Topic-Oriented Open Relation Extraction with *A Priori* Seed Generation

Linyi Ding* **Jinfeng Xiao*** **Sizhe Zhou** **Chaoqi Yang** **Jiawei Han**
University of Illinois at Urbana-Champaign
{linyid2, jxiao13, sizhez, chaoqiy2, hanj}@illinois.edu

Abstract

The field of open relation extraction (ORE) has recently observed significant advancement thanks to the growing capability of large language models (LLMs). Nevertheless, challenges persist when ORE is performed on specific topics. Existing methods give sub-optimal results in five dimensions: *factualness*, *topic relevance*, *informativeness*, *coverage*, and *uniformity*. To improve topic-oriented ORE, we propose a zero-shot approach called *PriORE*: Open Relation Extraction with *a Priori* seed generation. PriORE leverages the built-in knowledge of LLMs to maintain a dynamic seed relation dictionary for the topic. The dictionary is initialized by seed relations generated from topic-relevant entity types and expanded during contextualized ORE. PriORE then reduces the randomness in generative ORE by converting it to a more robust relation classification task. Experiments show the approach empowers better topic-oriented control over the generated relations and thus improves ORE performance along the five dimensions, especially on specialized and narrow topics.

1 Introduction

Relation extraction (RE) (Zhao et al., 2024), which aims to recognize semantic relations between entities under certain contexts, is essential for various downstream tasks such as knowledge graph construction (Shi and Weninger, 2018) and question answering (Yan et al., 2021). Traditional RE mostly focuses on the closed setting (Closed RE), where the task is to choose the best relation from a pre-defined relation set. Another line of work proceeds to the open relation extraction (ORE) task that extracts relations without being constrained by a pre-defined set.

Despite these developments, existing ORE approaches often give sub-optimal results on specific topics. Take Figure 1 as an example. Given a passage under a chemical topic *redox reactions*, existing methods fail to robustly extract the desired relation (*oxidized by*) between the entities *magnesium* and *oxygen*. Closed RE fails when the desired relation is not in the pre-defined set. Extractive ORE is restricted by the expression of the relation in contexts. It thus can be easily misled by topic-irrelevant relations (e.g., *exposed to* is not directly relevant to *redox reactions*) or implicit expressions. The generative method mostly utilizes the general knowledge in LLMs to interpret the context, leading to the generation of overly general relations (e.g., *forms compound with*).

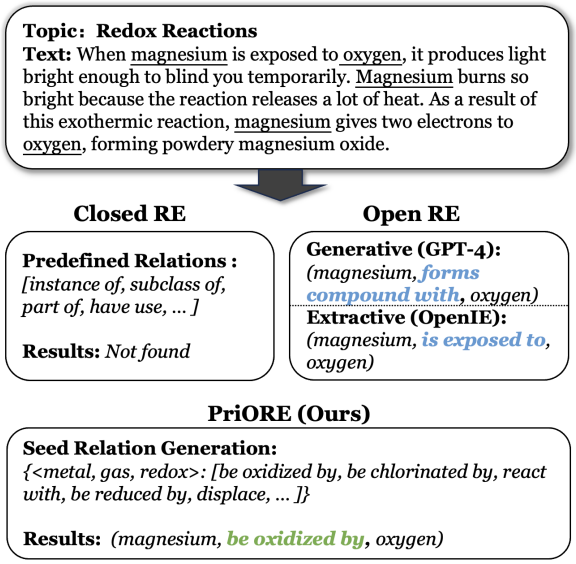

Figure 1: Extracted relations on the topic *redox reactions* for the entity pair *(magnesium, oxygen)* by closed RE, open RE and our topic-oriented RE.

An important yet less studied direction is how to make ORE more aligned with the users' needs under the specific topics (e.g., *redox reactions, electric vehicle batteries*). On the one hand, existing domain-specific ORE work (Aljamel et al., 2015) usually relies heavily on domain features and can

*Equal contribution

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 13834–13845

Dimension	Extracted Relation	Explanation
Factualness	gives to	Wrong given text and topic
Relevance	is exposed to	Correct given text but not directly relevant to topic
Informativeness	form compound with	Correct but conceptually too general given text and topic
Coverage	form powdery magnesium oxide	Correct but too specific, covering too few instances
Uniformity	{be oxidized by, give electrons to, ...}	Multiple correct expressions extracted for the same relation

Table 1: Existing ORE methods can cause errors in five dimensions. This example is a continuation of Figure 1 in extracting the relations between *magnesium* and *oxygen* under the topic *redox reactions*. The desired relation is *be oxidized by*.

hardly be generalized to out-of-domain topics. On the other hand, general methods (Wadhwa et al., 2023b; Angeli et al., 2015) are designed and evaluated without considering the alignment to user-specified topics.

Under specific topics, we summarize the main errors of existing ORE methods in five dimensions which cause the extracted relations to be sub-optimal (shown in Table 1). (i) Factualness: the extracted relations should be correct when examined with the text and the topic. (ii) Relevance: the relations should be directly relevant to the topic. (iii) Informativeness: the relations should contain proper levels of detail in context to avoid overly general expressions. (iv) Coverage: the relations should have general applicability under this topic to avoid overly specific expressions. (v) Uniformity: varied ways of expressing the same relation should have a uniform and consistent representation (i.e., synonym relations should be normalized to a single expression).

Some existing methods work well on certain dimensions. Take Figure 1 as an example. Although *forms compound with* meets the requirements of factualness, coverage, and topic relevance, it fails in informativeness. The extractive result *is expose to* also meets the factualness, while failing in topic relevance. No single mechanism has been optimized simultaneously across all five dimensions for topic-oriented ORE.

To address this gap, we propose *PriORE*[1], an approach that utilizes *a priori* seed generation to improve ORE for topic-oriented applications. The term *a priori* refers to: before considering any contexts for RE, PriORE first utilizes the LLMs' built-in knowledge about topics to generate seed relations based on topic-relevant entity types. We observe that the context-agnostic *a priori* generation can get more robust and coherent relations because randomness derived from context is alleviated. Type-based generation further enhances the robustness of long-tail scenarios. The seed relations are kept as a relation dictionary for the topic and can be dynamically updated. Then, with context, PriORE determines whether to choose a relation from the dictionary or update the dictionary with a new relation. In other words, PriORE converts ORE to relation classification (RC), which can reduce the randomness of ORE and generate more coherent expressions. The label space of RC is the seed relations first generated independent of the contexts and can be dynamically updated.

We evaluated PriORE on four topic-oriented datasets ranging from big, general topics to small, narrowly focused ones. We find that PriORE optimizes all five dimensions, especially informativeness, for topic-oriented ORE. Our advantage is particularly remarkable on specialized, narrow topics.

2 Related Work

Open RE Traditional open RE work can be categorized primarily into sequence labeling and clustering-based methods. Sequence labeling methods use syntactic or semantic features to extract relational phrases from the text as relations (Banko et al., 2007; Fader et al., 2011; Stanovsky et al., 2018; Cui et al., 2018). The relations can be limited in expressiveness, and this line of work can hardly capture global context. Clustering-based approaches model the relation representations and cluster them into relation types with masked language modeling (Wang et al., 2022; Li et al., 2022; Hogan et al., 2023) or relational feature extraction (Zhou et al., 2023). Although enhancing the coherence of expressions, they are sensitive to data quality, dataset size, and distribution of relations, especially for narrow topics.

Generative RE More recent works take advantage of generative models for ORE by formulating ORE into sequence-to-sequence tasks

[1]Our code is available at https://github.com/d01d01/PriORE.

(Huguet Cabot and Navigli, 2021; Ni et al., 2022), whose generalization to new topics is limited by training data. With the development of LLMs, Wadhwa et al. (2023a) and Jiang et al. (2024) show the effectiveness of LLMs on ORE. Other LLM-based RE methods focus on the closed-domain setting, with the paradigm of *filter-then-rerank* (Ma et al., 2023), the formulation of summarization (Zhang et al., 2023), or data synthesis (Li et al., 2023; Zhou et al., 2024).

Both generative and extractive methods for topic-oriented ORE may suffer from redundant/insufficient information and incoherent expressions. To align with users' needs and reduce ambiguity, some may apply relation canonicalization after extraction based on token-level statistics, entity information, triplet-level structure information, or external knowledge bases along with clustering methods (Galárraga et al., 2014; Vashishth et al., 2018; Putri et al., 2019). Instead of the post-extraction canonicalization, which is sensitive to data volume and distribution, our work can extract high-quality relations by *a priori* seed generation to convert the ORE task to a more robust relation classification task.

3 Method

3.1 Task Formulation

Different from general ORE, topic-oriented ORE takes a topic as additional input for guidance. The topic granularity can vary from general levels and domain levels (e.g., *chemistry*) to even more specialized and narrower ones (e.g., *redox reactions*).

Formally, the input consists of a topic τ and a set of unlabelled instances $\mathcal{D} = \{(s_i, h_i, t_i)\}_{i=1}^{N}$, where each instance contains a text passage s_i, a head entity h_i and a tail entity t_i. Our model, denoted as $P(r_i|\tau, s_i, h_i, t_i)$, aims to generate relations r_i to form the triples (h_i, r_i, t_i). The output of this task is the topic-oriented triple set $\mathcal{T} = \{(h_i, r_i, t_i)\}_{i=1}^{N}$.

We expect the generated relations to follow the principles of factualness, topic relevance, informativeness, coverage, and uniformity, as illustrated in Table 1.

3.2 Overall Framework

As Figure 2 shows, our proposed *PriORE* mainly consists of two parts: *A Priori Seed Relation Generation*, and *A Posteriori Seed-Guided Relation Extraction*. It first runs the *a priori* component to generate a relation dictionary without taking the topic-specific corpus. Then, it uses the *a posteriori* component to go through the corpus and extract relations.

3.3 *A Priori* Seed Relation Generation

Directly applying LLMs to extract relations from texts may lead to inferior results, as shown in Table 1. Some errors result from randomness in text generation with different contexts and entities as input. To reduce such randomness and improve the quality of relations, we generate candidate relations from entity types without considering contexts. Specifically, given a topic, PriORE initiates a *a dynamic relation dictionary* by querying LLMs to generate seed relations for possible pairs of topic-relevant entity types. An example for topic *Electrical Vehicle (EV) Batteries* is shown in the upper part of Figure 2.

Seed relations are high-quality candidate relations under a topic. For example, under topic *redox reactions*, *[be oxidized by, be reduced by, . . .]* can be a set of seed relations. As another example, *[be power source of, be recycled from, be managed by, . . .]* are high-quality seed relations under the *EV Batteries* topic.

We use a two-step approach to generate seed relations. First, we extract the topic-relevant entity hierarchy from external knowledge bases. Then, we utilize LLMs to generate relations from relevant entity types.

Motivations The reasons for using entity types instead of entities to generate seed relations include the following. (i) Directly generating relations from entities may lead to redundant phrases (e.g., *be oxidized by* and *give electrons to* have exactly the same meaning). The entity type space is much smaller than the entity space, which can significantly reduce the randomness of LLM generations. (ii) On narrow topics, understanding highly specialized or newly emerging entities may be hard for LLMs, while entity types can be easier. (iii) The hierarchy of entity types can enable different granularity in seed relations.

Entity Type Hierarchy Collection Given a topic, we collect the *topic-relevant hierarchies* of entity types from custom or existing external knowledge bases (e.g., Wikipedia Categories[2]). Starting from

[2] https://en.wikipedia.org/wiki/Wikipedia:Contents/Categories

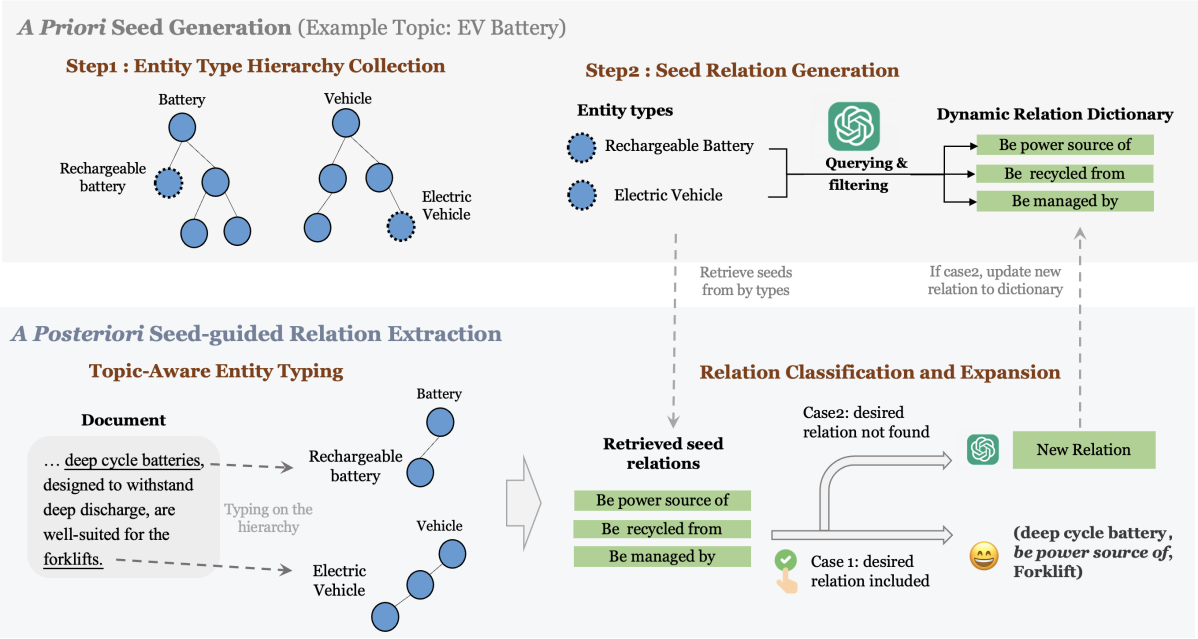

Figure 2: The Overview of our PriORE for the topic-oriented RE task. ***A Priori* Seed Generation**: Given the topic *EV battery*, we first initialize an entity type hierarchy and a seed relation dictionary for pairs of entity types. ***A Posteriori* Seed-guided Relation Extraction**: We then type the entities and retrieve seed relations from the dictionary as candidates for classification. In this case, the relation between the underlined entity pair is classified into *be power source of*. The relation dictionary can be dynamically updated if the desired relation is not found.

the topic node, we explore their relevant descendant and ancestor nodes as the initial type hierarchy. We clean the hierarchy by keeping the "is-a" relations and resolving the noise and inconsistency (Aouicha et al., 2016). Although the process may miss some necessary entity types of the topic, they can be dynamically discovered from the corpus in the subsequent relation extraction process.

Seed Relation Generation from Entity Types For a pair of entity types (T_1, T_2) on the topic hierarchy, we query the LLMs for the candidate relations under the specified topic as seed relations $(T_1, T_2, \tau) \rightarrow \mathcal{R}$. By providing the topic and entity types, the LLM can infer and generate the possible relations between two types in a topic-oriented way. The prompt we have used is shown in Appendix C.

To further enhance fault tolerance on narrow topics and provide more choices on different granularity, for a pair of nodes in the hierarchy, we also merge the relations generated from their ancestor nodes (i.e., parent types) into the set of seed relations $\mathcal{R}$. For example, the relations between *Battery* and *Vehicle* naturally apply to *Rechargeable battery* and *Electric vehicle*. To further improve the uniformity, we utilize a pre-trained sentence transformer (Reimers and Gurevych, 2019)[3] to check the relation similarity and filter out the relations that are highly similar to existing relations. The seed relations are used in the subsequent steps to guide the relation classification and expansion.

3.4 *A Posteriori* Seed-guided Relation Extraction

Given a topic-relevant document and the relation dictionary constructed by the *a priori* module, PriORE goes through the following two procedures to extract precise relations. (i) *Topic-Aware Entity Typing*. We first type the entities onto the type hierarchy. (ii) *Relation Classification and Expansion*. We use the entity types to retrieve the best matches from the relation dictionary and expand the dictionary as necessary.

Topic-Aware Entity Typing The type of entities can vary depending on the topic. For example, the entity *oxygen* can be typed as a *gas* under a chemistry topic, or as an *XML editor* under a computer science topic. Therefore, for topic-oriented ORE, we start with topic-aware entity typing.

Given the head or tail entity in the context, we first apply ZOE (Zhou et al., 2018), a zero-shot

[3]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

open entity typing method, to rank the types based on our *entity type hierarchy* to get the top K entity types. As an open method, ZOE is able to discover missing types of the topic hierarchy from text. We set a threshold to filter out the topic-relevant types based on their similarities to the topic. The similarity measurement is based on the sentence transformer[3]. The entity type hierarchy can be expanded by the newly discovered types if the entity has low similarity to all types in the hierarchy.

Relation Classification and Expansion Given a pair of entities and the context, we can retrieve the seed relations from the relation dictionary by their typing results. We then query LLMs to determine whether the target relation is included in the seed relations, and if so, which relation is the best match.

- **Case 1 (Classification):** If LLMs believe that a target relation r_i is in the retrieved seed relations, we can directly return the triple (h_i, r_i, t_i).

- **Case 2 (Expansion):** If none of the seed relations are suitable for the current entity pairs under the context (i.e., target relation is not found), then we query LLMs again with the context and seed relations to generate novel relations, which will be updated into the relation dictionary.

During expansion, to maintain the quality of newly generated relations, we query LLMs with the seed relations as examples. We also use the same similarity measurement to check the similarity of the new relation to all existing seed relations in the dictionary. To maintain the uniformity of the relation dictionary, a new relation is added only if it has low similarity to all existing ones.

All prompts we use are reported in Appendix C.

4 Experiment

4.1 Datasets

We evaluate our PriORE for topic-oriented RE task across different levels of topic specificity ranging from general to specific topics. For specific topics, we construct two topical datasets to show the performance on long-tail scenarios.

General-level At a general level, we use a general domain document-level RE dataset DocRED (Yao et al., 2019). The dataset has a relatively wide coverage of topics.

Domain-level At a domain-specific level, we use the FewRel Domain Adaption (FewRel DA) dataset[4] (Yao et al., 2022), and use the *"Biomedical"* as the domain-level topic.

Theme-level At a more specific level, we select two topics: *"Electric vehicle (EV) batteries"* and *"Redox reactions"*, which include long-tail terminologies and require theme-specified knowledge. The evaluation of the performance of PriORE on these topics aims to provide information on its potential to support real-world applications. For each theme, we collect instances from online databases. The datasets are annotated by domain experts.

Statistics and the annotation process of the datasets are reported in Appendix A.

4.2 Evaluation

4.2.1 Evaluation on Topic-Oriented ORE

It has been reported recently that traditional metrics like precision and recall fall short in evaluating generative RE (Jiang et al., 2024). Following our discussion in Section 1, we define the following metrics to evaluate the quality of extracted relations for the topic-oriented ORE task. For all metrics, the **higher**, the **better**.

Factualness We adopt the factualness score defined in previous work (Jiang et al., 2024) to evaluate the extent to which the extracted relation is supported by the context. It uses an LLM as a fact-checking tool.

Topic Relevance We query LLMs to decide if the extracted relations are directly relevant to the topic. The relevance score is defined by the percentage of extracted relations with a positive LLM response. The prompt template is given in Appendix C.

Informativeness This score assesses if relations contain sufficient details compared to the ground truth relations (e.g., *form a compound with* is less detailed than *be oxidized by*), which can be evaluated by querying LLM. The score is the percentage of informative relations in all instances. The prompt template is given in Appendix C.

Coverage This score evaluates the generality of a predicted relation (i.e., the capability of covering a significant proportion of instances in a topic-relevant corpus). For example, in Table 1, the predicted relation *form powdery magnesium oxide* has

[4]https://thunlp.github.io/2/fewrel2_da.html

a low coverage because it is only true in a limited number of instances. For each instance s_i, we consider all instances $\{s_j\}$ labeled with the same ground truth relation r_i as s_i, and calculate the percentage of instances in $\{s_j\}$ that support the relation $\hat{r_i}$ extracted from s_i. Formally, we define the coverage of $\hat{r_i}$ as follows:

$$C_{\hat{r_i}} = \frac{\sum_j \mathbb{1}\left(r_j = r_i\right) \mathbb{1}\left[f\left(\hat{r_i}, s_j\right)\right]}{\sum_j \mathbb{1}\left(r_j = r_i\right)} \quad (1)$$

where $\mathbb{1}(*)$ equals 1 if $*$ is true and 0 otherwise, $f\left(\hat{r_i}, s_j\right)$ denotes the factualness score (0 or 1) between $\hat{r_i}$ and the j-th instance s_j. The coverage score of a method on a dataset is the percentage of extracted relations that have a higher coverage than a predefined threshold t_c in Appendix A.

Overall Accuracy This score evaluates the *correctness* of the results with the above four metrics combined. A predicted relation is considered correct if it has good factualness, topic relevance, informativeness, and coverage.

Uniformity Different from the above correctness-focused metrics, this score inversely reflects the *redundancy* of the results. It is defined by the percentage of the most frequently predicted relation phrase in the group of instances labeled with the same ground-truth relation.

4.2.2 Evaluation on Seed Relations

Our method performs *a priori* generation of seed relations. To evaluate the quality of seed relations, we use the Precision (P), Recall (R), and F1 scores.

Precision The percentage of correct relations in generated seed relations. Correct relations refer to topic-relevant and logical relations between the given entity types.

Recall The percentage of ground-truth relations that are generated.

4.3 Experimented Methods

We report the results of the following methods. (1) `GPT-4`: the prompted GPT-4-1106-preview[5] (Achiam et al., 2023) for ORE task. (2) `LLama-3`: the prompted Llama-3-8B for ORE task[6], which is open-source with fewer parameters. Besides the topic-agnostic methods, we add two topic-aware methods based on GPT-4 and Llama-3 by providing topic information in prompts: (3) `Topic GPT-4`, and (4) `Topic Llama-3`. GPT-4-turbo is used as the LLM for evaluating the factualness, topic relevance, and informativeness of all methods.

To validate the effectiveness of methods, we further compared the following methods for ablation studies on GPT-4: (1) `PriORE+GPT4 w/o expan`, which removes the dynamic relation expansion of PriORE and merely uses seed relations for extraction. (2) `PriORE+GPT4 w/o type`, which directly uses entities instead of entity types to generate seed relations for extraction. The prompts we used for baselines can be found in Appendix C.

4.4 Results

Table 2 shows the correctness-based metrics for the experiment results on the general domain dataset, DocRED, and the biomedical domain dataset, FewREL DA. On the general domain, when our PriORE approach is applied with GPT-4 / Llama-3, we do not observe significant differences in the metrics compared with vanilla LLMs. However, notable improvement ($\sim 20\%$) over vanilla LLMs is observed in informativeness and overall accuracy on the biomedical topic. This reveals that when LLMs utilize their general world knowledge to interpret the context of specialized topics, they suffer from generating overly general results that miss topic-relevant, detailed information in the context.

We further report results on two narrowly focused, theme-level topics in Table 3. The margin of the improvement that PriORE brings to informativeness and overall accuracy now further grows to $\sim 30\%$. Moreover, PriORE starts to show an advantage of $\sim 10\%$ in the coverage metric without losing on the topic relevance and factualness metrics. As the topic narrows, the relations become more fine-grained, and correct extraction becomes more challenging for generative ORE, causing the performance of vanilla LLMs and topic-informed LLMs to decay significantly (Table 2 vs. Table 3). However, the performance of our PriORE remains stable and reliable when the topic granularity goes down from the biomedical domain to the battery/redox themes.

We also report the uniformity metric in Table 4. Our PriORE approach shows an improvement of $\sim 10\%$ on the general dataset and $\sim 20\%$ on domain-level and theme-level topics.

Directly adding the topic to the LLM queries fails to show significant and consistent improve-

[5]https://openai.com/gpt-4

[6]https://huggingface.co/meta-llama/Meta-Llama-3-8B

Table 2: Correctness-based metrics on coarse-grained topics. The metrics are abbreviated. Info: informativemess. TRel: topic relevance. Fac: factualness. Cov: coverage. Acc: overall accuracy. Since there's no limitation in topics on general datasets, the TRel of DocRED is omitted, and the topic LLMs are the same as vanilla LLMs for DocRED.

Method	**DocRED** (general)				**FewREL DA** (bio-medical)				
	Info	**Fac**	**Cov**	**Acc**	**Info**	**TRel**	**Fac**	**Cov**	**Acc**
Llama-3	0.91	**0.98**	0.92	0.82	0.57	0.97	0.88	0.98	0.49
Topic Llama-3	0.91	**0.98**	0.92	0.82	0.59	0.95	0.87	0.92	0.48
PriORE+Llama-3	0.90	0.92	0.97	0.81	0.79	0.98	0.85	0.97	0.70
GPT-4	0.89	0.96	0.94	0.78	0.61	0.94	0.90	0.91	0.49
Topic GPT-4	0.89	0.96	0.94	0.78	0.63	0.97	**0.90**	0.87	0.51
PriORE+GPT4 w/o type	**0.94**	0.93	0.86	0.80	0.66	0.97	0.79	0.83	0.52
PriORE+GPT4 w/o expan	0.86	0.88	**0.99**	0.74	0.74	**0.99**	0.75	**0.98**	0.64
PriORE+GPT4	0.90	0.94	0.98	**0.84**	**0.82**	0.98	0.88	0.94	**0.73**

Table 3: Correctness-based metrics on theme-level topics: Electric vehicle batteries, and redox reactions.

Method	**EV Battery**					**Redox Reaction**				
	Info	**TRel**	**Fac**	**Cov**	**Acc**	**Info**	**TRel**	**Fac**	**Cov**	**Acc**
Llama-3	0.49	0.89	0.82	0.84	0.37	0.47	0.84	0.85	0.86	0.38
Topic Llama-3	0.54	0.89	0.79	0.79	0.40	0.53	0.88	0.81	0.83	0.43
PriORE+Llama-3	0.82	0.92	0.81	0.90	0.69	0.77	0.87	0.83	0.95	0.67
GPT-4	0.52	0.89	0.84	0.81	0.40	0.48	0.85	0.87	0.80	0.37
Topic GPT-4	0.55	0.90	0.80	0.77	0.39	0.51	**0.93**	**0.89**	0.75	0.39
PriORE+GPT4 w/o type	0.63	0.88	0.75	0.80	0.45	0.59	0.86	0.71	0.77	0.36
PriORE+GPT4 w/o expan	0.68	**0.93**	0.74	**0.94**	0.61	0.65	0.90	0.68	**0.96**	0.57
PriORE+GPT4	**0.83**	0.91	**0.85**	0.90	**0.70**	**0.80**	0.92	0.88	0.92	**0.69**

Table 4: Evaluation of uniformity. Columns headers are abbreviated dataset names.

Method	**Doc.**	**Few.**	**EVB.**	**Redox.**
Llama3	0.41	0.25	0.27	0.20
Topic Llama3	0.41	0.22	0.25	0.24
PriORE+Llama3	**0.56**	**0.45**	0.46	0.41
GPT 4	0.43	0.26	0.27	0.25
Topic GPT	0.43	0.24	0.26	0.25
PriORE+GPT4	0.54	0.43	**0.47**	**0.42**

ment over topic-agnostic prompts, as indicated by the *Topic Llama3/GPT* rows in Tables 2 to 4. Therefore, the topic-oriented ORE task is harder than what simple prompt engineering can solve.

On the other hand, PriORE achieves state-of-the-art performance on all metrics and shows notable margins on some without directly instructing LLMs to optimize any of them. This shows that our design of *a priori* relation generation followed by *a posterior* relation classification is a systematic solution to topic-oriented ORE.

4.5 Ablation Study

We evaluate the quality of the dynamic relation dictionary in Table 5. The recall rate of the dynamic relation dictionary determines whether the ground truth relation can be retrieved and extracted. Without the expansion step (`w/o expan`), the recall rate of generated seed relations can already achieve $\sim 80\%$ on the general topic and $\sim 70\%$ on specific ones. For PriORE with expansion, the recall rate can achieve more than 90%. The precision achieves more than 80%. The wrong relations in the dictionary have limited effects on the final results because they are unlikely to be selected in the relation classification step. Columns under `w/o type` show the quality of relations generated from entities instead of entity types. Although the recall is high on general data, it drops to $\sim 55\%$ on the datasets with narrower topics. The overall performance of the ablated variants of PriORE (`w/o expan` and `w/o type`) is given in Tables 2 and 3. Using entities instead of types for relation generation (`w/o type`) significantly impairs the performance on domain-level and theme-level topics, while its effect on the general dataset is not that obvious.

4.6 Case Study

We analyze the performance of vanilla GPT4 and PriORE+GPT4 on relations with different granularity. We observe that PriORE brings major improvement to the informativeness especially on the specialized relations, as shown in Figure 3. The gen-

Table 5: Evaluation on precision, recall, and F1 of the dynamic relation dictionary.

Dataset	Method	PriORE			w/o expan			w/o type		
		P	R	F1	P	R	F1	P	R	F1
DocRED	PriORE+Llama-3	0.98	0.95	0.96	0.97	0.79	0.87	0.87	0.91	0.89
	PriORE+GPT4	0.95	0.96	0.95	0.94	0.86	0.89	0.90	0.91	0.90
FewREL DA	PriORE+Llama-3	0.85	0.91	0.88	0.86	0.61	0.71	0.71	0.58	0.64
	PriORE+GPT4	0.92	0.94	0.93	0.94	0.72	0.82	0.69	0.62	0.65
EV Battery	PriORE+Llama-3	0.83	0.88	0.85	0.85	0.62	0.72	0.68	0.57	0.62
	PriORE+GPT4	0.85	0.92	0.88	0.83	0.69	0.75	0.71	0.54	0.61
Redox Reaction	PriORE+Llama-3	0.84	0.85	0.84	0.83	0.61	0.70	0.70	0.49	0.58
	PriORE+GPT4	0.82	0.90	0.86	0.82	0.75	0.78	0.64	0.55	0.59

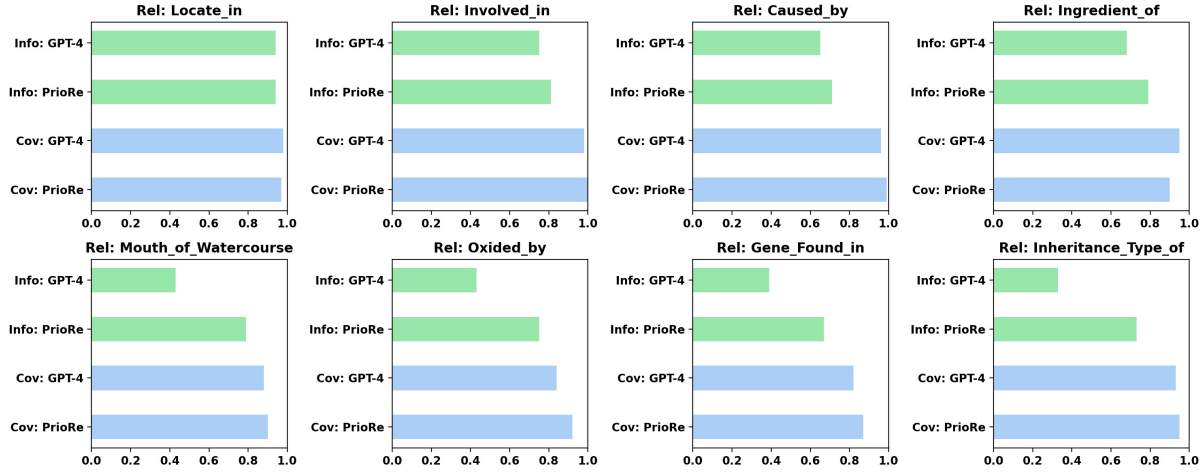

Figure 3: Case study. The informativeness and coverage of vanilla GPT4 and PriORE+GPT4 for relations with different granularity. The upper half shows some common and general relations. The lower half shows more specific relations.

Table 6: Some extracted relations for the ground truth relation *inheritance type of*. Red relations lose informativeness. Blue relations are correct. Note that the correct relations extracted by PriORE have better uniformity (i.e., fewer variations).

Method	Extracted Relations from Sampled Instances
GPT-4	(arterial tortuosity syndrome, **is a type of**, autosomal recessive), (tar syndrome, **is inherited in**, autosomal recessive), (hypohidrotic ectodermal dysplasia,**is transmitted as**, x-linked recessive trait), (Fanconi anemia, **is**, autosomal recessive) (cact deficiency, **is**, autosomal recessive)
PriORE	(arterial tortuosity syndrome, **is inherited through**, autosomal recessive), (tar syndrome, **is inherited through**, autosomal recessive), (hypohidrotic ectodermal dysplasia, **is inherited through**, x-linked recessive trait), (Fanconi anemia, **is associated with**, autosomal recessive) (cact deficiency, **is inherited through**, autosomal recessive)

eral and common relations (e.g., *locate in, involved in, caused by, ingredient of*) naturally have less information. Compared with specialized ones (e.g., *mouth of watercourse, oxidized by, gene found in, inheritance type of*), it is easier to achieve good informativeness. Therefore, on the upper half of Figure 3, we can see both vanilla GPT4 and PriORE+GPT4 can achieve higher informativeness scores than those in the lower half.

For general relations, PriORE brings marginal improvement to GPT-4, while for more specific ones, the advantage of PriORE becomes notable. The reason lies in the fact that GPT-4 can hardly control the granularity of the relations it generates. For example, GPT-4 sometimes generates "*is*" for the relation "*inheritance type of*" (Table 6), which misses some information. In some other cases, when GPT-4 generates overly specific relations that can hardly be applied to other instances, the coverage score is hurt.

We also show the advantage of PriORE in uniformity in Table 6 with examples. As the blue phrases demonstrate, PriORE keeps a coherent ex-

pression for the same relation, while vanilla GPT-4 is prone to generating multiple expressions (e.g., *is inherited in* and *is transmitted as*).

5 Conclusion

In this paper, we propose PriORE, a zero-shot approach for the task of topic-oriented ORE. Existing methods lead to inferior results on the dimensions of factualness, topic relevance, informativeness, coverage, and uniformity. By generating seed relations with LLMs for the topic in an *a priori* way, we reduce the randomness in generative LLMs by converting the ORE task into a more robust relation classification task. Experiments show we can outperform state-of-the-art generative LLMs along the five dimensions. Our advantage is particularly remarkable on specialized topics along the informativeness, coverage, and uniformity dimensions.

Limitation

Our method applies LLMs in the seed relation generation, relation classification, and relation expansion steps. It assumes that the topic is well discussed in the pre-training corpora of LLMs. If this is not the case, one may get inferior results. To optimize the results, one can fine-tune LLMs or apply retrieval-augmented generation (Lewis et al., 2020) on the topic-focused corpus before adopting our approach.

The entity types of the topics are the keys to generating seed relations. We assume that they are covered in external knowledge bases, which is largely true in many applications. For example, Wikipedia is sufficient for our experiments on the general-level, domain-level, and theme-level topics. For some applications requiring very fine-grained or specialized types that are not covered by Wikipedia, there may still exist suitable knowledge bases, e.g., the Microsoft Academic Graph (Wang et al., 2020) for scientific topics and the Medical Subject Headings (Lipscomb, 2000) for biomedical topics. Providing custom type hierarchies or applying automatic knowledge base construction methods can also help in case the desired types cannot be fetched.

As discussed throughout the paper and experimented on the general-level dataset DocRED, the improvement of our method mainly lies in topic-oriented ORE scenarios. In the general domain, we do not observe a significant advantage over vanilla state-of-the-art LLMs, which can be more convenient to use in such cases.

Acknowledgements

Research was supported in part by US DARPA INCAS Program No. HR0011-21-C0165 and BRIES Program No. HR0011-24-3-0325, National Science Foundation IIS-19-56151, the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) by NSF under Award No. 2118329, and the IBM-Illinois Discovery Accelerator Institute (IIDAI). 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 IBM, DARPA, or the U.S. Government.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Abduladem Aljamel, Taha Osman, and Giovanni Acampora. 2015. Domain-specific relation extraction: Using distant supervision machine learning. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, volume 1, pages 92–103. IEEE.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 344–354. The Association for Computer Linguistics.

Mohamed Ben Aouicha, Mohamed Ali Hadj Taieb, and Malek Ezzeddine. 2016. Derivation of "is a" taxonomy from wikipedia category graph. *Engineering Applications of Artificial Intelligence*, 50:265–286.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 2670–2676, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lei Cui, Furu Wei, and Ming Zhou. 2018. Neural open information extraction. *arXiv preprint arXiv:1805.04270*.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 1535–1545, USA. Association for Computational Linguistics.

Luis Galárraga, Geremy Heitz, Kevin Murphy, and Fabian M. Suchanek. 2014. Canonicalizing open knowledge bases. New York, NY, USA. Association for Computing Machinery.

William Hogan, Jiacheng Li, and Jingbo Shang. 2023. Open-world semi-supervised generalized relation discovery aligned in a real-world setting. *arXiv preprint arXiv:2305.13533*.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of EMNLP*, pages 2370–2381.

Pengcheng Jiang, Jiacheng Lin, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2024. Genres: Rethinking evaluation for generative relation extraction in the era of large language models. *arXiv preprint arXiv:2402.10744*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Junpeng Li, Zixia Jia, and Zilong Zheng. 2023. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. In *EMNLP*, pages 5495–5505.

Sha Li, Heng Ji, and Jiawei Han. 2022. Open relation and event type discovery with type abstraction. In *EMNLP*, pages 6864–6877.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

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! *arXiv preprint arXiv:2303.08559*.

Jian Ni, Gaetano Rossiello, Alfio Gliozzo, and Radu Florian. 2022. A generative model for relation extraction and classification. *arXiv preprint arXiv:2202.13229*.

Rifki Afina Putri, Giwon Hong, and Sung-Hyon Myaeng. 2019. Aligning open IE relations and KB relations using a Siamese network based on word embedding. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 142–153, Gothenburg, Sweden. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Baoxu Shi and Tim Weninger. 2018. Open-world knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Shikhar Vashishth, Prince Jain, and Partha Talukdar. 2018. Cesi: Canonicalizing open knowledge bases using embeddings and side information. In *Proceedings of the 2018 World Wide Web Conference*, pages 1317–1327.

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

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023b. Revisiting relation extraction in the era of large language models. *arXiv preprint arXiv:2305.05003*.

Jiaxin Wang, Lingling Zhang, Jun Liu, Xi Liang, Yujie Zhong, and Yaqiang Wu. 2022. MatchPrompt: Prompt-based open relation extraction with semantic consistency guided clustering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7875–7888, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413.

Yuanmeng Yan, Rumei Li, Sirui Wang, Hongzhi Zhang, Zan Daoguang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Large-scale relation learning for question answering over knowledge bases with pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3653–3660, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of ACL 2019*.

Yunzhi Yao, Shengyu Mao, Xiang Chen, Ningyu Zhang, Shumin Deng, and Huajun Chen. 2022. Schema-aware reference as prompt improves data-efficient relational triple and event extraction. *arXiv preprint arXiv:2210.10709*.

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of ACL*, pages 794–812.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Computing Surveys*, 56(11):1–39.

Ben Zhou, Daniel Khashabi, Chen-Tse Tsai, and Dan Roth. 2018. Zero-shot open entity typing as type-compatible grounding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2065–2076. Association for Computational Linguistics.

Sizhe Zhou, Suyu Ge, Jiaming Shen, and Jiawei Han. 2023. Corpus-based relation extraction by identifying and refining relation patterns. In *Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023, Proceedings, Part IV*, page 20–38, Berlin, Heidelberg. Springer-Verlag.

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

A Experimental Details

A.1 Datasets

Table 7: Statistics of datasets.

Datasets	Instances	Relations
DocRED	12275	96
FewREL DA	900	9
EVBattery	280	15
RedoxReaction	235	12

Data Statistics We construct two theme-level datasets: Electric Vehicle Batteries and Redox Reactions. We collect data from online sources and select relations for each theme. The data statistics are shown in Table 7.

Annotation Each theme has two expert annotators. The guideline of annotation follows the principles in Table 1. The primary annotator selects documents and performs the first round of annotation. Then, the predicted results from all models are pooled together and delivered to the primary annotator for a second round of annotation. After this, the secondary annotator checks the results and resolves disagreements with the primary annotator.

A.2 Evaluation Parameter

We set the coverage threshold t_c to 0.1, which yields results more consistent with preliminary manual assessments. Note that the threshold is an evaluation parameter rather than a model hyperparameter. When conducting coverage evaluation on different methods, it can be set to a reasonable value according to users' desired relation granularity.

B LLM-Based Evaluation

To establish the reliability of our LLM-based evaluation metrics, we report the alignment between the three LLM-computed metrics and human evaluation in Tables 8 and 9. The results are based on 120 sampled instances from all four datasets. Table 8 includes the average values of the corresponding metric over this sample. Table 9 evaluates the LLM-generated labels using the human-created labels as the ground truth. For example, for a relation instance, if LLM gives 1 for the factualness metric but the human gives 0, this instance will reduce the accuracy/precision/F1 values in the Factualness row. We observe a very high alignment between LLM-computed and human-labeled evaluation results.

Table 8: Mean metric values.

Metric	Human Labeled	LLM Labeled
Factualness	0.69	0.68
Informativeness	0.59	0.59
Topic Relevance	0.70	0.68

Table 9: Evaluation on automatically created $0/1$ labels with human labels as the ground truth.

Metric	Accuracy	Precision	Recall	F1
Factualness	0.98	0.99	0.98	0.98
Informativeness	0.96	0.97	0.97	0.97
Relevance	0.96	0.99	0.95	0.97

C Prompt Templates

All prompts we used are listed in Table 10.

Prompt Name	Prompt Template
`Type-Centric Generation`	You are a helpful assistant in generating all possible relations from entity types for knowledge base. *[Examples can be listed based on your need]*. Under the topic *[topic]*, what are the possible exclusive relations from *[type 1]* to *[type 2]*? List the possible relations in the format: (*[type 1]*, ____, *[type 2]*). Please strictly follow the format and only fill in the blank.
`Relation Classification`	You are a helpful assistant for relation classification about *[topic]*. I will give you the relation set and context. You should choose the most specific relation from the relation set that can best describe the accurate relation from *[entity 1]* to *[entity 2]* according to the context. Context: *[context]*. Relation set: *[triple 1, triple 2, . . . , none of the above (If the result is inaccurate according to the context, you should output none of the above)]*. Please direct output the result without other words.
`Dynamic Relation Expansion`	You are a helpful assistant for extracting relation of two entities in the context about *[topic]*. The example of relations can be: *[seed relations]*. Please extract the relation from *[entity 1]* to *[entity 2]* in the format of (*[entity 1]*, ___, *[entity 2]*). Context:*[context]*. Please strictly follow the format without other words.
`Informativeness`	You are a helpful assistance to evaluate the informativeness of relations. Given the topic *[topic]*, is the relation *[extracted relation]* having significantly more specific meaning than the relation *[ground truth relation]*? Output yes or no.
`Topic Relevance`	You are a helpful assistance of knowledge base construction for *[topic]*. Is "*[extracted relation]*" a relation directed related to the topic *[topic]*? Output yes or no.
`Topic-agnostic Baselines`	You are a helpful assistant for extracting relation of two entities in the context. Please extract the relation from *[entity 1]* to *[entity 2]* in the format of (*[entity 1]*, ___, *[entity 2]*). Context:*[context]*. Please strictly follow the format without other words.
`Topic-aware Baselines`	You are a helpful assistant for extracting relation of two entities in the context about *[topic]*. Please extract the relation from *[entity 1]* to *[entity 2]* in the format of (*[entity 1]*, ___, *[entity 2]*). Context:*[context]*. Please strictly follow the format without other words.

Table 10: Prompt templates used in this work. *Italic words* denote the placeholders for filling in contents indicated by their surface names.