# *Grasping the Essentials*: Tailoring Large Language Models for Zero-Shot Relation Extraction

**Sizhe Zhou**[1], **Yu Meng**[2], **Bowen Jin**[1], **Jiawei Han**[1]
[1] University of Illinois Urbana-Champaign [2] University of Virginia
{sizhez, bowenj4, hanj}@illinois.edu, yumeng5@virginia.edu

## Abstract

Relation extraction (RE) aims to identify semantic relationships between entities within text. Despite considerable advancements, existing models predominantly require extensive annotated training data, which is both costly and labor-intensive to collect. Moreover, these models often struggle to adapt to new or unseen relations. Few-shot learning, aiming to lessen annotation demands, typically provides incomplete and biased supervision for target relations, leading to degraded and unstable performance. To accurately and explicitly describe relation semantics while minimizing annotation demands, we explore the *definition only zero-shot RE* setting where *only* relation definitions expressed in natural language are used to train a RE model. We introduce REPAL, comprising three stages: (1) We leverage large language models (LLMs) to generate initial seed instances from relation definitions and an unlabeled corpus. (2) We fine-tune a bidirectional Small Language Model (SLM) with initial seeds to learn relations for the target domain. (3) We expand pattern coverage and mitigate bias from initial seeds by integrating feedback from the SLM's predictions on the unlabeled corpus and the synthesis history. To accomplish this, we leverage the multi-turn conversation ability of LLMs to generate new instances in follow-up dialogues, informed by both the feedback and synthesis history. Studies reveal that definition-oriented seed synthesis enhances pattern coverage whereas indiscriminately increasing seed quantity leads to performance saturation. Experiments on two datasets show REPAL significantly improved cost-effective zero-shot performance by large margins.

## 1 Introduction

Relation extraction is a pivotal task in Information Extraction (IE) that involves identifying and classifying semantic relationships between entities within texts. It has wide applications in various downstream tasks including knowledge graph construction (Lin et al., 2015; Yu et al., 2020), question answering (QA) (Wang et al., 2012, 2016) and event mining (Jiao et al., 2022; Li et al., 2023c). Traditional RE works train models with human-labeled data (Han et al., 2018a, 2020; Yamada et al., 2020). However, acquiring large-scale, high-quality datasets is challenging and costly in reality. To address this data scarcity, few- and zero-shot RE works try to leverage knowledge from LM pre-training or auxiliary sources (Petroni et al., 2019; Chia et al., 2022; Han et al., 2022; Chen et al., 2022a; Zhao et al., 2023; Zhou et al., 2023a; Wan et al., 2023; Li et al., 2023b; Sun et al., 2024).

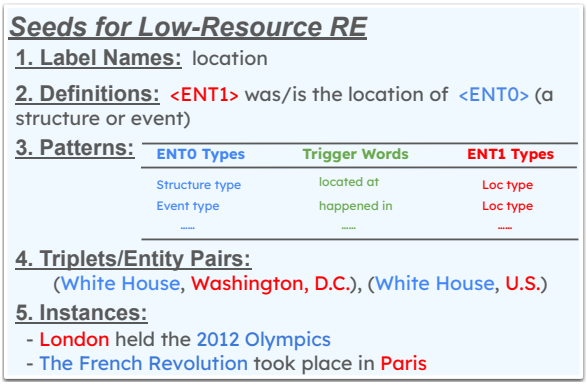


Figure 1: Different types of initial seeds for low-resource RE approaches for example relation P276. It shows using only two instances as seeds fail to cover *structure* type head entities.

Despite these advancements, two issues persist in low-resource RE. The first issue is *the underutilization of relation definitions*. Relation semantics are generally directional and multifaceted which involve entity-entity interactions and entity-related requirements (see Fig. 1). Thus, target relation semantics typically can only be partially reflected by most low-resource supervision, such as seed instances, triples, or label names. Such relation semantic complexity requires detailed elaborations described by relation definitions. Another issue



*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13462–13486
November 12-16, 2024 

is the *underutilization of LLMs for zero-shot RE*. Most LLMs are designed to perform multi-turn conversations and excel in seeking feedback from the dialogue history. Such a feature has shown great potential in knowledge-intensive or complex reasoning question-answering tasks (Trivedi et al., 2022; Zhou et al., 2023b,c). Nevertheless, LLM-based low-resource RE works typically rely on single-turn usages.

To address the first issue, this work introduces a new zero-shot RE setting where only relation definitions, instead of seen instances, are provided. In addition to the fact that relation definitions serve as more precise and less biased initial seeds, such a task setup is realistic as: (1) downstream applications such as QA tasks already have explicit definitions of interested relations and obtaining such supervision is generally straightforward; and (2) such a setting highlights the importance for RE systems to continuously adapt to new relation types based on corresponding definitions without maintaining a large amount of seen instances and re-training models.

To address the second issue, we propose a novel zero-shot RE framework, REPAL. REPAL initiates by prompting LLMs to generate positive instances based on predefined definitions and samples negative instances from an unlabeled corpus, forming an initial training set. This set is then used to train an SLM for inference efficiency and performance. Secondly, REPAL acquires and incorporates feedback to address coverage and bias issues from instance generation and SLM training. For robustness, the feedback consists of two independent components: the synthesis dialogues and sampled SLM's inference results on a large unlabeled corpus. The feedback is utilized to: (1) leverage LLMs' multi-turn conversational ability to recognize the pattern coverage bias, synthesis error, and then generate instances with new or rectified positive patterns, and (2) leverage LLMs' reasoning ability to diagnose the SLM's bias and further generate targeted or near-miss negative instances to rectify such bias by explicitly deriving negative definitions. The whole framework performs iterative refinements in which more and better-quality relation instances are accumulated to improve the task-specialized RE model.

Our data and codes are available here[1] and our contributions are as follows:

- We demonstrate the partial coverage issue of few-shot RE's initial seeds. Our studied definition-oriented RE setting can seamlessly leverage few-shot supervision for better pattern coverage and better performance by definition derivation and instance augmentation.
- We propose a novel zero-shot RE framework, REPAL, that only requires relation definitions and an unlabeled corpus. REPAL iteratively synthesizes both positive and negative instances to enhance pattern coverage and addresses biases by automatically mining and reflecting on feedback from multiple sources, leveraging the multi-turn conversation capability of LLMs.
- Extensive quantitative and qualitative experiments demonstrate the effectiveness and the potential of our task setup and framework.

## 2 Background

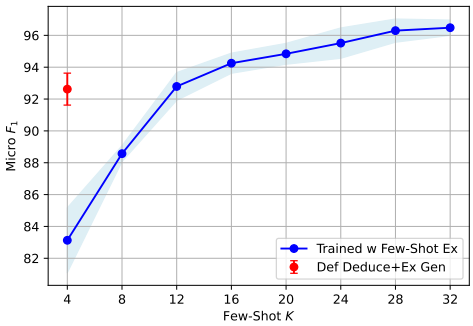


Figure 2: Micro $F_1$ (%) score of model trained with few-shot instances (*Trained w Few-Shot Ex*) and model trained with instances from our relation definition derivation and instance generation approach (*Def Deduce+Ex Gen*). The error bar/band denotes averaged value $\pm$ standard deviation.

### 2.1 Definition Only Zero-Shot Relation Extraction Task

**Task Definition 2.1. Definition Only Zero-Shot Relation Extraction.** We assume that, for any target relation $r(E_0, E_1) \in \mathcal{R}(E_0, E_1)$[2], only one associated relation definition $d(E_0, E_1)$ is given. Here $\mathcal{R}(E_0, E_1)$ denotes the whole binary relation space and $d(E_0, E_1)$ can be a single sentence or a document specifying the target relation $r(E_0, E_1)$. $E_0$ and $E_1$ are two entity placeholders.

The goal of *Definition Only Zero-Shot Relation Extraction* task is to extract all relation instances that belong to target relation $r(E_0, E_1)$ from any

[1] https://github.com/KevinSRR/REPaL

[2] This work addresses sentence-level binary relation extraction, where each instance involves evaluating the relationship between two specific entity mentions.

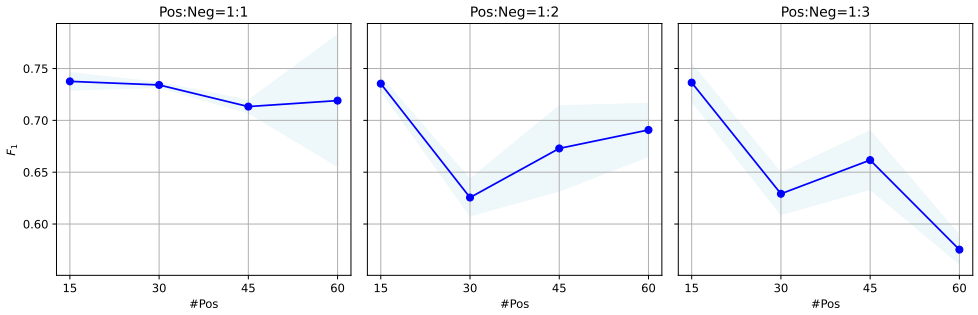


Figure 3: $F_1$ (%) scores for different setups on the number and ratio of training instances.

given set of relation instances $\{(s^i, e_0^i, e_1^i)\}_{i=1}^n$ in a binary classification manner. Here $s^i$ denotes the $i$-th context text, while $e_0^i$ and $e_1^i$ denote two entity mentions inside $s^i$. Compared with the traditional zero-shot RE task settings, our task assumes no seen data but only target relation definitions. Our task also assumes the unknown negative relation space while mainstream zero-shot RE assumes known information of all test relations. This makes our task more challenging, but aligns better with real-world scenarios. Further details and discussions are in Appendix A.

### 2.2 From Few-Shot to Definition-Oriented Low-Resource RE

In this section, we reveal the partial relation pattern coverage issue of commonly adopted RE few-shot setup and further show that the few-shot setup can be converted to our definition-oriented setup while achieving much better results. We take KNOWPROMPT (Chen et al., 2022b) as the underlying N-way classification model where N equals the number of test relations. It's a prompt tuning model with robust and strong few-shot performance and does not require hand-crafted prompts.

We experiment on two groups: (1) *Trained w Few-Shot Ex* (baseline group) which has KNOWPROMPT trained on few-shot instances for evaluation, and (2) *Def Deduce+Ex Gen* which uses GPT-4 to derive each relation's definition given 4-shot instances, generate 15 new instances based on the definition, and train KNOWPROMPT for evaluation.

The LLM derived relation definitions compared with original relation definitions are shown in Appendix G.1 and the experiment results are shown in Fig. 2 and Fig. 7 (in Appendix G.2). The derived relation definitions show that LLM is capable of deducing suitable yet generalizable relation definitions based on few-shot instances. However, the coverage of derived definitions is limited by the coverage of few-shot instances. This is in accordance with our motivation for definition-based low-resource RE setup.

Fig. 2 and Fig. 7 show that our definition derivation and instance generation approach achieves much better performance than the model trained only on few-shot instances. This indicates the approach extends the relation patterns conveyed by the few-shot instances. However, we can see our (15 generated + 4 gold shots) trained model has slightly lower performance than 16 gold shots trained model which is due to the partial coverage of relation semantics conveyed by the 4 gold shots instances. This further illustrates the importance of capturing actual relation definitions instead of few-shot data for low-resource RE approaches.

### 2.3 Effect of More Initial Seeds

As recent LLMs have enabled larger context windows, a naïve method for improving framework performance is to directly query LLMs for more generated instances. So we conduct such trials on our definition only zero-shot RE task (Sec. 3.1) to investigate whether more initially generated seeds can bring more benefits. The experiments are based on one split of the DefOn-FewRel and leverage GPT-4 as synthesis backbone with 3 random seeds. The positive instances are generated with prompt templates shown in Table 6, Table 7, and Table 8 while the negative instances are randomly sampled from the unlabeled corpus. The instances are used to train a RE model adapted from `roberta-large-mnli` (Liu et al., 2019). The quantitative results are shown in Fig. 5, Fig. 6, and Fig. 3. We can see that synthetic data by LLM is generally beneficial, but generating more initial seeds does not guarantee better results. Larger p&n[3] or larger n:p ratio both lead to higher precision and lower recall.

One explanation for these trends is that more

[3]We abbreviate the number of positive instances for each target relation as p and the number of negative instances as n.

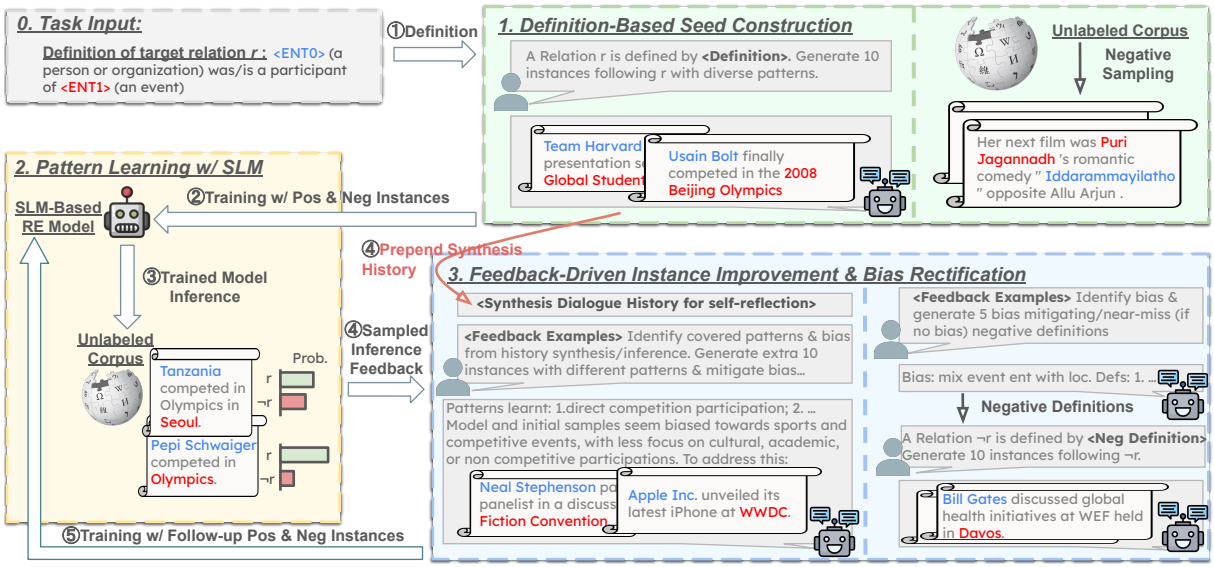


Figure 4: REPAL framework. The trained SLM-Based RE Model is used in inference stage.

positive seeds does not bring larger positive pattern coverage but results in over-fitting. However, more randomly sampled negative seeds lead to larger negative pattern coverage. Therefore the model tends to give positive predictions on instances with some dominant positive patterns but identify other minority patterns as negative. Note that when n&p are small, usually recall is pretty high while precision is low. This may also be caused by our task assumption on unknown negative relation space. The above observations and analysis motivates our design of a feedback-driven follow-up generation and refinement paradigm for instance synthesis.

## 3 Methodology

Our proposed REPAL is an iterative framework which consists of three major components: (1) Definition-based seed construction with LLM and the unlabeled corpus (Sec. 3.1). (2) Pattern learning with RE-specialized SLM (Sec. 3.2) which trains a SLM extractor with initial seeds for inference efficiency and performance. (3) Feedback-driven instance improvement and bias rectification (Sec. 3.3) by providing LLMs with independent feedback and leveraging LLMs' multi-turn conversations for followup positive and negative instance synthesis. The framework overview is shown in Fig. 4.

### 3.1 Definition-Based Seed Construction with LLM and Unlabeled Corpus

LLMs, such as the GPT family, are pre-trained for the domain adaptation ability (Radford et al., 2019). They have demonstrated to contain factual relation knowledge (Petroni et al., 2019) and are widely used as data generator for various tasks such as text classification (Meng et al., 2022) and dialogue systems (Abdullin et al., 2024). Follow-up evaluation studies have shown that LLMs are relatively skilled at constrained content generation, storytelling, and rationale generation (Sun et al., 2023; Keskar et al., 2019). Following such observations as well as the goal to tackle the data scarcity issue of zero-shot RE, REPAL first leverages LLMs to generate $K_{p_0}$ positive seeds, $\{(s_p^i, e_{0,p}^i, e_{1,p}^i)\}_{i=1}^{K_{p_0}}$, conditioned on the target relation definition $r(E_0, E_1)$. We carefully design three prompts for this step (shown in Appendix B) to encourage the LLM to generate comprehensive patterns from three perspectives in terms of complexity: `brief`, `medium`, and `implicit`. Based on exploratory experiments, they yield different types of relation patterns in accordance with our design expectations. Details of experiments and quantitatively how choice of different prompts affect the results are in Sec. 5.3.

In addition to positive seed generation, constructing negative seeds is also vital for better extraction for the target relation $r(E_0, E_1)$, as our task setting assumes no prior knowledge on the negative relation space in order to mimic the real-life deployments (Li et al., 2022). Our negative seed construction is based on the hypothesis that, with a normal large-scale unlabeled corpus, the proportion of target relation instances is relatively small. Thus REPAL conducts random sampling over the unlabeled corpus, which is expected to yield an acceptable set of negative instances denoted as $\{(s_n^i, e_{0,n}^i, e_{1,n}^i)\}_{i=1}^{K_{n_0}}$. To address potential issues in extreme cases, REPAL designs countermeasures

in Sec. 3.3.

### 3.2 Pattern Learning with RE-Specialized SLM

Given the relatively high computational costs of fine-tuning and inference with LLMs and the limitation of vanilla in-context learning (ICL) due to LLMs' difficulties in fully processing long-context demonstrations (Ma et al., 2023), we leverage SLMs as tunable and task-specialized extractors. SLMs also enable faster inference on massive unlabeled corpus to mine feedback in Sec. 3.3. Specifically, we adapt a natural language inference (NLI) model (Obamuyide and Vlachos, 2018; Liu et al., 2019; Sainz et al., 2021) to be relation-specific binary classifiers for simplicity and leave other architectures or LLMs for future research.

For each training instance $(s^j, e_0^j, e_1^j)$, the input is formulated as in NLI style:

$$\texttt{Premise}_j := s^j,$$
$$\texttt{Hypothesis}_j := d(E_0 = e_0^j, E_1 = e_1^j).$$

Given a SLM model $\mathcal{M}$, we obtain the encoded sequence hidden states $\mathbf{H}$ by:

$$\mathbf{H} = \mathcal{M}(\texttt{Premise}_j \texttt{ [SEP][SEP] Hypothesis}_j)$$

and the NLI logits $\mathbf{z} = [z_E, z_N, z_C] \in \mathbb{R}^3$ is computed as:

$$\mathbf{z} = \mathbf{W} \cdot \mathbf{H}_{[CLS]} + \mathbf{b}.$$

Finally, $P_j$, the probability of instance $(s^j, e_0^j, e_1^j)$ following relation $r(E_0, E_1)$, is computed as the normalized logit of ENTAILMENT label:

$$P_j = \frac{e^{z_E}}{\sum_{c \in \{C,N,E\}} e^{z_c}},$$

where $C, N, E$ denote logits $\mathbf{z}$'s indices for NLI label CONTRADICTION, NEUTRAL, ENTAILMENT respectively. The binary classification loss (Shannon, 1948) for relation $r(E_0, E_1)$ is defined as:

$$\mathcal{L} = -\frac{1}{|B|} \sum_{(s^j, e_0^j, e_1^j) \in B} [y_j \log(P_j) + (1 - y_j) \log(1 - P_j)], \quad (1)$$

where $B$ denotes the batched data sampled from $\{(s_p^i, e_{0,p}^i, e_{1,p}^i)\}_{i=1}^{K_{p_0}} \bigcup \{(s_n^i, e_{0,n}^i, e_{1,n}^i)\}_{i=1}^{K_{n_0}}$.

### 3.3 Feedback-Driven Instance Improvement and Bias Rectification

After obtaining the trained SLM relation extractor, REPAL further introduces a follow-up feedback-driven iterative refinement approach with intuitions listed as follows: (1) The initially generated positive seeds might only have a partial relation pattern coverage or concentrate on common patterns due to LLM's longtail deficiency (Mallen et al., 2022). Instead of one-time generation, iterative follow-up generations by incorporating feedback from examining previously generated results are better for pattern coverage. (2) Bias or errors can be introduced by previous two steps (Sec. 3.1 and 3.2). One potential source of bias is the generated positive seeds or the randomly sampled negative seeds. Another potential source of bias is the randomness of SLM training over limited training samples (see Fig. 8 and Fig. 9). Identifying and rectifying bias requires a feedback-driven follow-up refinement approach.

Specifically, we first leverage the trained SLM extractor to obtain the inference results on the unlabeled corpus. The first component of the feedback is constructed by randomly sampling instances within a predicted probability range (see Appendix F for details). It is constructed for follow-up positive instance and follow-up negative instance generation respectively, differing only on the probability ranges.

The other component of the feedback is LLMs' synthesis history. For follow-up positive instance generation, we leverage the multi-turn conversational feature of LLMs where the positive instance synthesis history is prepended to a new turn of conversation asking for positive instances with different patterns. The SLM's inference feedback is integrated into the new turn of conversation so that LLMs can identify the correct and incorrect patterns learned by SLMs. The double channel feedback design, incorporating both SLM inference and synthesis history, enhances the robustness of this step as the followup synthesis can still rely on LLMs' self-reflection if the inference feedback is uninformative.

For follow-up negative instance generation, we divide it into two sub-steps following the divide-and-conquer philosophy: the first sub-step is feedback-driven negative relation definition generation and the subsequent sub-step is negative instance generation based on the negative relation

definitions. LLM is queried to examine the SLM's inference feedback and identify the incorrect patterns learned by SLM so that newly generated negative relations can address such patterns. Similarly, for robustness in cases where SLM's inference feedback is uninformative, we query LLM to generate near-miss negative relation definitions based on the positive relation definition to better distinguish hard cases. After acquiring the negative relation definitions, the second sub-step is the same as the initial positive seed generation process in Sec. 3.1. The prompts used in this section are placed in Appendix C. After obtaining all the follow-up positive and negative instances, we iteratively train the SLM extractor.

# 4 Experiments

## 4.1 Evaluation Dataset Construction

We first construct datasets for model evaluation on the *Definition Only Zero-Shot Relation Extraction* task. Our constructed datasets are based on FewRel (Han et al., 2018b) and Wiki-ZSL (Chen and Li, 2021) respectively[4]. The public portion of FewRel contains 80 relations, each comes with 700 instances annotated by crowd workers. Wiki-ZSL consists of 113 relations with total 93,383 instances obtained by distant supervision. As their relations are from WikiData's[5] properties, we query the property definitions and slightly modify them to become complete sentences with entity placeholders `[ENT0]` and `[ENT1]` (corresponding to $E_0$ and $E_1$ as in definition notation $d(E_0, E_1)$). For evaluation, we sample 5 groups of 14 FewRel relations and 3 groups of 15 Wiki-ZSL relations as test sets.

To construct unlabeled corpus, for each group of the test set, we down-sample 10,000 instances from the unlabeled corpora used by Gao et al. for few-shot relation learning which contains 744 distantly supervised relations (labels are not used in this work) and in total 899,996 instances (2020). The final test sets with the corpus are denoted as DefOn-FewRel and Defon-Wiki-ZSL respectively.

[4]An ideal test set for our setting should follow: (1) annotations should follow the officially documented relation definitions and the sentences should be sufficient to deduce the target relations without external knowledge; (2) relations are better semantically disjoint without ambiguity so that we do not need to manually adjust binary test labels for overlapped relations in each test iteration. These principles also apply to existing multi-way classification RE datasets. However, we leave these for future dataset related work.

[5]WikiData main page

## 4.2 Baseline Methods

We compare our method with the following baselines under *Fully-Supervised* and *Zero-Shot* settings: (1) RANDOM GUESS: gives uniformly random binary predictions. (2) GPT-3.5 (Brown et al., 2020; Ouyang et al., 2022): uses `gpt-3.5-turbo-0125` model and formulates RE as a double choice problem, answering whether two entity mentions in each test instance follow the given target relation definition. (3) RE AS QA: similar to QA4RE (Zhang et al., 2023), we design more QA-like prompt templates based on the relation definitions and `gpt-3.5-turbo-0125` model to formulate RE as a multiple choice problem (double choice for *Definition Only Zero-Shot Relation Extraction* task and multi-choice for traditional zero-shot RE). (4) ROBERTA NLI (Devlin et al., 2018; Liu et al., 2019): our adopted SLM RE backbone model (Sec. 3.2). It adopts 100 gold positive and 100 gold negative instances for each relation under the *Fully-Supervised* setting. (5) ZS-BERT (Chen and Li, 2021): a Siamese-based zero-shot RE framework trained on seen labeled instances and gives prediction by nearest neighbour search comparing encoded sentence representations and relation definition representations. (6) RELATIONPROMPT (Chia et al., 2022): a Seq2Seq-based zero-shot RE framework that trains GPT-2 (Radford et al., 2019) to generate relation instances conditioned on relation names and train BART (Lewis et al., 2020) to generate the extracted relation triple on seen data. For unseen data, it fine-tune BART on synthetic data generated by trained GPT and then predicts. (7) RE-MATCHING (Zhao et al., 2023): a Siamese-based zero-shot RE model that encodes entity types and relation definitions for each relation on seen data and conducts nearest neighbour search for prediction on unseen data.

Note that ZS-BERT, RELATIONPROMPT and RE-MATCHING require being trained on labeled seen data and their predictions on unseen data need to be conducted in a multi-way classification manner where information of all relations is leveraged. These three baselines are trained on the relation instances not in the test set while leaving 5 relations' instances for dev set.

## 4.3 Experiment Setup

**Evaluation Methods.** Following our *Definition Only Zero-Shot Relation Extraction* setting, evaluation is conducted similar to cross validation where

| Model | DefOn-FewRel | | | | DefOn-Wiki-ZSL | | | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | $F_1$ | Macro-$F_1$ | Precision | Recall | $F_1$ | Macro-$F_1$ |
| *Fully-Supervised* | | | | | | | | |
| RoBERTa NLI | 79.36 | 98.46 | 86.99 | - | 68.66 | 97.14 | 78.46 | - |
| *Zero-Shot* | | | | | | | | |
| Random Guess | 7.14 | 50.77 | 12.52 | - | 6.67 | 51.01 | 11.67 | - |
| GPT-3.5 | 55.09 | 61.81 | 53.94 | - | 42.64 | 45.70 | 39.60 | - |
| RE as QA | 84.57 | 73.42 | 74.34 | 84.67 | 75.72 | 61.48 | **60.89** | 70.27 |
| RoBERTa NLI | 48.79 | 81.17 | 55.07 | - | 36.23 | 62.65 | 35.28 | - |
| ZS-BERT | 35.91 | 35.70 | - | 35.78 | 37.11 | 33.85 | - | 35.26 |
| RelationPrompt | 74.39 | 66.89 | - | 67.78 | 71.89 | 60.50 | - | 61.08 |
| RE-Matching | 77.49 | 72.95 | - | 75.11 | 73.10 | 68.99 | - | 70.97 |
| **REPaL (w GPT-4o mini)** | 73.27 | 73.75 | 70.46 | 90.01 | 65.67 | 43.53 | 44.78 | 78.47 |
| **REPaL (w GPT-4o)** | 78.86 | 77.28 | **74.61** | **91.71** | 68.98 | 47.63 | 47.80 | **80.96** |

Table 1: Evaluation results of relation extraction models under zero-shot and fully-supervised settings. REPaL is trained on 15p15n initial instances to acquire SLM inference feedback and then trained with additional 15p15n follow-up instances (30p30n in total). We show the results with different backbone synthesis LLMs (GPT-4o and GPT-4o mini). Note that the results for GPT-3.5 and RE as QA baselines on the *Definition Only Zero-Shot Relation Extraction* setting are based on the down-sampled subsets with 30 samples/relation to reduce inference cost.

each iteration one test relation is treated as the gold positive target relation and remaining test relations serve as gold negative test relations (our setting assumes all negative relations are unknown in terms of relation definitions and any distribution information). Each test iteration is a binary classification problem with precision, recall and $F_1$ calculated. Table 1 shows the main results where Precision, Recall and $F_1$ indicates the corresponding binary classification metric scores averaged across all test iterations. If the cell for $F_1$ column is empty, then the Precision and Recall represents the Macro Precision and the Macro Recall corresponding to Macro $F_1$. Since ZS-BERT, RelationPrompt and RE-Matching require the traditional zero-shot RE setting where multi-way classification is assumed, we further train the KnowPrompt (Chen et al., 2022b) multi-way RE classifier with all the positive instances of test relations synthesized by REPaL in the *Definition Only Zero-Shot Relation Extraction* setting for comparison. This leads to the Macro-$F_1$ column. See Appendix F for further implementation details and see Appendix A for details of our definition only zero-shot RE setting.

## 5 Results and Analysis

### 5.1 Main Results

The main results are shown in Table 1. REPaL achieves generally better performance compared with zero-shot baselines with large margins for both evaluation datasets. This shows the effectiveness of our method in low-resource settings and the robustness across different benchmarks.

The advance of REPaL is slightly lower in DefOn-Wiki-ZSL compared to DefOn-FewRel which is because Wiki-ZSL is much more noisy due to distant supervised annotations. By comparing the absolute values of $F_1$ scores derived for our zero-shot RE setting with the Macro-$F_1$ scores derived for traditional zero-shot RE setting, it can be concluded that our definition only zero-shot setting is much more challenging. Once we break the assumption of unknown negative relation space, models can take shortcuts to distinguish different relations without actually comprehending the relations.

Compared with fully-supervised baselines, it shows there is still room for improvement. This is related to our task's assumption on unknown negative test relation space which is normally overlooked in mainstream zero-shot RE works. Note that, on our *Definition Only Zero-Shot RE setting*, RE as QA achieves better performance on Defon-Wiki-ZSL but lower on traditional zero-shot RE setting. This shows that LLMs are capable of judging whether an instance follows a certain relation if given a clear relation definition. However, they are less competent for dealing with multiple relations. Meanwhile, inference with LLMs are much more costly as the RE as QA (w/ GPT-3.5) would cost around $260 for one DefOn-FewRel split on our setting if without down-sampling. For traditional setting, the RE as QA would cost around $2.3. REPaL, in contrast, even with GPT-4-0125-preview, costs around $3.7 for generating 30p30n

train and 30p30n dev examples[6]. The generating instances are further reused for both settings.

### 5.2 Ablation Study

| Model | DefOn-FewRel | | |
|---|---|---|---|
| | Precision | Recall | $F_1$ |
| REPAL | **78.86** | 77.28 | **74.61** |
| - N_feedback | 76.74 | 77.47 | 73.04 |
| - P_feedback | 73.59 | 75.90 | 70.03 |
| - P_init | 48.79 | **81.17** | 55.07 |

Table 2: Ablation results of REPAL. P_init, P_feedback, and N_feedback denote initial positive generation, feedback-driven follow-up positive generation and negative generation respectively. The ablations are based on REPAL w GPT-4o.

To investigate the effectiveness of our framework design, we conduct ablation studies with results shown in Table 2. The initial seed generation brings considerable performance advance which shows LLM's power of domain adaptation is fully leveraged given a clear relation definition. Compared to the results shown in Sec. 5.3 where we generate 30p30n instances all at once, REPAL achieves better recall and $F_1$ both with and without follow-up negative instance generation. This indicates the importance of feedback-driven generation design.

Furthermore, the negative follow-up instance generation further boosts the precision, demonstrating its effectiveness on rectifying SLM's bias in distinguishing positive and negative relations.

### 5.3 Effect of Positive Templates

| Model | DefOn-FewRel | | |
|---|---|---|---|
| | Precision | Recall | $F_1$ |
| REPAL | 78.63 | **74.29** | **71.71** |
| - `implicit` | **80.60** | 71.15 | 71.40 |
| - `brief` | 80.00 | 71.08 | 71.56 |

Table 3: Evaluation results w/o follow-up generation and conditioned on different initial positive generation templates. 30p30n training instances are gathered across all settings.

Table 6, Table 7, and Table 8 in Appendix B shows the adopted initial positive seed construction templates. Note that no ICL is adopted in REPAL's generation step. Analysis in Appendix B yields the conclusion that the generated instance's patterns generally follow the prompt instructions well, covering instances with brief, medium-length, and implicit patterns respectively.

The quantitative results with different combinations of positive instance generation prompts are shown in Table 3. We can see that results are generally robust against different prompt combinations. Leveraging all the `medium`, `brief`, and `implicit` prompts yields the best recall and slightly better overall $F_1$. In our main experiments, all three prompts are adopted to enhance diversity of generated relation patterns.

### 5.4 Effect of More Iterations

| Iteration | DefOn-FewRel | | |
|---|---|---|---|
| | Precision | Recall | $F_1$ |
| 1 | 73.59 | 75.90 | 70.03 |
| 2 | 78.86 | 77.28 | 74.61 |
| 3 | 80.61 | 74.91 | 74.57 |
| 4 | 78.85 | 76.93 | 74.96 |

Table 4: Results on DefOn-FewRel with different iterations based on REPAL w GPT-4o. Iteration 1 refers to the round of initial instance generation, and Iteration 2 refers to the first feedback-driven instance generation which is taken for the main experiment.

We further run REPAL with more iterations on the DefOn-FewRel dataset with results shown on Table 4. The results indicates that more iterations can further improve the performance, but it exhibits a dynamic trade-off between precision and recall, a common challenge of learning with limited supervision, especially given our assumption of unknown negative relation space: As the model learns to recognize more true positive instances through the extended pattern coverage brought by synthesized samples, it may also include more noise (false positives). When the model corrects the false positives in the next iteration, it may become more conservative and lower the recall. Since conducting more rounds of iterative refinement incurs more costs, we leave more comprehensive explorations (e.g., performance change by more iterations, impact of LLM's long-context capabilities) for future work.

### 5.5 Error Analysis

Table 5 shows the major source of false positive predictions of the final tuned SLM RE model, it can be seen that the majority of false positive predictions are concentrated on a few similar negative relations. As our proposed task setting assumes the unknown negative relation space, one challenge appears to be learning the positive relation against the unknown

[6]Approximated by multiplying the number of input and output tokens of GPT-4o with the price of GPT-4-0125-preview.

| Example Target Relation | Majority False Positive Predicted Relations | Example Instance of False Positive Predicted Relations |
|---|---|---|
| P40: <ENT1> was/is the child (not stepchild) of <ENT0> | 1. P26 (129): <ENT1> was/is the married spouse (husband, wife, partner, etc.) of <ENT0><br>2. P3373 (183): <ENT1> and <ENT0> had/have at least one common parent (<ENT1> is the sibling, brother, sister, etc. including half-sibling of <ENT0>) | 1. Daughter of <ENT1> Sancho IV </ENT1> and of <ENT0> María de Molina </ENT0> , Infanta Beatrice was born in Toro . (Gold: P26 \| Pred: P40 \| Pos Prob: 0.861)<br>2. Sofia Coppola was born in New York City , New York , the youngest child and only daughter of set decorator / artist <ENT1> Eleanor Coppola </ENT1> ( née Neil ) and director <ENT0> Francis Ford Coppola </ENT0> . (Gold: P26 \| Pred: P40 \| Pos Prob: 0.665)<br>3. <ENT1> Ruby Aldridge </ENT1> is the daughter of former Playboy playmate Laura Lyons and artist and graphic designer Alan Aldridge , and younger sister of fashion model <ENT0> Lily Aldridge </ENT0> . (Gold: P3373 \| Pred: P40 \| Pos Prob: 0.885)<br>4. Amongst his regular visitors were his younger brothers <ENT0> Jyotirindranath Tagore </ENT0> ( 1849–1925 ) and Rabindranath Tagore ( 1861–1941 ) , the Nobel Prize – winning poet , and his sister <ENT1> Swarnakumari Devi </ENT1> . (Gold: P3373 \| Pred: P40 \| Pos Prob: 0.945) |
| P410: <ENT1> was/is the military rank achieved by or associated with <ENT0> (a person or a position) | P241 (210): <ENT1> was/is the military branch to which <ENT0> (a military unit, award, office, or person) belonged/belongs | 1. In November 1966 , retired <ENT1> USMC </ENT1> Major <ENT0> Donald Keyhoe </ENT0> and Richard H. Hall , both of NICAP , briefed the panel . (Gold: P241 \| Pred: P410 \| Pos Prob: 0.906)<br>2. <ENT0> Ricardo Sanchez </ENT0> ( born 1953 ) is a retired <ENT1> United States Army </ENT1> lieutenant general . (Gold: P241 \| Pred: P410 \| Pos Prob: 0.768) |

Table 5: Error analysis of the predictions made by SLM-based RE model. The contents in red denote the number of false positive predictions for a specific relation. The contents in blue denote the prediction details made by SLM-based RE model. *Gold* refers to the gold relation label of an instance. *Pred* refers to the predictions made by our model. *Pos Porb* means the predicted probability of the instance following the target relation.

and infinitely many negative relations. To address such challenge, our model derive targeted negative relations based on the feedback of model inference. Based on the results in Table 5, we can see that the challenge is not fully eliminated which serves as a promising future research direction. Another feature seen from the false positive predicted instances is that some typical false positive instances actually express the target relation in addition to its gold relation. However, the target relation is not expressed by the tagged entity mention pair. This may indicate that better RE architectures which well model the position awareness of target entity pairs can be adopted for improving the overall performance.

## 6 Related Work

**Zero-Shot Relation Extraction** Our work is related to zero-shot RE (Levy et al., 2017). Majority zero-shot RE approaches mainly leverage clustering, label-verbalization, or Siamese-based architectures (Rahimi and Surdeanu, 2023; Chen and Li, 2021; Chia et al., 2022; Li et al., 2023a) which seek for the instance-instance similarity or the similarity between the relation instances and the unseen relations' information. Chen and Li (2021) utilize relation descriptions for zero-shot RE but their approach still relies on seen data to align relation descriptions with instances in a supervised manner. Li et al. (2023b) adopt the relation descriptions but only for verifying synthesized data with the instance-level seeds. LLM-based RE works focus on designing prompting strategies or LLM alignment to tackle zero-shot RE (Li et al., 2023a; Zhang et al., 2023; Wei et al., 2023; Wadhwa et al., 2023). Our work is distinguished from pre-LLM zero-shot RE works as they heavily rely on the supervision from massive seen data and the complete negative relation space. And the majority do not focus on relation definitions. Our work is different from LLM-based zero-shot RE works as we emphasize both the rich relation definitions for data synthesis and synergy between SLM and LLM.

**Definition-Driven Text Mining** BERTNet (Hao et al., 2023) applies definitions for distilling entities from LM parametric knowledge. Label definitions/descriptions have also been proven to be powerful in text classification (Gao et al., 2023). In zero-shot RE, several PLM-based works have utilized relation definitions (Chen and Li, 2021; Zhao et al., 2023) but they mainly focus on computing instance-definition similarities. In our work, LLM is used to distill patterns and extend or rectify the learning of SLM based on definitions.

## 7 Conclusion

In this work, we have introduced a new zero-shot RE task where only relation definitions instead of seen-unseen relation instances are provided. Correspondingly, we have proposed REPAL which leverages LLMs and unlabeled corpora to generate relation instances and iteratively self-improves the generation pattern coverage while rectifying the bias by automatically acquiring and reflecting over sampled feedback from multiple sources. Quantitative experiments and qualitative analysis on our two modified datasets show the effectiveness and robustness of our framework as well as our large-margin advance over most baselines. Exploratory experiments show that generating more data in a single-turn conversation does not yield proportionally larger pattern coverage. We also proposed a derive-definition-then-generate approach which achieves much better performance than just utilizing few-shot instances. This gives insights into low-resource RE works to capture the complete relation semantics to avoid partial coverage by few-shot instances.

## Limitations

In this work, we mainly experimented on GPT for data synthesis as their instruction following performance is competent so that we do not need to introduce in-context learning in most of the time. Therefore, one follow-up work is to explore other LLMs to see their generation capability compared to the GPT series. Besides, new RE datasets tailored for our definition only zero-shot RE still can be created as it still lacks large scale yet high quality datasets. Thirdly, prompt engineering and hyperparameter search are not conducted. For the sake of better performance in downstream tasks, future works could compensate this.

## Ethics Statements

Since our goal is to solve sentence-level RE tasks where the text contexts are sufficient to derive the relation, factualness of the relation triples is not a strict requirement or a vital factor for the training instances. Therefore, in generative data synthesis, we do not further verify the factualness of the generation results and we simply count on the LLMs. Therefore follow-up works could explore this and other related approaches should also be careful if they want to adapt our work to downstream tasks necessitating factualness such as factual question answering.

## Acknowledgements

Research was supported in part by the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) by NSF under Award No. 2118329, US DARPA INCAS Program No. HR0011-21-C0165 and BRIES Program No. HR0011-24-3-0325, National Science Foundation IIS-19-56151, and the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily represent the views, either expressed or implied, of DARPA or the U.S. Government.

## References

Yelaman Abdullin, Diego Molla-Aliod, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2024. Synthetic dialogue dataset generation using llm agents. *arXiv preprint arXiv:2401.17461*.

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

Chih-Yao Chen and Cheng-Te Li. 2021. Zs-bert: Towards zero-shot relation extraction with attribute representation learning. *arXiv preprint arXiv:2104.04697*.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022a. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web conference 2022*, pages 2778–2788.

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

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022. Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. *arXiv preprint arXiv:2203.09101*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lingyu Gao, Debanjan Ghosh, and Kevin Gimpel. 2023. The benefits of label-description training for zero-shot text classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13823–13844, Singapore. Association for Computational Linguistics.

Tianyu Gao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2020. Neural snowball for few-shot relation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7772–7779.

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2020. More data, more relations, more context and more openness: A review and outlook for relation extraction. *arXiv preprint arXiv:2004.03186*.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018a. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2245.

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

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

Shibo Hao, Bowen Tan, Kaiwen Tang, Bin Ni, Xiyan Shao, Hengzhe Zhang, Eric Xing, and Zhiting Hu. 2023. BertNet: Harvesting knowledge graphs with arbitrary relations from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5000–5015, Toronto, Canada. Association for Computational Linguistics.

Yizhu Jiao, Sha Li, Yiqing Xie, Ming Zhong, Heng Ji, and Jiawei Han. 2022. Open-vocabulary argument role prediction for event extraction. *arXiv preprint arXiv:2211.01577*.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. *arXiv preprint arXiv:1706.04115*.

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

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

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

Sha Li, Heng Ji, and Jiawei Han. 2022. Open relation and event type discovery with type abstraction. *arXiv preprint arXiv:2212.00178*.

Sha Li, Ruining Zhao, Manling Li, Heng Ji, Chris Callison-Burch, and Jiawei Han. 2023c. Open-domain hierarchical event schema induction by incremental prompting and verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5677–5697, Toronto, Canada. Association for Computational Linguistics.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Conference on Empirical Methods in Natural Language Processing*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the first workshop on fact extraction and VERification (FEVER)*, pages 72–78.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero- and few-shot relation extraction. *arXiv preprint arXiv:2109.03659*.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.

Qi Sun, Kun Huang, Xiaocui Yang, Rong Tong, Kun Zhang, and Soujanya Poria. 2024. Consistency guided knowledge retrieval and denoising in llms for zero-shot document-level relation triplet extraction. *arXiv preprint arXiv:2401.13598*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *ACL*, pages 15566–15589.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. Gpt-re: In-context learning for relation extraction using large language models. *arXiv preprint arXiv:2305.02105*.

Chang Wang, Aditya Kalyanpur, James Fan, Branimir K Boguraev, and DC Gondek. 2012. Relation extraction and scoring in deepqa. *IBM Journal of Research and Development*, 56(3.4):9–1.

Huazheng Wang, Fei Tian, Bin Gao, Chengjieren Zhu, Jiang Bian, and Tie-Yan Liu. 2016. Solving verbal questions in iq test by knowledge-powered word embedding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 541–550.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

Haoze Yu, Haisheng Li, Dianhui Mao, and Qiang Cai. 2020. A relationship extraction method for domain knowledge graph construction. *World Wide Web*, 23:735–753.

Kai Zhang, Bernal Jiménez Gutiérrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. *arXiv preprint arXiv:2305.11159*.

Jun Zhao, Wenyu Zhan, Xin Zhao, Qi Zhang, Tao Gui, Zhongyu Wei, Junzhe Wang, Minlong Peng, and Mingming Sun. 2023. Re-matching: A fine-grained semantic matching method for zero-shot relation extraction. *arXiv preprint arXiv:2306.04954*.

Sizhe Zhou, Suyu Ge, Jiaming Shen, and Jiawei Han. 2023a. Corpus-based relation extraction by identifying and refining relation patterns. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 20–38. Springer.

Sizhe Zhou, Siru Ouyang, Zhuosheng Zhang, and Hai Zhao. 2023b. Towards end-to-end open conversational machine reading. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2064–2076, Dubrovnik, Croatia. Association for Computational Linguistics.

Yucheng Zhou, Xiubo Geng, Tao Shen, Chongyang Tao, Guodong Long, Jian-Guang Lou, and Jianbing Shen. 2023c. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734*.

# A Detailed Discussions on Definition Only Zero-Shot Relation Extraction Setup

## A.1 Task Setup and Evaluation Process

Suppose the RE test set contains $R$ relations and each relation $r$ has $N_r$ instances. Then the evaluation will be conducted in $R$ iterations and the final score will be calculated by averaging over the individual scores from the total $R$ iterations.

For each test iteration, we take one test relation $r$ as the target relation (gold positive relation) with $N_r$ gold positive instances. All the other test relations will be treated as negative relations and all their associated instances are gold negative instances. So each iteration is in the format of binary classification, targeting at the set of gold positive instances and gold negative instances. Additionally, we assume the negative relation space is unknown, which means that for each test iteration, we only know the relation definition of the gold positive relation $r$, and we do not know any information about the negative test relations. Such cross-validation evaluation is designed for robustness as the results are averaged over all different test relations. For the evaluation complexity, if we assume each test relation has balanced $N$ instances, the complexity is proportional to $N \cdot R^2$.

## A.2 Distinctions with Traditional Zero-Shot RE

Traditional zero-shot relation extraction models are trained on a large set of seen relations with corresponding labeled instances. During evaluation, the trained model will deal with a set of unseen relations with corresponding relation instances. The trained model will have access to the information of all unseen relations (at the same time) in the format such as relation label names, or relation descriptions/definitions, or other more fine-grained information (e.g., extended possible entity type lists). Then the final model is required to conduct multi-way classification over all unseen relations.

For our newly defined Definition Only Zero-Shot Relation Extraction task:

- We do not rely on any seen relation or any labeled relation instance.
- We only assume a clear and complete relation definition for each target positive relation and an unlabeled corpus.
- We assume unknown negative relation space which means for each test iteration, in addition to the single positive target test relation, all the other test relations are treated as negative relations and we do not know any information about how many negative relations are and what the negative relations are.
- The evaluation process is completely different from traditional zero-shot RE as described above.

## A.3 Motivations and Practical Values

We design our Definition Only Zero-Shot Relation Extraction in order to accommodate more realistic and more challenging applications as the fast developing LMs, especially LLMs, are enabling the design of such advanced systems. Here we will emphasize the motivations and values of the setup of our Definition Only Zero-Shot Relation Extraction task.

First, the assumption that the definition of the target unseen relation is given instead of assuming and using annotated data of other seen relations accommodates various applications scenarios. When people deal with domain specific problems, the definitions of interested relations are normally clear and explicit. For instance, an expert in the geographic information system (GIS) domain might want to model textual patterns which describe two geospatial entities "touches" with each other (their interiors do not intersect and only their boundaries intersect). They already have well defined terminologies and associated definitions for such relations but the annotations are expensive. Another example lies in the question answering task where one can derive the "<ENT0> is youngest birth child of <ENT1>" definition if they are interested in extracting instances (text contexts and tuples) for the question "Who is the youngest child of Person A?". Additionally, when crowdsource workers are annotating RE samples, they are often provided with the relation definitions to guide their annotation work. Therefore it's realistic to assume that a clear and explicit relation definition is given. From the above cases, it can also be seen that the potential relations of interest are infinite which emerge with different problems in different domains. But annotating in-domain samples are expensive and time consuming. It's hence also meaningful to tackle RE from the source. Namely, focusing on definitions which capture the complete relation semantics compared to other types of starting seeds and leveraging the domain adaptation power and the constrained gen-

eration power of LLMs to alleviate the annotation scarcity issue.

Second, the assumption of unknown negative relation space is to mimic the real world setting where the number of interested relations is extremely small compared to the number of negative relations between entities. For example, one may be interested in one Wikidata property relation but the number of total Wikidata properties is massive. In such cases, approximately, we barely know anything about the negative relation space. But to train a good relation extractor that can distinguish interested target relations against all the other negative relations in the corpus, we will need some method designs to deal with such unknown negative relation space. Our assumption of unknown negative relation space can also be considered as a more challenging version of "None" or "No Relation" relation labels for traditional multi-classification RE datasets.

Third, the assumption of a large unlabeled corpus is natural as the unlabeled corpus usually come together with specific domains. Still taking the geospatial RE as an example, it's relatively easy to acquire documents that mention about geospatial entities and potentially their mutual relations. Some experts from GIS could also be able to provide such corpus. Since the corpus do not need to be labeled with relations, it's much more convenient and efficient to get such unlabeled corpus set up instead of gathering domain-specific annotations. Besides, our framework does not solely rely on the unlabeled corpus as we have designed follow-up positive and negative instance generation processes to rectify the bias and extend the relation pattern coverage which synthesizes follow-up positive and negative instances. LLMs will conduct self-reflection on the given relation definition, synthesis dialogue history, and the sampled inference results on the unlabeled corpora. Among those sampled inference results, LLM will judge whether the prediction is correct or wrong. If the sampled inference results contain correct predictions, LLM can analyze the instances which convey the information on the patterns learnt by SLM. If the sampled inference results contain wrong predictions, LLM can summarize their relations and generate follow-up similar negative relations and corresponding instances to correct the learning of SLM. Even if the feedback does not contain useful information, LLM can still generate more positive relation patterns based on the previously generated instances. LLM can also generate near-miss negative relations simply based on the definition of the positive target relation. With LLMs becoming more powerful in inference and having longer context window sizes, the performance gain of follow-up feedback driven generation process can be further improved.

Finally, our setting of binary NLI is versatile and has great potential to adapt to multi-way classification and multi-label classification settings. Based on the task setup and evaluation process, we can see there will be one binary relation classifier for each target relation. This is versatile because if we stick to the multi-classification setting based on data synthesis approach, we would need to retrain our multi-way classifier on synthesized data if there is an additional new relation coming in. Furthermore, the setup of $R$ binary classifiers accommodates the scenarios where there are relations entailed by other relations or the contexts indicate two possible relations which are not allowed by multi-way classification setup. If we have multiple interested relations and the number of such relations are large, there are various methods to reduce the cost of adapting binary classifiers for multi-class/label problems. One method is to use some rules (e.g., mismatched entity types) or some coarse-grained NLI methods to first filter impossible relation candidates to reduce the candidate relation space and then apply our trained relation classifiers.

## B Prompt Templates Used in Definition-Based Seed Construction with LLM

Table 6, Table 7, and Table 8 contains the three prompt templates used for generating initial positive seeds using LLMs.

Example instances generated by corresponding prompts are shown in Table 9. Our goal of designing such prompts is to cover all the patterns for target relations. From the generated example instances, it is evident that the pattern complexity (or, more simply, the sentence length) shows considerable differences, particularly between the `implicit` prompt and the other two prompts. The pattern complexity (or more naïvely, the sentence length) well follows the instructions conveyed by each type of prompt and well represents the prompt name, `brief`, `medium`, and `implicit` correspondingly.

| PROMPT TEMPLATE FOR INITIAL POSITIVE INSTANCE GENERATION (BRIEF) |
|---|
| A binary relation between entity placeholders <ENT0> and <ENT1> is defined by "{relation_definition}". Under sentence-level relation extraction setting, generate {number_of_examples} examples (numbered from 1 to {number_of_examples}) expressing the same relation, where <ENT0> is replaced with actual entity mention and is prefixed with tag <ENT0> and suffixed with tag <ENT0> , <ENT1> is replaced with actual entity mention and is prefixed with tag <ENT1> and suffixed with <ENT1> . Do not overfit the pattern of the definition. Try as many different relation patterns or relation expressions as possible. |

Table 6: Prompt templates (`brief`) used in Definition-Based Seed Construction with LLM (Sec. 3.1). Words highlighted denote the placeholders for filling in contents indicated by their surface names.

| PROMPT TEMPLATE FOR INITIAL POSITIVE INSTANCE GENERATION (MEDIUM) |
|---|
| A binary relation between entity placeholders <ENT0> and <ENT1> is defined by "{relation_definition}". Under sentence-level relation extraction setting, generate {number_of_examples} examples (numbered from 1 to {number_of_examples}) expressing the same relation, where <ENT0> is replaced with actual entity mention and is prefixed with tag <ENT0> and suffixed with tag <ENT0> , <ENT1> is replaced with actual entity mention and is prefixed with tag <ENT1> and suffixed with <ENT1> . Other content requirements:<br><br>1. Do not overfit the pattern of definition. Try as many different relation patterns or relation expressions as possible.<br>2. Generate rich and informative related contexts before and after each entity. |

Table 7: Prompt templates (`medium`) used in Definition-Based Seed Construction with LLM (Sec. 3.1). Words highlighted denote the placeholders for filling in contents indicated by their surface names.

## C Prompt Templates Used in Feedback-Driven Instance improvement and Bias Rectification

Table 10, Table 11, and Table 12 contain prompt templates for our feedback-driven follow-up positive instance generation and negative relation definition generation respectively. Note that Table 12, which assumes there are previously generated negative relation definitions, is required for REPAL iterations following the first feedback-driven generation iteration. After obtaining the negative relation definitions, we simply leverage the `medium` instance generation template in Table 7 to generate negative relation instances. We take this template as our purpose of negative instance generation is to rectify the existing bias instead of pursuing complete negative relation pattern coverage. Furthermore, Table 3 demonstrates that the performance difference between the usages of different templates is minor.

## D Details of Constructed DefOn-FewRel and DefOn-Wiki-ZSL Datasets

Table 13 shows the example relation labels and constructed definitions. Please refer to our Github repository for detailed relations and definitions. To get quality evaluation samples, we conduct test data cleaning with the requirements as: (1) The two entity mentions should not overlap; (2) The entity mentions should not be pronouns such as *I*, *he*, and *she*. Note these two requirements only give negligible impact on the number of relation instances. Furthermore, we clean the unlabeled corpora before down-sampling by requiring that selected unlabeled samples should not be repeated. Namely, for any two unlabeled samples, the sentence, the head entity mention and the tail entity mention can not all be the same.

## E Effect of More Initial Seeds

The precision (%) scores and recall (%) scores discussed in Sec. 2.3 are shown in Fig. 5 and Fig. 6 respectively.

## F Implementation Details

### F.1 Baselines

In consideration of OpenAI API calling expense, the GPT-3.5 baseline results are from the evaluation over down-sampled test sets (30 down-sampled test instances for each relation). The prompt template used for GPT-3.5 baseline is shown by Table 14. The prompt templates used for RE AS QA on our *Definition Only Zero-Shot RE* setting and on traditional multi-class classification setting are shown by Table 15 and Table 16 respectively. For inference, we use the temperature as 0. For all the other baselines, we adopted their original hyperparameters reported in corresponding publications.

**PROMPT TEMPLATE FOR INITIAL POSITIVE INSTANCE GENERATION (IMPLICIT)**

A binary relation between entity placeholders <ENT0> and <ENT1> is defined by "{relation_definition}". Under sentence-level relation extraction setting, generate {number_of_examples} examples (numbered from 1 to {number_of_examples}) expressing the same relation, where <ENT0> is replaced with actual entity mention and is prefixed with tag <ENT0> and suffixed with tag <ENT0> , <ENT1> is replaced with actual entity mention and is prefixed with tag <ENT1> and suffixed with <ENT1> . Other content requirements:

1. Do not overfit the pattern of definition. Try as many different relation patterns or relation expressions as possible.
2. Generate rich and informative related contexts before and after each entity.
3. The relation patterns or relation expressions should be implicit or complicated

Table 8: Prompt templates (`implicit`) used in Definition-Based Seed Construction with LLM (Sec. 3.1). Words highlighted denote the placeholders for filling in contents indicated by their surface names.

| Prompt Template Name | Example Generated Instances (Relation=P241) |
|---|---|
| `brief` | 1. During World War II, <ENT0>The 101st Airborne Division</ENT0> proudly served under <ENT1>The United States Army</ENT1>, where they played a crucial role in several pivotal battles.<br>2. The prestigious <ENT0>Victoria Cross</ENT0> is awarded by <ENT1>The British Army</ENT1> for acts of extraordinary valor in the presence of the enemy.<br>3. <ENT0>General Douglas MacArthur</ENT0> was a prominent figure in <ENT1>The United States Army</ENT1>, noted for his leadership in the Pacific Theater during World War II. |
| `medium` | 1. <ENT0>The U.S. Navy SEALs</ENT0>, an elite special operations force, is a notable unit that belongs to the <ENT1>United States Navy</ENT1>, which plays a crucial role in maritime security and warfare.<br>2. During his illustrious career, <ENT0>General Dwight D. Eisenhower</ENT0> served in the <ENT1>United States Army</ENT1>, playing a key role during World War II and later becoming the 34th President of the United States.<br>3. The prestigious <ENT0>Distinguished Flying Cross</ENT0> is an honor awarded to personnel of the <ENT1>United States Air Force</ENT1> for acts of heroism or extraordinary achievement during aerial flight. |
| `implicit` | 1. During the freezing winter of 1944, the decisive Battle of the Bulge tested the mettle of many military entities, among them <ENT0>the 101st Airborne Division</ENT0>. Engaged in ferocious combat, the valor of these troops was on full display under the aegis of the <ENT1>United States Army</ENT1>.<br>2. Last summer, the grand ceremony at the Capitol honored various noteworthy figures, including <ENT0>General Dwight D. Eisenhower</ENT0>, whose illustrious career and leadership were long-standing pillars of the <ENT1>United States Army</ENT1>.<br>3. On Veterans Day, numerous speeches commemorated those it was instituted to serve, like <ENT0>Sergeant John Doe</ENT0>, a brave soul who once operated under the proud tradition and command structure of the <ENT1>Marine Corps</ENT1>. |

Table 9: Prompt templates used in Definition-Based Seed Construction with LLM (Sec. 3.1). Example generated instances are based on the relation P241: <ENT1> was/is the military branch to which <ENT0> (a military unit, award, office, or person) belonged/belongs.

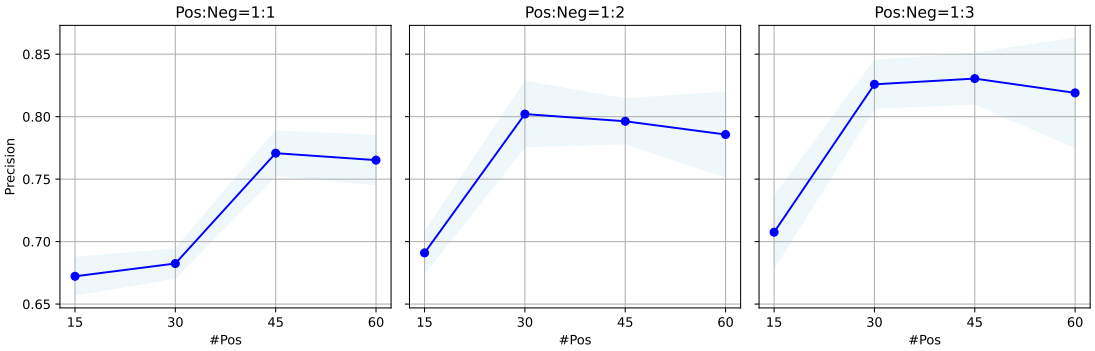


Figure 5: Precision (%) scores for different setups on the number and ratio of training instances.

## F.2 REPAL

We use `gpt-4o-2024-05-13` for GPT-4o and `gpt-4o-mini-2024-07-18` for GPT-4o mini. We have max tokens set to 4,096, presence penalty set to 0 and temperature set to 0.6. The SLM checkpoint used is ROBERTA-LARGE-MNLI.

For main experiments and analysis with REPAL, in addition to constructing the train set, we also construct a dev set with same number of positive instances and negative instances as the train set to automatically select the SLM model checkpoint for test. The dev positive instances are obtained by running the corresponding positive training instance generation with exactly the same prompt input and same setup for another time. The dev negative in-

**PROMPT TEMPLATE FOR FOLLOW-UP POSITIVE INSTANCE GENERATION**

Sampled examples which are predicted as positive by my relation extraction model are:

{feedback_examples}

Based on these predicted examples and your previously generated examples, generate {number_of_examples} additional examples (numbered from 1 to {number_of_examples}) expressing the same pre-defined relation: "{relation_definition}". Other requirements are:

1. Identify what relation patterns have been learnt by my model and what relation patterns have been covered by your previously generated examples. Your newly generated examples should have different and diverse relation patterns.
2. Identify model's bias from the sampled predicted examples which do not express the correct relation definition and your newly generated examples should try to mitigate the bias.
3. If the sampled predicted examples are uninformative, focus on the dialogue history, especially examples that were previously generated, to generate new examples with different and more diverse patterns.

Table 10: Prompt template used for follow-up positive instance generation in Feedback-Driven Instance Improvement and Bias Rectification (Sec. 3.3). Words highlighted denote the placeholder for filling in contents indicated by their surface names.

**PROMPT TEMPLATE FOR FOLLOW-UP NEGATIVE RELATION DEFINITION GENERATION**

A binary relation between entity placeholders <ENT0> and <ENT1> is defined by: "{positive_relation_definition}". In relation examples or relation instances, <ENT0> is replaced with actual entity mention and is prefixed with tag <ENT0> and suffixed with tag </ENT0> , <ENT1> is replaced with actual entity mention and is prefixed with tag <ENT1> and suffixed with </ENT1> .
Typical examples predicted as positive by current relation extraction model are:

{feedback_examples}

Based on the positive relation definition and the typical predicted examples, generate {number_of_negative_relations} negative binary relation definitions (numbered from 1 to {number_of_negative_relations}) in the same format as the above positive relation definition (including entity placeholders and entity type constraints). Other requirements are:

1. Identify false positive predictions from the typical predicted examples and your generated negative relations should teach model to mitigate such bias.
2. After addressing the previous requirement or if there is no false positive prediction, consider generating near-miss negative relations.

Table 11: Prompt template used for follow-up negative relation definition generation (without previously generated negative relation definitions) in Feedback-Driven Instance Improvement and Bias Rectification (Sec. 3.3). Words highlighted denote the placeholder for filling in contents indicated by their surface names.

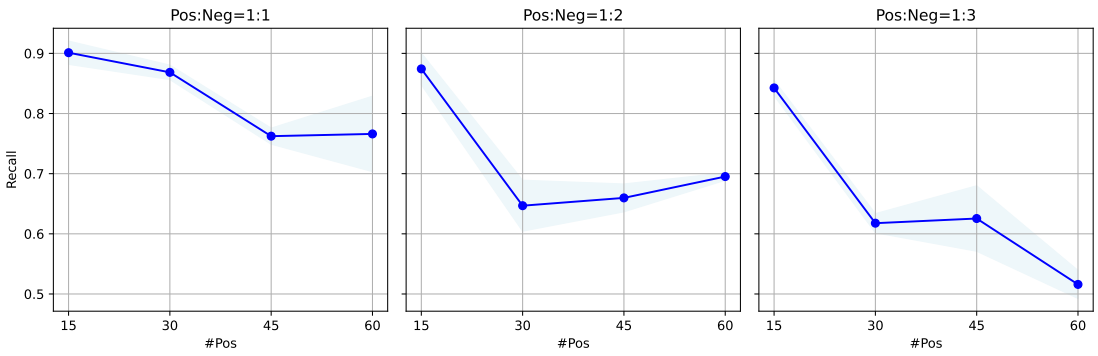


Figure 6: Recall (%) scores for different setups on the number and ratio of training instances.

stances are obtained by random sampling from the unlabeled corpus.

For the step of *Definition-Based Seed Construction with LLM and Unlabeled Corpus*, we follow the templates in Table 6, Table 7, and Table 8. For the results reported in Table 1, we choose the setting of 15 initial positive seeds and 15 initial negative seeds based on the analysis in Sec. 2.3. As we have three types of prompts for positive seeds generation, each prompt will contribute to gener-

**PROMPT TEMPLATE FOR FOLLOW-UP NEGATIVE RELATION DEFINITION GENERATION**

A binary relation between entity placeholders <ENT0> and <ENT1> is defined by: "{positive_relation_definition}" (in relation examples, <ENT0> is replaced with actual entity mention and is prefixed with tag <ENT0> and suffixed with tag </ENT0> , <ENT1> is replaced with actual entity mention and is prefixed with tag <ENT1> and suffixed with </ENT1> .).
Typical examples predicted as positive by current relation extraction model are:

{feedback_examples}

Existing generated negative relation definitions are:

{previously_generated_negative_relation_definitions}

Based on the positive relation definition, the typical predicted examples and the existing generated negative relation definitions, generate {number_of_negative_relations} additional negative binary relation definitions (numbered from 1 to {number_of_negative_relations}) in the same format as the above positive relation definition (including entity placeholders and entity type constraints). Other requirements are:

1. Identify false positive predictions from the typical predicted examples and your generated negative relations should teach model to mitigate such bias.
2. After addressing the previous requirement or if there is no false positive prediction, consider generating near-miss negative relations.
3. Your generated negative relation definitions should not be the same as existing negative relation definitions.

Table 12: Prompt template used for follow-up negative relation definition generation (with previously generated negative relation definitions) in Feedback-Driven Instance Improvement and Bias Rectification (Sec. 3.3). Words highlighted denote the placeholder for filling in contents indicated by their surface names.

| Original Dataset | Relation Label : Definition | Frequency |
|---|---|---|
| FewRel | P106: <ENT1> was/is the occupation of <ENT0> (a person) | 700 |
| FewRel | P1344: <ENT0> (a person or organization) was/is a participant of <ENT1> (an event) | 700 |
| FewRel | P136: <ENT1> was/is the genre or the field of work of <ENT0> (a creative work or an artist) | 700 |
| FewRel | P1411: <ENT1> was/is the award nomination received by <ENT0> (a person, organisation, or creative work) | 698 |
| FewRel | P241: <ENT1> was/is the military branch to which <ENT0> (a military unit, award, office, or person) belonged/belongs | 700 |
| FewRel | P26: <ENT1> was/is the married spouse (husband, wife, partner, etc.) of <ENT0> | 700 |
| FewRel | P276: <ENT1> was/is the location of <ENT0> (an object, structure or event) | 700 |
| FewRel | P3373: <ENT1> and <ENT0> had/have at least one common parent (<ENT1> is the sibling, brother, sister, etc. including half-sibling of <ENT0>) | 700 |
| FewRel | P40: <ENT1> was/is the child (not stepchild) of <ENT0> | 700 |
| FewRel | P400: <ENT1> was/is the platform or platform version for which <ENT0> (a work or a software product) was/is developed or released | 700 |
| FewRel | P410: <ENT1> was/is the military rank achieved by or associated with <ENT0> (a person or a position) | 700 |
| FewRel | P57: <ENT1> was/is the director(s) of <ENT0> (a film, TV-series, stageplay, video game or similar) | 700 |
| FewRel | P84: <ENT1> was/is the architect or architectural firm responsible for designing <ENT0> (a building) | 700 |
| FewRel | P974: <ENT1> was/is the watercourse that flowed/flows into <ENT0> (a watercourse) | 700 |

Table 13: Example DefOn-FewRel relation labels, definitions, and corresponding instance frequencies.

ating 5 initial positive seeds. For the step of *Pattern Learning with RE-Specialized SLM*, we train the SLM with 12 epochs using AdamW optimizer (Loshchilov and Hutter, 2017) with learning rate equal to 3e-5 and batch size equal to 64.

For the step of *Feedback-Driven Instance Im-*

**PROMPT TEMPLATE FOR INFERENCE WITH GPT-3.5 BASELINE**

A binary relation between entity placeholders <ENT0> and <ENT1> is defined by "{relation_definition}". In following relation instances, <ENT0> will be replaced with actual entity mention and prefixed with tag <ENT0> and suffixed with tag </ENT0> , <ENT1> will be replaced with actual entity mention and prefixed with tag <ENT1> and suffixed with </ENT1> . Now given an instance: "{instance_sentence_with_entities_enclosed_by_tags}", choose one option to answer: is the relation between two entities in the instance the same as the defined positive relation?
Option 1: Yes
Option 2: No

Answer:

Table 14: Prompt templates used for GPT-3.5 baseline in main experiments. Words highlighted denote the placeholders for filling in contents indicated by their surface names.

**PROMPT TEMPLATE FOR RE AS QA INFERENCE ON DEFINITION ONLY ZERO-SHOT RE**

A binary relation between entity placeholders <ENT0> and <ENT1> is defined by "{relation_definition}". In the following relation instances, <ENT0> will be replaced with actual entity mention and prefixed with tag <ENT0> and suffixed with tag </ENT0> , <ENT1> will be replaced with actual entity mention and prefixed with tag <ENT1> and suffixed with </ENT1> .
Given an instance, choose one option to answer: based on the new instance, do the entity mention enclosed by <ENT0> and </ENT0> and the entity mention enclosed by <ENT1> and </ENT1> follow the above positive relation definition?
The chosen option should come from:
Option 1: Yes
Option 2: No

Now answer:
Instance: {instance_sentence_with_entities_enclosed_by_tags}
Question: is the following statement true based on the instance: {relation_definition_filled_with_instance_entities}
Answer:

Table 15: Prompt templates used for RE AS QA baseline in main experiments following *Definition Only Zero-Shot Relation Extraction* setting. Words highlighted denote the placeholders for filling in contents indicated by their surface names.

*provement and Bias Rectification*[7], we first leverage the trained SLM to conduct inference on the unlabeled corpus after which each unlabeled instance will be associated with a score as the probability of being positive. Then, for follow-up positive instance generation, we conduct random sampling from all instance with score higher than 0.85 as we want the sampled feedback instances to reflect the model's learning outcome for the target positive relation. For follow-up negative relation definition generation, we conduct random sampling from all instances with score higher than 0.50 as we want to see both the confident predictions and less confident predictions to identify the existing bias. Note that continuing from the initial positive seed generation, there will be three threads of dialogue history for follow-up positive instance generation corresponding to three types of prompts in Table 6, Table 7, and Table 8 respectively. So for each thread of dialogue, we fill in different groups of sampled feedback instances to maximize the feedback coverage. For both the follow-up positive instance generation and the follow-up negative instance generation, the number of feedback instances for each prompt input is set to 10. For follow-up negative relation definition generation, we set the number of generated negative relation definitions to be 5 and the number of total follow-up negative instances to be 15. After obtaining the feedback-driven follow-up instances, we repeat the SLM training with all the accumulated training instances and all the other hyperparameters the same as our previous SLM training step.

## G LLM-Based Relation Definition Derivation

Our adopted prompt template for deriving relation definitions based on few-shot instances in Sec 2.2 is shown in Table 17. Note that we leveraged a fixed 3 relation definition demonstrations for in-context learning across all relations so that the LLM can give the relation definition in our desired format

[7]Note that for this step, we adopt GPT-4o mini as a parser to obtain the list of generated negative relation definitions or generated instances from the output string of synthesis LLM. This can help separate the generated definitions/instances from the additional output analysis.

**PROMPT TEMPLATE FOR RE AS QA INFERENCE ON MULTI-CLASS ZERO-SHOT RE**

Determine which option can be inferred from the given sentence.

Sentence: {instance_sentence}

Options:
A. {definition_of_relation_1_filled_with_instance_entities}
B. {definition_of_relation_2_filled_with_instance_entities}
...omitted for clarity...
N. {definition_of_relation_14_filled_with_instance_entities}

Respond with one letter from "A"-"N".

Table 16: Prompt templates used for RE AS QA baseline in main experiments following traditional multi-class classification zero-shot RE setting. Words highlighted denote the placeholders for filling in contents indicated by their surface names. This template is adapted from QA4RE (Zhang et al., 2023). This prompt template assumes the number of all relations is 14 for demonstration purpose.

| Prompt Name | Prompt Template |
|---|---|
| `Few-Shot Definition Derivation` | Given a list of relation instances/examples of a binary relation defined between two entities <ENT0> and <ENT1>, derive the relation definition in a single sentence. Note that in relation instances/examples, actual entity mention for <ENT0> is prefixed with tag <ENT0> and suffixed with tag </ENT0>, and actual entity mention for <ENT1> is prefixed with tag <ENT1> and suffixed with </ENT1> . Your derived relation definition should use entity placeholders <ENT0> and <ENT1> to refer to the two entities and the relation definition should try to contain entity type constraints. Example relation definitions are:\n\n1. <ENT1> is the league in which <ENT0> (team or player) plays or has played in.\n\n2. <ENT1> is the organization or person responsible for publishing <ENT0> (books, periodicals, printed music, podcasts, games or software).\n\n3. <ENT1> is the city, where <ENT0> (an organization)'s headquarters is or has been situated.\n\nThe list of relation instances/examples is:\n\n${Few-Shot Instances for One Relation}\n\n |
| `Train Instance Generation` | A binary relation between entity placeholders <ENT0> and <ENT1> is defined by "${Derived Relation Definition}". Under sentence-level relation extraction setting, generate additional ${Number Of Additional Examples to Generate} examples (numbered from 1 to ${Number Of Additional Examples to Generate} expressing the same relation, where <ENT0> is replaced with actual entity mention and is prefixed with tag <ENT0> and suffixed with tag </ENT0> , <ENT1> is replaced with actual entity mention and is prefixed with tag <ENT1> and suffixed with </ENT1> . ${Gold Few-Shot Examples for ICL} Do not overfit the pattern of the definition. Try as many different relation patterns or relation expressions as possible. |

Table 17: Prompt template used in deriving original relation definitions given few-shot relation instances and generating new relation instances based on the derived relation definition and gold few-shot instances (Sec. 2.2). Words in blue denote the placeholder for filling in contents indicated by their surface names.

for automatic parsing. After getting the relation definition, we use the prompt template in Table 17 to generate 15 instances for each derived relation. Note that the instance generation prompt is basically the same as `brief` prompt in Table 6 except that it integrates the gold few-shot instances as in-context learning demonstrations.

### G.1 Relation Definitions Derived by LLM From Few-Shot Instances

Table 18 and Table 19 show the LLM derived relation definitions based on the gold 4-shot instances. The table also contains the ground truth relation definitions for reference. We can see that for most of the FewRel relations, LLM successfully recovers the gold relation definitions. The derived definitions also reveal that one major difficulty is to specify the entity type constraints as few-shot instances may only convey a partial set of entity types which misguides LLMs to deduce a partial entity type constraints in the derived relation definitions.

### G.2 Macro F1 Scores of Few-Shot Method against Definition-Based Method

The macro $F_1$ scores of the experiments conducted in Sec. 2.2 are shown in Table 7. Since the DefOn-FewRel dataset is almost balanced, the micro $F_1$ and macro $F_1$ are close. So we put macro $F_1$ here for reference.

## H Fewshot Performance on DefOn-FewRel

We experiment inference with RE AS QA on one split of DefOn-FewRel dataset by increasing the

| Gold Definition | Gold Few-Shot Instances For Derivation | Derived Definition |
|---|---|---|
| <ENT1> was/is the occupation of <ENT0> (a person) | 1. <ENT0>Pierre Maudru</ENT0> ( 1892\u20131992 ) was a French <ENT1>screenwriter</ENT1> . Goble p.189 He also directed three films .<br>2. WWF Hall of Famer Bob Backlund and Extreme Championship Wrestling <ENT1>manager</ENT1> <ENT0>Bill Alfonso</ENT0> also made surprise appearances during the event .<br>3. In May 2010 , Paratici moved from Sampdoria to Juventus , along with Director General Giuseppe Marotta and <ENT1>Manager</ENT1> <ENT0>Luigi Delneri</ENT0> .<br>4. <ENT0>Else Reval</ENT0> ( 14 June 1893 \u2013 25 January 1978 ) was a German <ENT1>film actress</ENT1> . Giesen p.210 | <ENT1> is the profession in which <ENT0> (a person) works or has worked. |
| <ENT0> (a person or organization) was/is a participant of <ENT1> (an event) | 1. He only saw limited action in <ENT1>Euro 2000</ENT1> as cover for left - back <ENT0>Arthur Numan</ENT0> .<br>2. <ENT0>Francesco Cameli</ENT0> was a sailor from Italy , who represented his country at the <ENT1>1928 Summer Olympics</ENT1> in Amsterdam , Netherlands .<br>3. <ENT0>Giannin Andreossi</ENT0> ( born July 2 , 1902 , date of death unknown ) was a Swiss ice hockey player who competed in the <ENT1>1928 Winter Olympics</ENT1> .<br>4. <ENT0>Ren\u00e9 Sch\u00f6fisch</ENT0> ( born February 3 , 1962 ) is a German speed skater who competed for East Germany in the <ENT1>1984 Winter Olympics</ENT1> . | <ENT1> is the major international sports competition in which <ENT0> (an athlete) has competed. |
| <ENT1> was/is the genre or the field of work of <ENT0> (a creative work or an artist) | 1. Another version , dating from c. 1616 , was given in c. 1790 to <ENT0>Joshua Reynolds</ENT0> by the Duke of Leeds in exchange for a Reynolds self - <ENT1>portrait</ENT1> .<br>2. Teixeira is a former member of indie rock bands Ik Mux and Boris Ex - Machina , as well as the <ENT1>hip hop</ENT1> group <ENT0>Da Weasel</ENT0> and industrial metal band Bizarra Locomotiva .<br>3. Beautiful Stories for Ugly ChildrenMUSHROOMHEAD To Release ' Beautiful Stories For Ugly Children ' In September is the seventh studio album by <ENT1>industrial metal</ENT1> band <ENT0>Mushroomhead</ENT0> .<br>4. Wales is portrayed in the 1976 <ENT1>western film</ENT1> " <ENT0>The Outlaw Josey Wales</ENT0> " by actor and director Clint Eastwood . | <ENT1> is the genre or type of art (music, painting, film) associated with <ENT0> (an artist, band, or cultural artifact). |
| <ENT1> was/is the award nomination received by <ENT0> (a person, organisation, or creative work) | 1. On January 24 , 2012 , he was nominated for an <ENT1>Academy Award for Best Adapted Screenplay</ENT1> for the movie " <ENT0>Moneyball</ENT0> " .<br>2. " <ENT0>The Great Santini</ENT0> " received two Academy Award nominations : <ENT1>Best Actor in a Leading Role</ENT1> ( Duvall ) and Best Actor in a Supporting Role ( O'Keefe ) .<br>3. " <ENT0>Born This Way</ENT0> " ( 2011 ) , Gaga 's second studio album , accrued three nominations at the 54th Annual Grammy Awards , including her third consecutive nomination for <ENT1>Album of the Year</ENT1> .<br>4. As a producer , he has been nominated for <ENT1>Best Picture</ENT1> for three other films : " Raging Bull " , " <ENT0>The Right Stuff</ENT0> " , and " Goodfellas " . | <ENT1> is the award category for which <ENT0> (films, albums, or individuals associated with entertainment productions) has been nominated. |
| <ENT1> was/is the military branch to which <ENT0> (a military unit, award, office, or person) belonged/belongs | 1. General Sir ( William ) <ENT0>Henry Mackinnon</ENT0> , ( 15 December 1852 \u2013 17 March 1929 ) was a <ENT1>British Army</ENT1> General during World War I.<br>2. Lieutenant - Colonel <ENT0>Gordon Graham Donaldson</ENT0> was a senior officer in the <ENT1>British Army</ENT1> who died as a result of illness contracted during the disastrous Walcheren Campaign in 1809 .<br>3. <ENT0>Raphael Semmes</ENT0> was an officer in the <ENT1>United States Navy</ENT1> from 1826 to 1860 and the Confederate States Navy from 1860 to 1865 .<br>4. <ENT0>Isaac Townsend</ENT0> ( " c. " 1685 \u2013 21 November 1765 ) was an admiral in the <ENT1>British Royal Navy</ENT1> and a Member of Parliament . | <ENT1> is the military organization (such as an army or navy) with which <ENT0> (an individual, specified by their role or rank) has served or been associated. |
| <ENT1> was/is the married spouse (husband, wife, partner, etc.) of <ENT0> | 1. The film is about <ENT0>Carolyn Cassady</ENT0> 's recollection of life with husband <ENT1>Neal Cassady</ENT1> and Jack Kerouac , and her concern that the truth about these men is being lost in their mythos .<br>2. Maximilian married Duchess Helene in Bavaria , daughter of <ENT1>Duke Maximilian Joseph in Bavaria</ENT1> and his wife <ENT0>Princess Ludovika of Bavaria</ENT0> , on 24 August 1858 at Possenhofen Castle .<br>3. In 1916 his younger daughter , <ENT0>Nadejda</ENT0> ( " Nada " ) married <ENT1>Prince George of Battenberg</ENT1> , older son of Prince Louis by Queen Victoria 's granddaughter , Princess Victoria of Hesse - Darmstadt .<br>4. The fourth and youngest son of King <ENT1>John II of France</ENT1> and his wife , <ENT0>Bonne of Luxembourg</ENT0> , Philip was the founder of the Burgundian branch of the House of Valois . | <ENT1> is the spouse or partner of <ENT0> (an individual), indicating a marital, romantic, or partnership connection between the two entities. |
| <ENT1> was/is the location of <ENT0> (an object, structure or event) | 1. At the <ENT0>2014 Winter Olympics</ENT0> , Hudec won the bronze medal in the super - G at <ENT1>Rosa Khutor</ENT1> .<br>2. On the night of 22 January 1942 during the <ENT0>Battle of the Points</ENT0> , Japanese troops of the 16th Division attempted a landing on the west coast of southern <ENT1>Bataan</ENT1> .<br>3. Since the Netherlands did boycott the Moscow Olympic Games Brasser represented his National Olympic Committee at the <ENT0>1980 Summer Olympics</ENT0> in <ENT1>Tallinn</ENT1> , USSR under the Dutch NOC flag .<br>4. The bridge Norrbro stretches past the Riksdag on <ENT0>Helgeandsholmen</ENT0> and further south to <ENT1>Stockholm Old Town</ENT1> and the Royal Palace . | <ENT1> is the location or venue where <ENT0> (an event such as sports competitions, battles, or significant historical or cultural events) took place or was hosted. |
| <ENT1> and <ENT0> had/have at least one common parent (<ENT1> is the sibling, brother, sister, etc. including half-sibling of <ENT0>) | 1. Together they had three sons : Antonio , <ENT1>Arturo</ENT1> , and <ENT0>Alejandro</ENT0> .<br>2. Portuguese and Spanish conquerors made use of these weapons , including Vasco da Gama and his sons <ENT1>Crist\u00f3v\u00e3o da Gama</ENT1> and the younger brother <ENT0>Est\u00eav\u00e3o da Gama</ENT0> .<br>3. <ENT1>Arjuna</ENT1> was the fourth one to fall after Draupadi , <ENT0>Sahadeva</ENT0> and Nakula .<br>4. His nephews , Andr\u00e9 , <ENT0>Jordan</ENT0> and <ENT1>Rahim</ENT1> , also played the sport professionally . | <ENT1> is the sibling, specifically the brother, of <ENT0>. |

Table 18: Comparison between gold relation definitions and few-shot (4-shot) derived relation definitions (random seed=1).

number of in-context learning exemplars. The results are shown in Table 20. It shows that certain amount of in-context learning exemplars can lead to better performance while more exemplars do not necessarily result in better performance. It indicates that relation definitions are the most essential part of our *Definition Only Zero-Shot RE* setting as definitions convey more complete relation semantics while exemplars could guide models to better comprehend the task and enhance the reasoning process.

Besides, our work mainly focuses on zero-shot RE which is a different setting from few-shot RE, and most existing zero-shot RE works do not include few-shot PLM/LLM methods as comparison baselines. Meanwhile, LLM-based inference is

| Gold Definition | Gold Few-Shot Instances For Derivation | Derived Definition |
|---|---|---|
| <ENT1> was/is the child (not stepchild) of <ENT0> | 1. He was the son of Flemish painter <ENT1>Jan Massys , Matsys , or Metsys</ENT1> and the grandson and namesake of <ENT0>Quentin Massys or Metsys</ENT0> .<br>2. She married <ENT1>Lu Jing</ENT1> , who was born to <ENT0>Lu Kang</ENT0> and another daughter of Zhang Cheng ; both Sun He 's daughter and Lu Jing therefore were Zhang Cheng 's maternal grandchildren .<br>3. She is the wife of Bollywood actor , <ENT0>Jackie Shroff</ENT0> and mother of <ENT1>Tiger Shroff</ENT1> and Krishna Shroff .<br>4. His uncle was polymath <ENT0>Lionel Penrose</ENT0> , whose children include mathematician <ENT1>Oliver Penrose</ENT1> , polymath Sir Roger Penrose , chess grandmaster Jonathan Penrose , and geneticist Shirley Hodgson . | <ENT1> is a direct family member (such as a son, grandson, wife, or mother) of <ENT0>, specified by their familial relationship. |
| <ENT1> was/is the platform or platform version for which <ENT0> (a work or a software product) was/is developed or released | 1. The <ENT1>NES</ENT1> version of <ENT0>Shadowgate</ENT0> also carries the distinction of being one of the few NES games to be available in a Swedish language version .<br>2. In case of incidents <ENT0>Plumbr</ENT0> provides its users with information on problem severity , problem 's root cause location in source code or <ENT1>JVM</ENT1> configuration and lists steps needed to take to remediate the problem .<br>3. In 2013 , " <ENT0>Mega Man Xtreme</ENT0> " was made available on the Virtual Console of Japan 's Nintendo eShop for the <ENT1>Nintendo 3DS</ENT1> .<br>4. Prior to <ENT0>Windows 2000</ENT0> , Windows NT ( and thus PE ) supported the MIPS , Alpha , and <ENT1>PowerPC</ENT1> ISAs . | <ENT1> is the platform, console, or environment for which <ENT0> (software applications, games, or operating systems) is designed or available. |
| <ENT1> was/is the military rank achieved by or associated with <ENT0> (a person or a position) | 1. The son of Robert Langton Douglas , he was a half - brother to <ENT1>Marshal of the Royal Air Force</ENT1> <ENT0>William Sholto Douglas , 1st Baron Douglas of Kirtleside</ENT0> .<br>2. <ENT0>Dwight Edward Aultman</ENT0> , <ENT1>Brigadier General</ENT1> , United States Army .<br>3. He then served in the 27th U - boat Flotilla along with " <ENT1>Korvettenkapit00e4n</ENT1> " <ENT0>Erich Topp</ENT0> .<br>4. <ENT0>Axel Schimpf</ENT0> ( born 1 October 1952 ) is a retired " <ENT1>Vizeadmiral</ENT1> " ( vice admiral ) of the German Navy . | <ENT1> is the military rank of <ENT0> (a military personnel). |
| <ENT1> was/is the director(s) of <ENT0> (a film, TV-series, stage-play, video game or similar) | 1. Cummins 's photographs have been used extensively in cinema and TV documentaries including <ENT1>Grant Gee</ENT1> 's <ENT0>Joy Division</ENT0> and John Dower 's .<br>2. In 2014 , Zhang starred in <ENT1>Tsui Hark</ENT1> 's wuxia film " <ENT0>The Taking of Tiger Mountain</ENT0> " .<br>3. Starting her career in 2005 , she acted in the Malayalam film " <ENT0>Boyy Friennd</ENT0> " directed by <ENT1>Vinayan</ENT1> .<br>4. Kaif had her first success in Bollywood when she appeared opposite Salman Khan in <ENT1>David Dhawan</ENT1> 's romantic comedy " <ENT0>Maine Pyaar Kyun Kiya ?</ENT0> " . | <ENT1> is the director who directed the movie <ENT0>. |
| <ENT1> was/is the architect or architectural firm responsible for designing <ENT0> (a building) | 1. The <ENT0>Hotel Attraction</ENT0> project by Catalan architect <ENT1>Antoni Gaud\u00ed</ENT1> was built in 1908 in the parallel universe , whereas in our world it never went past initial planning .<br>2. Tampere Cathedral ( Lars Sonck , 1900 ) , <ENT0>National Museum</ENT0> , Helsinki ( <ENT1>Herman Gesellius , Armas Lindgren and Eliel Saarinen</ENT1> , 1902 ) .<br>3. Its designer was <ENT1>George Gilbert Scott</ENT1> , <ENT0>Busbridge Church</ENT0> \2013 Church of England Official gateway to the church .<br>4. He served a seven - year apprenticeship with <ENT1>Sir Charles Barry</ENT1> , the architect of the <ENT0>Houses of Parliament</ENT0> and Manchester Art Gallery . | <ENT1> is the architect or group of architects who designed <ENT0> (a building or architectural project). |
| <ENT1> was/is the watercourse that flowed/flows into <ENT0> (a watercourse) | 1. The <ENT1>Cerchez River</ENT1> is a tributary of the <ENT0>Ceair River</ENT0> in Romania .<br>2. The lake flows into the <ENT1>River Mangfall</ENT1> , a tributary of the <ENT0>River Inn</ENT0> and thence the River Danube .<br>3. The <ENT1>Veljul Mic River</ENT1> is a tributary of the <ENT0>Veljul Mare River</ENT0> in Romania<br>4. A small part of the district along the eastern boundary drains into the east - flowing <ENT1>River Loud</ENT1> , a tributary of the <ENT0>Hodder</ENT0> . | <ENT1> is a tributary of the <ENT0> (rivers or water bodies). |

Table 19: (Continued from Table 18) Comparison between gold relation definitions and few-shot (4-shot) derived relation definitions (random seed=1).

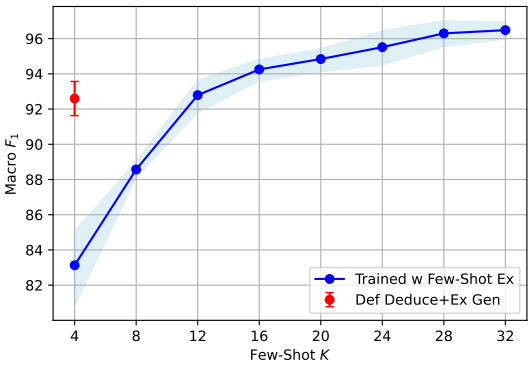


Figure 7: Macro $F_1$ (%) scores of model trained with few-shot instances (*Trained w Few-Shot Ex*) and model trained with instances from our relation definition derivation and instance generation approach (*Def Deduce+Ex Gen*). The error bar/band denotes averaged value $\pm$ standard deviation.

| ICL Demos | DefOn-FewRel | | |
|---|---|---|---|
| | Precision | Recall | $F_1$ |
| 0p0n | 87.26 | 69.76 | 72.00 |
| 2p2n | 75.15 | 86.43 | 76.81 |
| 5p5n | 70.35 | 93.33 | 76.60 |

Table 20: Results of RE AS QA on one split of DefOn-FewRel with increasing ICL demonstrations. 0p0n means zero-shot, while 2p2n means 2 sampled positive instance and 2 sampled gold negative instance as demonstration exemplars for each target relation. 5p5n means 5 sampled positive instance and 5 sampled gold negative instance as demonstration exemplars for each relation.

costly. REPAL could also replace its SLM with LLMs or introduce in-context learning strategies for better performance, but the increased computational cost and the reduced flexibility are against our low-resource positioning. Therefore, we stick with an efficient small encoder approach and leverage LLMs for zero-shot synthesis. The synthesized relation instances can be reused to train other relation extractors.

## I Case Study

In order to showcase how LLMs function in accordance with our design expectations, we provide some dialogue cases in Fig. 8 and Fig. 9. From the example dialogues, we can see that GPT-4o is considerably capable of identifying and summarizing the existing patterns in the dialogue history. The conversational feature also improves the interpretability of the results which can be further exploited with human-in-the-loop methods for future research or downstream applications.

It's also evident that feedback-driven negative relation definition generation is of great potential. As the feedback instances may contain false predictions which can be directly taken by LLM to generate new negative relations and effectively rectify the SLM's bias. Such feature directly boost the precision score as shown in Sec. 5.2. Hence, we can see that our design expectations are well fulfilled and this also qualitatively explains the performance boost of our proposed model.

Meanwhile, Fig. 9 also indicates that GPT-4o is still not perfect on giving the correct intermediate analysis of examining each of the feedback example. This indicates the future work can further improve the design of feedback acquisition and feedback processing.

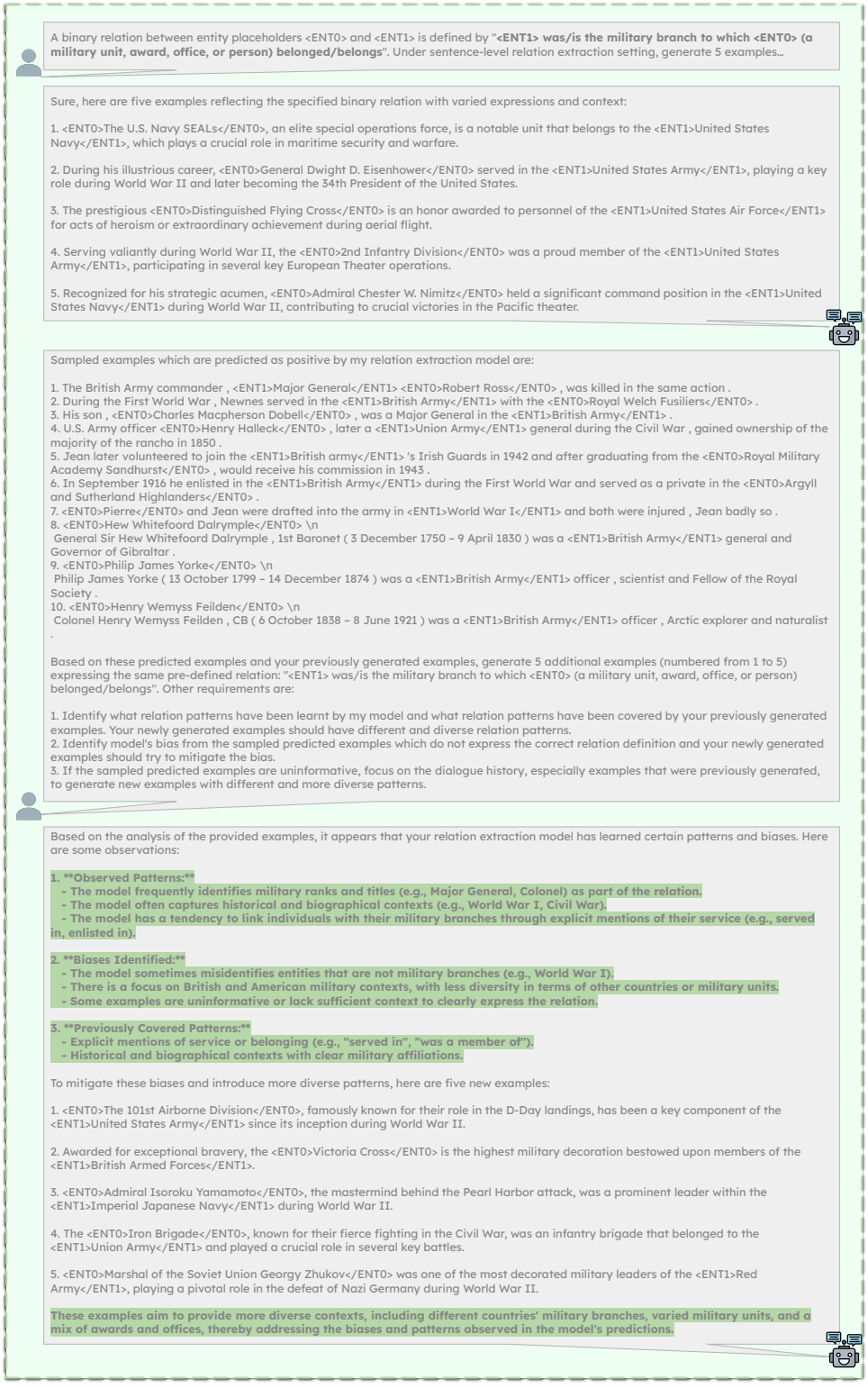


Figure 8: Example interaction dialogue which demonstrates the initial seed generation and feedback-driven follow-up positive instance generation.


Figure 9: Example interaction dialogue which demonstrates the feedback-driven generation of negative relation definitions.