626	162
located next to body of water	1752	179
lyrics by	1775	213
manifestation of	1647	156
member of political party	1623	177
military branch	1713	159
minor planet group	1653	150
mouth of the watercourse	1659	204
movement	1767	171
no_relation	147370	14745
occupant	1697	162
office contested	1721	174
official language	1694	171
organizer	1751	174
original network	1609	150
place of birth	1595	189
place of publication	1731	186
points/goal scored by	1770	207
political ideology	1725	210
position played on team / speciality	1720	197
presenter	1653	144
producer	1755	189
product	1662	171
professional or sports partner	1632	171
programming language	1748	174
publication date	1817	204
published in	1656	162
radio format	1686	156
said to be the same as	1568	171
sex or gender	1731	150
student	1771	168
subclass of	2019	220
surface played on	1734	219
track gauge	1665	144
winner	1776	177

Table 20: Distribution of relation labels in the one-shot WIKIDATA testing partition (original vs. sampled), part 3 of 3.

A.7 Prompt for CogRE framework

CogRE Prompt

You are given two sentences. Follow the three steps below to determine whether they express a similar relation.

—

Summarization: Focus on the main parts between subjects and objects in the sentences.
Summarization examples:

Summarize the relations between "Malcolm Peeler" and "Pangburn" in "Dr. Malcolm Peeler , grew in Pangburn, has continued the family tradition of practicing medicine in Jonesboro .".
Summarization: Malcolm Peeler came from Pangburn.

Summarize the relations between "Oceania" and "PECC" in "Oceania and the Western Hemisphere within the PECC region , as surplus food producers and exporters , confront unique consumer issues , such as lower food expenditure and higher caloric intake compared to Asia .".
Summarization: Oceania within region PECC.

Summarize the relations between "Global Climate Research Institute" and "GCRI" in "Climate change challenges remain a key concern at the annual summit. The outlook is concerning, according to the Global Climate Research Institute (GCRI), which coordinates the event each year.".
Summarization: Global Climate Research Institute is abbreviated as GCRI.

Summarize the relations between "Panasonic Corp" and "Tesla Inc" in "Tesla Inc. is a wholly-owned subsidiary of Panasonic Corp, focusing on energy storage solutions.".
Summarization: Panasonic Corp is a subsidiary of Tesla Inc.

—

Step 1: summarize the relations between "{support_sentence_subject}" and "{support_sentence_object}" in "{support_sentence}".
Label your result as: Relation_Summarization_1.

Step 2: summarize the relations between "{test_sentence_subject}" and "{test_sentence_object}" in "{test_sentence}".
Label your result as: Relation_Summarization_2.

Step 3: are the relations between "{support_sentence_subject}" and "{support_sentence_object}" in Relation_Summarization_1 and between "{test_sentence_subject}" and "{test_sentence_object}" in Relation_Summarization_2 similar?
Focus on the phrases in the Relation_Summarization_1 an Relation_Summarization_2 that convey relations.

Generate the understanding process, followed by Yes or No in a separate line.

A.8 Prompt for Relation Phrase Extraction

Phrases Extraction Prompt for GPT-4o

Relation: {relation}
Please extract the words or phrases that indicate trigger words or relation summaries from the following answers; the relation is {relation}.
Output a string list contain all the words.
output_case_1:
{content_1}
support_sentence: {support_sentence_1}
test_sentence: {test_sentence_1}

output_case_2:
{content_2}
support_sentence: {support_sentence_2}
test_sentence: {test_sentence_2}

output_case_3:
{content_3}
support_sentence: {support_sentence_3}
test_sentence: {test_sentence_3}

output_case_4:
{content_4}
support_sentence: {support_sentence_4}
test_sentence: {test_sentence_4}

output_case_5:
{content_5}
support_sentence: {support_sentence_5}
test_sentence: {test_sentence_5}

A.9 Prompt for Baselines

A.9.1 Prompts for SUMASK

Realistic FSRE SUMASK Prompt

You are given two sentences. Follow the three steps below to determine whether they express a similar relation.

—

Summarization examples:

Summarize the relations between "Malcolm Peeler" and "Pangburn" in "Dr. Malcolm Peeler , grew in Pangburn, has continued the family tradition of practicing medicine in Jonesboro .".
Summarization: Malcolm Peeler came from Pangburn.

Summarize the relations between "Oceania" and "PECC" in "Oceania and the Western Hemisphere within the PECC region , as surplus food producers and exporters , confront unique consumer issues , such as lower food expenditure and higher caloric intake compared to Asia .".
Summarization: Oceania within region PECC.

Summarize the relations between “Global Climate Research Institute” and “GCRI” in “Climate change challenges remain a key concern at the annual summit. The outlook is concerning, according to the Global Climate Research Institute (GCRI), which coordinates the event each year.”.
Summarization: Global Climate Research Institute is abbreviated as GCRI.

Summarize the relations between “Panasonic Corp” and “Tesla Inc” in “Tesla Inc. is a wholly-owned subsidiary of Panasonic Corp, focusing on energy storage solutions.”.
Summarization: Panasonic Corp is a subsidiary of Tesla Inc.

—

Step 1: summarize the relations between “{support_sentence_subject}” and “{support_sentence_object}” in “{support_sentence}”.
Label your result as: Relation_Summarization_1.

Step 2: summarize the relations between “{test_sentence_subject}” and “{test_sentence_object}” in “{test_sentence}”.
Label your result as: Relation_Summarization_2.

Step 3: generate a question as: are the relations between “{support_sentence_subject}” and “{support_sentence_object}” in Relation_Summarization_1 and between “{test_sentence_subject}” and “{test_sentence_object}” in Relation_Summarization_2 similar?

Step 4: directly answer the question with Yes or No in a separate line.

A.9.2 Direct-Answer Prompt

Direct-Answer Prompt

Are the relations between “{support_sentence_subject}” and “{support_sentence_object}” in “{support_sentence}” and between “{test_sentence_subject}” and “{test_sentence_object}” in {test_sentence} similar?
Directly answer Yes or No in a separate line.
—
IMPORTANT: must answer with just Yes or No.

A.9.3 Simple-Reasoning Prompt

Simple-Reasoning Prompt

Are the relations between “{support_sentence_subject}” and “{support_sentence_object}” in “{support_sentence}” and between “{test_sentence_subject}” and “{test_sentence_object}” in {test_sentence} similar?
Generate the understanding process, followed by Yes or No in a separate line.

A.10 Prompt for Ablation experiments

A.10.1 COGRE- w/o chunking Prompt

CoGRE- *w/o* chunking Prompt

You are given two sentences. Follow the three steps below to determine whether they express a similar relation.
Step1: are the relations between “{paraphrased_sentence_subject}” and “{paraphrased_sentence_object}” in {paraphrased_sentence} and between “{test_sentence_subject}” and “{test_sentence_object}” in {test_sentence} similar?
Focus on the phrases in the {paraphrased_sentence} an {test_sentence} that convey relations.
Generate the understanding process, followed by Yes or No in a separate line.

A.10.2 COGRE- w/o reasoning Prompt

CoGRE- *w/o* reasoning Prompt

You are given two sentences. Follow the three steps below to determine whether they express a similar relation.
—
Summarization examples:

Summarize the relations between “Malcolm Peeler” and “Pangburn” in “Dr. Malcolm Peeler , grew in Pangburn, has continued the family tradition of practicing medicine in Jonesboro .”.
Summarization: Malcolm Peeler came from Pangburn.

Summarize the relations between “Oceania” and “PECC” in “Oceania and the Western Hemisphere within the PECC region , as surplus food producers and exporters , confront unique consumer issues , such as lower food expenditure and higher caloric intake compared to Asia .”.
Summarization: Oceania within region PECC.

Summarize the relations between “Global Climate Research Institute” and “GCRI” in “Climate change challenges remain a key concern at the annual summit. The outlook is concerning, according to the Global Climate Research Institute (GCRI), which coordinates the event each year.”.
Summarization: Global Climate Research Institute is abbreviated as GCRI.

Summarize the relations between “Panasonic Corp” and “Tesla Inc” in “Tesla Inc. is a wholly-owned subsidiary of Panasonic Corp, focusing on energy storage solutions.”.
Summarization: Panasonic Corp is a subsidiary of Tesla Inc.
—
Step 1: summarize the relations between “{support_sentence_subject}” and “{support_sentence_object}” in “{support_sentence}”.
Label your result as: Relation_Summarization_1.

Step 2: summarize the relations between "{test_sentence_subject}" and "{test_sentence_object}" in "{test_sentence}".
Label your result as: Relation_Summarization_2.

Step 3: generate a question as: are the relations between "{support_sentence_subject}" and "{support_sentence_object}" in Relation_Summarization_1 and between "{test_sentence_subject}" and "{test_sentence_object}" in Relation_Summarization_2 similar?

Step 4: Focus on the phrases in the Relation_Summarization_1 an Relation_Summarization_2 that convey relations, and directly answer the question with Yes or No in a separate line.

A.10.3 COGRE- w/o phrases Prompt

COGRE- *w/o* phrase Prompt

You are given two sentences. Follow the three steps below to determine whether they express a similar relation.
—
Summarization examples:

Summarize the relations between "Malcolm Peeler" and "Pangburn" in "Dr. Malcolm Peeler , grew in Pangburn, has continued the family tradition of practicing medicine in Jonesboro .".
Summarization: Malcolm Peeler came from Pangburn.

Summarize the relations between "Oceania" and "PECC" in "Oceania and the Western Hemisphere within the PECC region , as surplus food producers and exporters , confront unique consumer issues , such as lower food expenditure and higher caloric intake compared to Asia .".
Summarization: Oceania within region PECC.

Summarize the relations between "Global Climate Research Institute" and "GCRI" in "Climate change challenges remain a key concern at the annual summit. The outlook is concerning, according to the Global Climate Research Institute (GCRI), which coordinates the event each year.". Summarization: Global Climate Research Institute is abbreviated as GCRI.

Summarize the relations between "Panasonic Corp" and "Tesla Inc" in "Tesla Inc. is a wholly-owned subsidiary of Panasonic Corp, focusing on energy storage solutions.".
Summarization: Panasonic Corp is a subsidiary of Tesla Inc.
—
Step 1: summarize the relations between "{support_sentence_subject}" and "{support_sentence_object}" in "{support_sentence}".
Label your result as: Relation_Summarization_1.

Step 2: summarize the relations between "{test_sentence_subject}" and "{test_sentence_object}" in "{test_sentence}".
Label your result as: Relation_Summarization_2.

Step 3: generate a question as: are the relations between "{support_sentence_subject}" and "{support_sentence_object}" in Relation_Summarization_1 and between "{test_sentence_subject}" and "{test_sentence_object}" in Relation_Summarization_2 similar?
Generate the understanding process, followed by Yes or No in a separate line.

A.11 Inference Results across Model Families and Sizes on realistic One-shot RE task

In addition to Phi-4 and Qwen2.5-14B-Instruct, we further evaluate our CoGRE reasoning method on models from four additional families of varying sizes. The results, shown in Table 21, reveal two key findings: (1) CoGRE consistently outperforms prompting-based baselines across both model families and model sizes; and (2) models with fewer than 10B parameters perform poorly on the one-shot RE task.

We also include three mainstream LLMs in Table 21 to evaluate the mainstream LLMs' performance in this realistic Few-shot Relation Extraction task.

A.12 Cases Comparison of Phi-4 on TACRED

As discussed in Section 5.5, we examine the quality of explanations generated by LLMs across three stages: (1) vanilla LLM, (2) LLM trained with RL using only the accuracy reward, and (3) LLM trained with RL using both the accuracy reward and the HIT@DICT reward. We conducted evaluations on four model–dataset combinations, sampling and analyzing 40 instances for each. Here, we present examples from Phi-4 on the one-shot TACRED dataset. Specifically, we provide six illustrative cases: in five of them, the LLM trained with both accuracy and HIT@DICT rewards produces more concise summaries, and in all six cases, its explanations include relational phrases that align more closely with the gold relation labels. In the following examples, we highlight poor behaviors in yellow and the corresponding improvements in blue.

Technique	One-shot TACRED		
	P	R	F1
Qwen2.5-7B-Instruct			
Direct Matching	22.58	9.21	13.08
Random Reasoning	3.42	43.42	6.35
Cognitive-Structured RE (*our*)	10.24	50.00	17.00
Qwen2.5-3B-Instruct			
Direct Matching	100.00	0.00	0.00
Random Reasoning	10.53	7.89	9.02
Cognitive-Structured RE (*our*)	4.69	36.84	8.32
Phi-4-mini-Instruct (3.8B)			
Direct Matching	0.88	28.95	1.70
Random Reasoning	1.65	52.63	3.21
Cognitive-Structured RE (*our*)	9.35	34.21	14.69
Llama-3.1-8B-Instruct			
Direct Matching	1.61	39.47	3.09
Random Reasoning	0.43	17.11	0.85
Cognitive-Structured RE (*our*)	0.72	27.63	1.40
Mistral-7B-Instruct-v0.2			
Direct Matching	5.00	17.11	7.74
Random Reasoning	0.95	26.32	1.83
Cognitive-Structured RE (*our*)	5.14	25.00	8.52
Deepseek-v3			
Direct Matching	35.62	37.14	36.36
Random Reasoning	38.46	21.43	27.52
GPT-4o			
Random Reasoning	20.00	23.00	21.33
GPT-3.5-turbo			
Random Reasoning	16.23	36.00	22.32

Table 21: Performance comparison of different model families and sizes on the one-shot TACRED task. We evaluate our CogRE framework against prompting-based baselines (Direct Matching and Random Reasoning).

Cases1 - Error Analysis from three Stages LLMs

► **Input:**

1. Federal regulators filed a civil fraud lawsuit Thursday against **Countrywide Financial** Corp co-founder **Angelo Mozilo** and two former associates.
2. It traveled to Rice University, where de Menil and his wife, **Dominique de Menil**, who later founded the **Menil Collection**, ran the art museum.

► **Gold Relation Labels and Ground Truth:**

1. org:founded_by
2. org:founded_by
3. Yes

► **Vanilla LLMs Output:**

Relation_Summarization_1:
Federal regulators filed a civil fraud lawsuit against Angelo Mozilo, co-founder of Countrywide Financial Corp.
Relation_Summarization_2: Dominique de Menil founded the Menil Collection.
Understanding Process: - In Relation_Summarization_1, the relation is between a legal action (lawsuit) and Angelo Mozilo, who is associated with Countrywide Financial Corp as a co-founder. - In Relation_Summarization_2, the relation is between Dominique de Menil and the Menil Collection, where Dominique is the founder. The relations differ in nature: one involves a legal action against a co-founder, and the other involves founding an institution.

► **Trained LLMs with Acc. Output:**

Relation_Summarization_1:
Angelo Mozilo is a co-founder of Countrywide Financial Corp.
Relation_Summarization_2:
Dominique de Menil founded the Menil Collection.
Understanding Process:
- In Relation_Summarization_1, the relation is between a person (Angelo Mozilo) and a corporation (Countrywide Financial Corp), where the person is identified as a co-founder of the corporation.
- In Relation_Summarization_2, the relation is between a person (Dominique de Menil) and a collection (Menil Collection), where the person is identified as the founder of the collection.
- Both summaries describe a person's role in an entity, but the nature of the entities differs: one is a corporation, and the other is a collection.

► **Trained LLMs with HIT@DICT Output:**

Relation_Summarization_1:
Angelo Mozilo is a co-founder of Countrywide Financial Corp.
Relation_Summarization_2: Dominique de Menil is the founder of Menil Collection.
Understanding Process:
In Relation_Summarization_1, the relation is organizational, indicating that Angelo Mozilo is a co-founder of Countrywide Financial Corp.
In Relation_Summarization_2, the relation is organizational, indicating that Dominique de Menil is the founder of Menil Collection.
In both cases, the relation is the same: a person is the founder of a company.

Cases2 - Error Analysis from three Stages LLMs

► **Input:**

1. The Honduran interim government would not be recognized by the international community, **Jacinto Suarez**, Nicaraguan deputy to the **Central American Parliament** (PARLACEN) said Monday.

2. SEC's McKown to Step Down as Chief Counsel for Enforcement Her replacement probably will inherit pending high-profile cases, including a fraud lawsuit against Goldman Sachs and an insider-trading case against **Angelo Mozilo**, the former chief executive officer of **Countrywide Financial** Corp.

► **Gold Relation Labels and Ground Truth:**

1. org:top_members/employees
2. org:top_members/employees
3. Yes

► **Vanilla LLMs Output:**

Relation_Summarization_1: Jacinto Suarez is a Nicaraguan deputy to the Central American Parliament.
Relation_Summarization_2: Angelo Mozilo is the former chief executive officer of Countrywide Financial Corp.
Understanding Process:
Relation_Summarization_1 describes Jacinto Suarez's role as a deputy to the Central American Parliament, indicating a membership or representative relationship.
Relation_Summarization_2 describes Angelo Mozilo's role as the former CEO of Countrywide Financial Corp, indicating a leadership or executive position within the company.
The nature of the relationships differs: one is a representative role within a parliamentary body , and the other is an executive role within a corporation .
Conclusion: No.

► **Trained LLMs with Acc. Output:**

Relation_Summarization_1: Jacinto Suarez is a Nicaraguan deputy to the Central American Parliament.
Relation_Summarization_2: Angelo Mozilo is the former chief executive officer of Countrywide Financial Corp.
Understanding Process:
Relation_Summarization_1 identifies Jacinto Suarez as holding a position (deputy) within the Central American Parliament.
Relation_Summarization_2 identifies Angelo Mozilo as having held a position (chief executive officer) within Countrywide Financial Corp.
Both summaries describe a person and their role or position within an organization.
The nature of the roles is similar in that both involve being part of an organization, but the specific roles differ (deputy vs. CEO).
Conclusion: No

► **Trained LLMs with HIT@DICT Output:**

Relation_Summarization_1: Jacinto Suarez is a deputy to the Central American Parliament.
Relation_Summarization_2: Angelo Mozilo is the former chief executive officer of Countrywide Financial Corp.
Understanding Process:
In Relation_Summarization_1, the relation is organizational, indicating that Jacinto Suarez holds the position of deputy to the Central American Parliament.
In Relation_Summarization_2, the relation is occupational, indicating that Angelo Mozilo is the former CEO of Countrywide Financial Corp.
In both cases, the relation is the same: a person is employed in a specific role within a company / organization .
Yes

Cases3 - Error Analysis from three Stages LLMs

► **Input:**

1. The religious re-awakening culminated in the 1998 visit by Catholic Pope John Paul II, but **Dworin** recalls with pride that the **Cuban** leader did not ignore the Jews.
2. In Port-au-Prince, **Laura Silsby**, the leader of the 10 **American** missionaries arrested at the Dominican border for trying to smuggle out a busload of children, appeared in court Monday to answer kidnapping charges.

► **Gold Relation Labels and Ground Truth:**

1. per:origin
2. per:origin
3. Yes

► **Vanilla LLMs Output:**

Relation_Summarization_1: Dworin recalls the Cuban leader acknowledging the Jews.
Relation_Summarization_2: Laura Silsby is the leader of the American missionaries.
Understanding Process:
1. Relation_Summarization_1 focuses on Dworin's recollection of the Cuban leader's acknowledgment of Jews, indicating a historical or cultural recognition .
2. Relation_Summarization_2 describes Laura Silsby's role as the leader of a group of American missionaries, indicating a leadership or organizational relationship.
The relations are not similar: the first is about historical acknowledgment, while the second is about leadership within a group.

► **Trained LLMs with Acc. Output:**

Relation_Summarization_1: Dworin recalls the Cuban leader acknowledging Jews.
Relation_Summarization_2: Laura Silsby is the leader of American missionaries.
Understanding Process:
Relation_Summarization_1 highlights a historical event where Dworin recalls a Cuban leader's acknowledgment of Jews, indicating a recognition or interaction.
Relation_Summarization_2 describes Laura Silsby's role as the leader of a group of American missionaries, indicating a leadership or organizational relationship.
The phrases in Relation_Summarization_1 ("recalls," "acknowledging") suggest a memory or recognition of an event involving a Cuban leader.
The phrases in Relation_Summarization_2 ("leader," "American missionaries") suggest a leadership role within a group.
The nature of the relationships is different: one is about historical acknowledgment , and the other is about leadership .

► **Trained LLMs with HIT@DICT Output:**

Relation_Summarization_1: Dworin is Cuban.
Relation_Summarization_2: Laura Silsby is American.
Understanding Process:
In Relation_Summarization_1, the relation is national identity , indicating that Dworin is Cuban.
In Relation_Summarization_2, the relation is identity , indicating that Laura Silsby is American.
In both cases, the relation is the same: A person is originally from a country.

Cases4 - Error Analysis from three Stages LLMs

► **Input:**
1. This news comes from **Karr Ingham**, an economist who created the **Texas Petro Index** (TPI), which is a service of the Texas Alliance of Energy Producers.
2. **Ahmed Rashid**, a Pakistani journalist, is the author of “Descent into Chaos: The United States and the **Failure of Nation Building** in Afghanistan, Pakistan and Central Asia.”

► **Gold Relation Labels and Ground Truth:**
1. org:founded_by
2. no_relation
3. No

► **Vanilla LLMs Output:**
Relation_Summarization_1: Karr Ingham created the Texas Petro Index.
Relation_Summarization_2: Ahmed Rashid is the author of "Failure of Nation Building."
Understanding Process:
Relation_Summarization_1 identifies Karr Ingham as the creator of the Texas Petro Index.
Relation_Summarization_2 identifies Ahmed Rashid as the author of "Failure of Nation Building."
The key relation in both summaries is one of creation or authorship.
In both cases, the first subject (Karr Ingham and Ahmed Rashid) is responsible for creating or authoring the second subject (Texas Petro Index and "Failure of Nation Building").
Conclusion: Yes.

► **Trained LLMs with Acc. Output:**
Relation_Summarization_1: Karr Ingham created the Texas Petro Index.
Relation_Summarization_2: Ahmed Rashid is the author of "Failure of Nation Building."
Understanding Process:
Relation_Summarization_1 highlights a creator-creation relationship where Karr Ingham is the creator of the Texas Petro Index.
Relation_Summarization_2 highlights an author-work relationship where Ahmed Rashid is the author of "Failure of Nation Building."
Both summaries describe a relationship where one entity (Karr Ingham/Ahmed Rashid) is responsible for producing or creating another entity (Texas Petro Index/"Failure of Nation Building").
Conclusion: Yes.

► **Trained LLMs with HIT@DICT Output:**
Relation_Summarization_1: Karr Ingham is the creator of the Texas Petro Index.
Relation_Summarization_2: Ahmed Rashid is the author of "Failure of Nation Building."
Understanding Process:
In Relation_Summarization_1, the relation is organizational , indicating that Karr Ingham is the person who created the Texas Petro Index.
In Relation_Summarization_2, the relation is authorship , indicating that Ahmed Rashid is the author of a specific book.
The nature of these relations is different: one is about creation of an index, and the other is about authorship of a book.
No

Cases5 - Error Analysis from three Stages LLMs

► **Input:**

1. (1:32) R. " William Kunstler : Disturbing the Universe " – Filmmakers **Emily** and **Sarah Kunstler** explore the life of their father, the late radical civil rights lawyer who fought with Martin Luther King, Jr. and represented the famed "Chicago 8" activists who protested the Vietnam War.
2. **He** is also survived by his parents and a sister, Karen Lange, of Washington, and a brother, **Adam Lange**, of St Louis.

► **Gold Relation Labels and Ground Truth:**

1. per:siblings
2. per:siblings
3. Yes

► **Vanilla LLMs Output:**

Relation_Summarization_1:
Sarah Kunstler and Emily are filmmakers exploring the life of their father.
Relation_Summarization_2: He is survived by his brother, Adam Lange.
Understanding Process:
- In Relation_Summarization_1, the relation is between Sarah Kunstler and Emily, who are collaborators in a filmmaking project.
- In Relation_Summarization_2, the relation is between "He" and Adam Lange, who are siblings.
The phrases in Relation_Summarization_1 ("filmmakers," "exploring") indicate a professional and collaborative relationship.
In Relation_Summarization_2 ("survived by," "brother"), the relationship is familial.
Conclusion: No

► **Trained LLMs with Acc. Output:**

Relation_Summarization_1:
Sarah Kunstler and Emily are filmmakers exploring the life of their father.
Relation_Summarization_2: He and Adam Lange are siblings.
Understanding Process:
Relation_Summarization_1 identifies Sarah Kunstler and Emily as collaborators in a filmmaking project about their father.
Relation_Summarization_2 identifies a familial relationship between "He" and Adam Lange as siblings.
The phrases in Relation_Summarization_1 are "filmmakers" and "exploring," indicating a professional and collaborative relationship.
The phrases in Relation_Summarization_2 are "survived by" and "brother," indicating a familial relationship.
The nature of the relationships in both summaries is different: one is professional/collaborative, and the other is familial.
No

► **Trained LLMs with HIT@DICT Output:**

Relation_Summarization_1: Sarah Kunstler is Emily's.
Relation_Summarization_2: He is Adam Lange's brother.
Understanding Process:
In Relation_Summarization_1, the relation is familial, indicating that Sarah Kunstler and Emily are sisters.
In Relation_Summarization_2, the relation is also familial, indicating that "he" is Adam Lange's brother.
In both cases, the relation is the same: they are siblings.
Yes

Cases6 - Error Analysis from three Stages LLMs

► **Input:**
1.**China**'s National Development and Reform Commission (**NDRC**) on July 1 released a tape recording as evidence of the traders' mung bean price manipulation.
2. A US recession is likely to affect the industry, but strong growth in key Asian economies such as China and South Korea will drive regional travel demand, the **Pacific Asia Travel Association** (**PATA**) said.

► **Gold Relation Labels and Ground Truth:**
1. org:member_of
2. no_relation
3. No

► **Vanilla LLMs Output:**
Relation_Summarization_1:
NDRC is an abbreviation for China's National Development and Reform Commission.
Relation_Summarization_2: PATA is an abbreviation for the Pacific Asia Travel Association.
The relations in both summaries involve an abbreviation: NDRC is an abbreviation for China's National Development and Reform Commission, and PATA is an abbreviation for the Pacific Asia Travel Association.
Both summaries describe the relationship as one of abbreviation .
Yes.

► **Trained LLMs with Acc. Output:**
Relation_Summarization_1:
NDRC is an abbreviation for China's National Development and Reform Commission.
Relation_Summarization_2: PATA is an abbreviation for the Pacific Asia Travel Association.
Understanding Process: - In Relation_Summarization_1, the relation is that NDRC is an abbreviation for a specific organization within China. - In Relation_Summarization_2, the relation is that PATA is an abbreviation for a specific organization. - Both summaries describe an abbreviation relationship between an acronym and its full organizational name.
Conclusion: Yes.

► **Trained LLMs with HIT@DICT Output:**
Relation_Summarization_1: NDRC is China's National Development and Reform Commission.
Relation_Summarization_2: Pacific Asia Travel Association is abbreviated as PATA.
Understanding Process: In Relation_Summarization_1, the relation is organizational, indicating that NDRC is an organization within China. In Relation_Summarization_2, the relation is linguistic, indicating that PATA is an abbreviation for Pacific Asia Travel Association. The nature of these relations is different: one is about organizational identity within a country, and the other is about nomenclature.
No

A.13 Comparison of Extracted Dictionaries by Open-source and Closed-source LLMs

We provide the complete phrase lists for all relation labels, extracted separately by Qwen2.5-14B-Instruct and GPT-4o. For each relation, we report the phrases identified by both models along with representative explanation cases, enabling a detailed comparison of phrase consistency across different LLMs.

Qwen2.5-14B	**GPT4o**
/location/country/capital	*/location/country/capital*
"city"	"city"
"capital"	"capital"
"location"	"location"
"government"	"government"
"center"	
"country"	
	"city"

Table 22: The phrases for the label */location/country/capital* extracted separately by Qwen2.5-14B and GPT-4o. The yellow indicates the same phrases generated by both open-source and closed-source models.

Qwen2.5-14B	**GPT4o**
/location/location/contains	*/location/location/contains*
"city"	"city"
"location"	"location"
"province"	"provincial"
"state"	"capital"
"located"	"located"
"part"	
	"contains"

Table 23: The phrases for the label */location/location/contains* extracted separately by Qwen2.5-14B and GPT-4o. The yellow indicates the same phrases generated by both open-source and closed-source models.

Qwen2.5-14B	**GPT4o**
/location/us_state/capital	*/location/us_state/capital*
"city"	"city"
"located"	"located"
"capital"	"capital"
"state"	
	"part"

Table 24: The phrases for the label */location/us_state/capital* extracted separately by Qwen2.5-14B and GPT-4o. The yellow indicates the same phrases generated by both open-source and closed-source models.

Qwen2.5-14B	GPT4o
/business/company/ place_founded	*/business/company/place_ founded*
"based"	"based"
"located"	"located"
"headquarters"	"headquarters"
"offices"	"offices"
"building"	"building"
"logo"	"logo"

Table 25: The phrases for the label */business/company/place_founded* extracted separately by Qwen2.5-14B and GPT-4o. The yellow indicates the same phrases generated by both open-source and closed-source models.

Qwen2.5-14B	GPT4o
/film/film_location/ featured_in_films	*/film/film_location/featured_in_ films*
"setting"	"setting"
"set"	"set"
"place"	"place"
"associated"	"associated"

Table 26: The phrases for the label */film/film_location/featured_in_films* extracted separately by Qwen2.5-14B and GPT-4o. The yellow indicates the same phrases generated by both open-source and closed-source models.

Qwen2.5-14B	GPT4o
/business/company/ founders	*/business/company/founders*
"chairman"	"chairman"
"founder"	"founder"
"associated"	"associated"
"chief"	"chief"
"officer"	"officer"
"executive"	"executive"
"co-founder"	"co-founder"
"president"	"president"
"member"	

Table 27: The phrases for the label */business/company/founders* extracted separately by Qwen2.5-14B and GPT-4o. The yellow indicates the same phrases generated by both open-source and closed-source models.

Qwen2.5-14B	GPT4o
/people/person/ nationality	*/people/person/nationality*
"President"	"president"
"Minister"	"minister"
"Prime"	"prime"
"chancellor"	"chancellor"

Table 28: The phrases for the label */people/person/nationality* extracted separately by Qwen2.5-14B and GPT-4o. The yellow indicates the same phrases generated by both open-source and closed-source models.

Qwen2.5-14B	GPT4o
/business/person/ company	*/business/person/company*
"Secretary"	"secretary"
"President"	"president"
"program"	"program"
"executive"	"executive"
"chief"	"chief"
"leader"	"leader"
"chairman"	"chairman"
"officer"	"officer"
"correspondent"	"correspondent"

Table 29: The phrases for the label */business/person/company* extracted separately by Qwen2.5-14B and GPT-4o. The yellow indicates the same phrases generated by both open-source and closed-source models.

Qwen2.5-14B	GPT4o
/location/us_county/ county_seat	*/location/us_county/county_seat*
"municipality"	"municipality"
"included"	"included"
"area"	"area"
"around"	"around"
"in"	"in"
"part"	"part"

Table 30: The phrases for the label */location/us_county/county_seat* extracted separately by Qwen2.5-14B and GPT-4o. The yellow indicates the same phrases generated by both open-source and closed-source models.

Qwen2.5-14B	GPT4o
/people/person/ place_lived	*/people/person/place_lived*
"lives"	"lives"
"hometown"	"hometown"
"from"	"from"
"resides"	"resides"
"governor"	"governor"
"of"	
	"at"

Table 31: The phrases for the label */people/person/place_lived* extracted separately by Qwen2.5-14B and GPT-4o. The yellow indicates the same phrases generated by both open-source and closed-source models.

Qwen2.5-14B	GPT4o
/people/person/ children	*/people/person/children*
"mother"	"mother"
"son"	"son"
"father"	"father"
"daughter"	"daughter"
"children"	"child"
	"niece"
	"relative"

Table 32: The phrases for the label */people/person/children* extracted separately by Qwen2.5-14B and GPT-4o. The yellow indicates the same phrases generated by both open-source and closed-source models.

Qwen2.5-14B	GPT4o
/people/ethnicity/ geographic_ distribution	*/people/ethnicity/geographic_ distribution*
"energy"	"energy"
"gas"	"gas"
"loan"	"loan"
"reporter"	"reporter"
"national"	"national"
"associated"	"associated"
"pipelines"	"pipelines"
"consortium"	
"identity"	
	"foreign"

Table 33: The phrases for the label */people/ethnicity/geographic_distribution* extracted separately by Qwen2.5-14B and GPT-4o. The yellow indicates the same phrases generated by both open-source and closed-source models.

Qwen2.5-14B	GPT4o
/people/person/ ethnicity	*/people/person/ethnicity*
"American"	"american"
"Italian"	"italian"
"Manchu"	"manchu"
"Russian"	"russian"
"Tejano"	"tejano"

Table 34: The phrases for the label */people/person/ethnicity* extracted separately by Qwen2.5-14B and GPT-4o. The yellow indicates the same phrases generated by both open-source and closed-source models.

Qwen2.5-14B	GPT4o
/people/deceased_ person/place_of_ burial	*/people/deceased_person/place_ of_burial*
"home"	"home"
	"sites"
"buried"	"buried"
	"grave",
"interred"	"interred"
"cemetery"	"cemetery"

Table 35: The phrases for the label */people/deceased_person/place_of_burial* extracted separately by Qwen2.5-14B and GPT-4o. The yellow indicates the same phrases generated by both open-source and closed-source models.

Qwen2.5-14B	GPT4o
/people/place_of_ interment/interred_ here	*/people/place_of_interment/ interred_here*
"home"	"home"
"tomb"	"tomb"
"buried"	"buried"
"site"	"sites"
"staging"	"staging"
"rites"	"rites"
	"birthplace"
"residence"	"resided"

Table 36: The phrases for the label */people/place_of_interment/interred_here* extracted separately by Qwen2.5-14B and GPT-4o. The yellow indicates the same phrases generated by both open-source and closed-source models.