

# Structured Semantic Information Helps Retrieve Better Examples for In-Context Learning Applied to Few-Shot Relation Extraction

Aunabil Chakma, Mihai Surdeanu, and Eduardo Blanco

University of Arizona, Tucson, AZ, USA

{aunabilchakma, msurdeanu, eduardoblanco}@arizona.edu

## Abstract

This paper presents several strategies to automatically obtain additional examples for in-context learning, effectively transforming relation extraction from a 1-shot to a few-shot setting. Specifically, we introduce a novel strategy for example selection, in which new examples are selected based on the similarity of their underlying syntactic-semantic structure to the provided 1-shot example. We show that our strategy results in complementary word choices and sentence structures compared to LLM-generated examples. When both strategies are combined, the resulting hybrid system achieves a more holistic picture of the relations of interest than either method alone. Our framework transfers well across datasets (FS-TACRED and FS-FewRel) and LLM families (Qwen and Gemma). Overall, our hybrid system consistently outperforms alternative strategies achieving state-of-the-art performance on FS-TACRED and strong gains on a customized FewRel subset.

## 1 Introduction

Relation extraction (RE) is a core NLP task that identifies relations between entity pairs (e.g., the place of birth or employer of a person). Extracting relations is often a component of other tasks such as knowledge graph construction (Schneider et al., 2022). In turn, knowledge graphs coupled with retrieval-augmented approaches have been shown to improve LLM reasoning (Sui et al., 2025) and reduce hallucinations (Wagner et al., 2025), thus potentially benefiting any end-user application including question answering (Xu et al., 2025).

Traditional relation extraction approaches require many training examples per relation (Marsh, 1998). More recently, few-shot relation extraction (FSRE) has received substantial attention (Sabo et al., 2021; Alam et al., 2024). In this setting, relations must be extracted given nothing but the relation name, a brief description, and a handful

of examples. Because of this challenging setting, previous work obtains relatively low results (e.g.,  $\approx 24$  F1 on FS-TACRED). Existing approaches to FSRE include those grounded in prototypes (Meng et al., 2023; Hu et al., 2025) and typically require training or fine-tuning (Section 2).

In-context learning (ICL) with LLMs (Brown et al., 2020) for FSRE is a natural choice yet how additional examples and diversity affect performance remains underexplored. In particular, the best strategy for example selection in FSRE remains unknown. To address this gap, in this paper we explore<sup>1</sup> several strategies to automatically obtain additional examples for the relation of interest. In particular, we explore two main strategies: using LLMs to generate new examples and retrieving examples from unannotated corpora *based on their deeper syntactic-semantic similarity* with the provided 1-shot example.

Figure 1 presents our strategies to obtain additional examples for 1-shot relation extraction. We automatically combine the 1-shot gold example with additional examples that are: (a) generated with LLMs or (b) retrieved from unannotated corpora using their syntactic-semantic similarity with the gold example. For LLM-based generation, we either paraphrase the gold example while preserving the subject, object, and relation, or generate new examples conditioned on the relation name and its natural language definition, encouraging alternative lexicalizations and sentence structures. Our key novelty is the use of lexico-syntactic rules in a semantic space for example retrieval. These structured rules capture how the subject and object are connected in a sentence through their core syntactic structure and semantic context (Section 4.2). Importantly, the semantic representations of these rules are learned in a self-supervised fashion that

<sup>1</sup>Code and data are available at <https://github.com/clulab/example-selection-icl-fsre>.

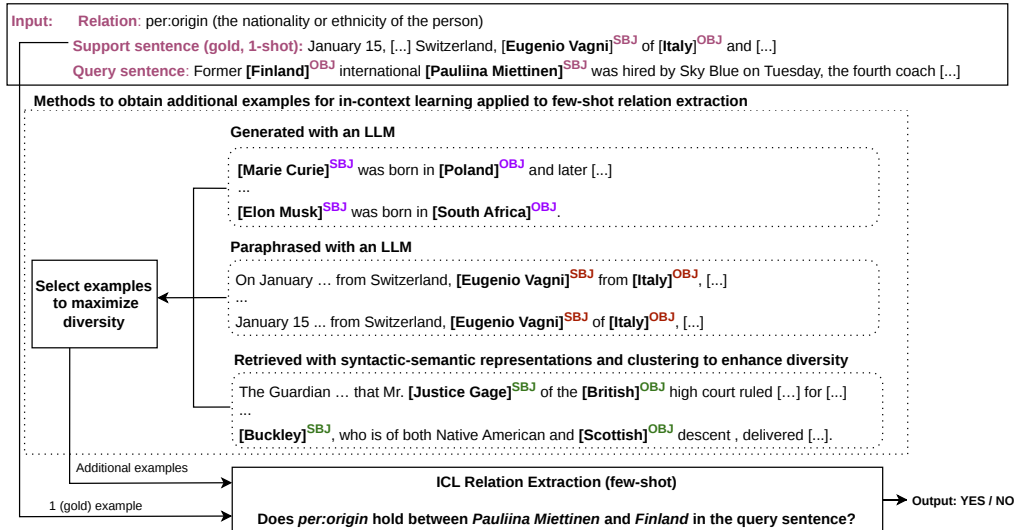


Figure 1: Our approach to in-context learning for 1-shot relation extraction. Our method complements the (gold) 1-shot ("Support sentence" in the figure) with 4 or 9 additional examples, turning the setup into a few-shot setting (5- or 10-shot), generated using shallower LLM-based methods (i.e., paraphrasing the gold example, or generating new examples based on the relation definition) or deeper methods using the underlying syntactic-semantic representations of candidate examples. When combining these strategies, the resulting few-shot relation extraction system achieves better results across two corpora (FS-TACRED and FS-FewRel) and two language models (Qwen3 and Gemma3).

is agnostic to the relations of interest, similar to word embeddings (Mikolov et al., 2013). Further, we emphasize diversity in the retrieved examples by: (a) clustering candidates using their semantic representations and selecting one representative example per cluster, and (b) using an LLM to select the most diverse examples from a pool of example candidates created with the above strategies.

The main contributions of this paper are:

- Several strategies to obtain additional examples for FSRE that range from shallower (e.g., LLM paraphrasing) to deeper (i.e., using representations of the syntactic-semantic structures of both the 1-shot gold example and the candidates themselves).
- Methods to emphasize example diversity (i.e., diverse word choices and sentence structures) from a pool of candidate examples constructed using the above strategies. We investigate two directions: (a) explicit clustering of examples and selecting one representative example per cluster, and (b) prompt engineering.
- Demonstrating that a hybrid approach balancing similarity and diversity obtains new state-of-the-art (SOTA) performance for FS-TACRED, especially with smaller language models, outperforming prior methods.

## 2 Related Work

Early approaches and datasets for RE targeted supervised learning (Marsh, 1998; Doddington et al., 2004). These approaches are effective when trained on hundreds or thousands of examples per relation (Miwa and Bansal, 2016; Zhou et al., 2005). However, they rarely generalize across domains (i.e., same relation types expressed in documents from different sources) and can only identify at inference time the relations they were trained on. Distant supervision (Mintz et al., 2009) has been explored to eliminate manual annotations, but other limitations remain.

**Few-Shot Relation Extraction** Most modern FSRE methods bypass the need to know the relations to be extracted at inference time and avoid manual annotations beyond a handful of examples per relation. Representative approaches include prototype-based classification using aggregated relation representations (Meng et al., 2023; Hu et al., 2025; Zhang et al., 2025; Li and Lyu, 2024; Luo et al., 2024). Embedding-based nearest-neighbor methods are also widely used (Sabo et al., 2021; Wan et al., 2023; Fei et al., 2022). While effective, these methods lack flexibility under relation or domain shifts. In contrast, neuro-symbolic approaches emphasize explicit relational

cues (Vacareanu et al., 2024; Singh and Blanco, 2024), which improves their in- and out-of-domain performance. These properties motivate our approach to retrieve examples grounded in syntactic-semantic representations. Moreover, most existing work on FSRE is evaluated on TACRED (Zhang et al., 2017), FewRel (Han et al., 2018), and NYT (Riedel et al., 2010). In this work, we focus on a more challenging variant introduced by Sabo et al. (2021), which contains a large proportion of NOTA (none-of-the-above) instances, making the task substantially harder (Section 5).

While we are not the first to use in-context learning and LLMs for FSRE (Ma et al., 2023; Xu et al., 2023), we are the first to use syntactic-semantic similarity to retrieve examples for ICL applied to FSRE. Our strategies for example selection improve both similarity and diversity. Moreover, we demonstrate that our approach transfers across benchmarks and LLMs, using the same prompt for each component across all models.<sup>2</sup>

### Example selection for In-Context Learning

Prior work has studied example selection for ICL applied to several NLP tasks including FSRE. Existing approaches mainly rely on embedding-based similarity (Wan et al., 2023; Mo et al., 2024) or learning-based selection (Scarlatos and Lan, 2024; Chen et al., 2024a; Wang et al., 2024) using task supervision, model feedback (Li and Qiu, 2023; Wang et al., 2024; Chen et al., 2024a), and diversity-aware criteria (Li and Qiu, 2023; Qin et al., 2024). In contrast, we leverage syntactic-semantic representations to guide example retrieval, an under-explored line of research in FSRE. LLM-based data generation techniques are also explored for RE to improve training and prediction (Li et al., 2025; Zheng et al., 2024). We explore LLM-based example generation as a simple baseline and also incorporate it into our hybrid method.

### Similarity and Diversity in Relation Extraction

Diversity plays a crucial role in in-context learning (ICL) (Xiao et al., 2025; Li et al., 2025; Wang et al., 2023a; Kapuriya et al., 2025). However, most prior work measures diversity using sentence-level embeddings, which often capture irrelevant context and surface variation rather than the differences in the relation holding between two entities. In con-

<sup>2</sup>We do not compare our results with Ma et al. (2023) and Xu et al. (2023) because they work with customized few-shot versions of TACRED and FewRel for which the exact few-shot evaluation setups are not publicly available.

trast, we (a) focus on the minimal context needed to express a relation by embedding the shortest path between entity pairs in the dependency-tree, and (b) enforcing diversity directly at the rule level, yielding a more informative similarity-diversity trade-off for relation extraction.

## 3 The Few-Shot Relation Extraction Task

Few-shot relation extraction is formulated as an episodic  $n$ -way  $k$ -shot classification task. A relation holds between two entities: the *subject* and *object*. Each episode consists of a  $k$ -shot learning instance defined by (a) a set of candidate relations  $\mathcal{R} = \{r_1, \dots, r_n\}$  along with  $k$  support sentences exemplifying each relation and (b) a query sentence with a subject and object named entities tagged. The task is to predict which of the  $n$  relations hold between the entities in the query sentence, or output `no_relation` if none apply. We target the standard 5-way 1-shot setting as defined in existing corpora (Section 5). In particular, we explore in-context learning and strategies to complement the single gold example per relation with automatically obtained additional examples.

### In-Context Learning for Few-Shot Relation Extraction

We reformulate the 5-way 1-shot relation extraction task as five independent 1-shot binary classifications. Each classification task predicts whether a relation holds between the subject and object in the query sentence (see few-shot prompt in Figure 2). We output `no_relation` if none hold, and a random relation out of all that hold if more than one is predicted.<sup>3</sup> Additionally, prior to in-context learning, we include a filter that enforces compatibility between the named entity types in the query sentence and the type constraints of the target relation. We use in-context learning (Figure 3) to match these named entity types. This filter was added because it substantially reduces compute (86.04% of candidate relations are discarded), and yields empirical benefits (Section 7).

We also experimented with in-context learning to decide which of the five candidate relations hold between the entities in the query sentence using a single LLM call, i.e., where the LLM functions as a multi-class classifier. We discarded it after observing consistently lower results than with the binary approach outlined above (Figure 4).

<sup>3</sup>This scenario is very unlikely: 0.26% of query sentences.

## 4 From 1-shot Relation Extraction to Many-Shot Relation Extraction

In this section, we propose strategies for automatically transforming the 1-shot relation extraction task into a many-shot task, where the additional examples are automatically obtained. We consider LLM-based strategies as well as retrieval-based strategies from unannotated corpora. Table 3 illustrates the additional examples obtained with all the strategies. As we shall see, a hybrid approach that balances similarity to the 1-shot example with contextual diversity is the most effective.

### 4.1 Obtaining Examples with LLMs

Our first strategies use LLMs to obtain additional examples and are arguably the simplest ones.

**Paraphrasing with LLMs:** We paraphrase the original (gold) support sentence for a relation with the prompt in Figure 5. The prompt is designed to ensure that (a) the same relation holds between the same named entities and (b) the subject and object entities remain tagged.

**Generating with LLMs:** We also explore a 1-shot example generation strategy with the prompt in Figure 6. This prompt uses the original (gold) support sentence as an example, and is designed to encourage different wordings and sentence structures while preserving the target relation. As we’ll see in Section 7, the LLMs tend to generate examples using their parametric knowledge (e.g., relations involving celebrities or politicians).

### 4.2 Retrieving Examples from Unannotated Corpora Using Syntactic-Semantic Representations

Our alternative to obtaining examples with LLMs is to retrieve them from large unannotated corpora. This approach has two advantages. First, retrieved examples reflect natural language usage patterns and can expose in-context learning to more realistic examples than LLM-synthesized examples. Second, retrieval can leverage precomputed representations to scale to millions of candidate sentences, enabling us to search for examples that are both semantically relevant and structurally diverse.

We retrieve examples from a subset of UMBC WebBase (Han et al., 2013), a corpus of over 3 billion words collected from a web crawl of 100 million web pages from more than 50,000 websites (Hirai et al., 2000). Specifically, we use the subset created by Vacareanu et al. (2024). After

removing duplicates and downsampling to ensure uniform frequency of named entity types, this subset contains 2.3M sentences containing two named entities belonging to types compatible with the corpora we work with (Section 5). Prior work has used this UMBC subset as a source of self-supervised training (Vacareanu et al., 2024). In contrast, we use this dataset to retrieve examples for ICL. The retrieval strategies always select examples based on cosine similarity. The main differences are (a) the representations for the gold support sentence and candidate sentences from UMBC (SBERT or syntactic-semantic rule representations) and (b) the methods to select candidate examples out of the closest ones.

**Sentence Representations from SBERT** We use SBERT (Reimers and Gurevych, 2019) as a baseline. Note that these representations are meant to embed the meaning of a sentence without emphasizing the subject and object named entities, the relation between them, or the context within the sentence that is most indicative of the relation.

**Syntactic-Semantic Rule Representations** In the real-world, it is often the case that a relation holds between two entities in a sentence, yet (a) only a portion of a sentence signals the relation between subject and object entities, and (b) context other than the tokens between subject and object entities are indicative of which relation holds (see examples in Table 3). Our rationale behind using syntactic-semantic representations is to pinpoint and leverage the specific context that is most beneficial for FSRE.

In particular, by lexico-syntactic rules we mean structured representations that encode the shortest syntactic dependency path between subject and object entities, combining lexical items, dependency relations, their directionality, and the named entity types of the subject and object entities. These paths are extracted from dependency parses produced by tools such as Stanford CoreNLP<sup>4</sup> or Odinson (Valenzuela-Escárcega et al., 2020), and are used in prior work (Vacareanu et al., 2024) to capture relational structure.

Semantic rule representations are created using the lexico-syntactic rule extractor and semantic rule matcher of Vacareanu et al. (2024). The rule matcher, named SoftMatcher, was trained following a self-supervised approach and is used in two

<sup>4</sup><https://stanfordnlp.github.io/CoreNLP/>

main stages during retrieval. First, semantic representations embedding the lexico-syntactic rule—not the sentence they originate from—are obtained for the (gold) support sentence and candidate sentences containing the same named entity types. Second, candidate sentences are selected based on cosine similarity between the rule representation of the support sentence and candidate sentences. The second step is facilitated at scale by a FAISS index (Douze et al., 2025) accounting for the tokens in a sentence, named entity types, and rule representation.

Consider the following support sentence for the *per:origin* relation: “Just this week we have the [European]<sup>OBJ</sup> G6 group, lead by the UK’s Home Secretary [John Reid]<sup>SBJ</sup>, promising to work together to make the internet a [...]” The lexico-syntactic rule is as follows: [entity=NATIONALITY]+ <amod group >acl:relcl lead >nmod\_by [entity=PERSON]+, which can be read as “a nationality named entity modifying (incoming syntactic dependency ‘amod’) *group*, which is modified by an outgoing dependency acl:relcl headed by *lead*, which is modified by (outgoing dependency ‘nmod’ headed by *by*) a person named entity.” After obtaining semantic rule representations, the SoftMatcher finds sentences with (a) the same named entity types and (b) different contexts expressing the same relation. For example, “[Buckley]<sup>SBJ</sup>, who is of both Native American and [Scottish]<sup>OBJ</sup> descent, delivered the sermon [...]” is the closest sentence obtained by rule representations.

**Enhancing Diversity** Selecting the most similar examples often yields redundant examples that express the relation in nearly identical ways (same entities, very similar word choices indicating the relation, etc.). To encourage diversity, we explore strategies to ensure differences among the selected examples. To this end, we first select the  $m$  candidate sentences above a similarity threshold  $\tau$  tuned with validation instances. Second, we use  $k$ -means or  $k$ -means++ to obtain  $k = \lfloor \sqrt{m} \rfloor$  clusters. Third, we select as many clusters as additional examples are desired by choosing (a) randomly, (b) the clusters closest to the support sentence, or (c) the clusters that maximize pairwise intra-cluster distances after initially selecting the furthest cluster to the support sentence. Fourth, we select the closest sentence to each cluster centroid as a representative ex-

ample. In total, we select examples following four diversity-oriented strategies ( $k$ -means with random or closest cluster selection, and  $k$ -means++ with farthest or closest cluster selection), which will be empirically validated (Section 6).

### 4.3 Hybrid Example Selection

Obtaining examples with LLMs and retrieving them from unannotated corpora offer complementary strengths. The former tend to have a very similar form to the original support sentence, while the latter represent more natural usage and diversity. We therefore explore a hybrid strategy that combines examples from both sources. Specifically, we shuffle the examples from both sources and select the desired amount of examples using the prompt in Figure 7. Intuitively, the hybrid approach balances the presence of (a) similar examples to the support sentences (obtained with LLMs) and (b) diverse examples with respect to the support sentence (retrieved from unannotated corpora).

### 4.4 Summarizing Examples

Prior work (Li et al., 2023) has shown that summarizing examples can improve in-context learning by reducing spurious content to the minimal content needed for the task at hand. In relation extraction, summarization would simplify examples to contain only the information needed to extract the relation of interest between the subject and object named entities. For example, *both Native American* and all tokens after *descent* could be deleted without altering the signal for the relation between *Buckley* and *Scottish* in “[Buckley]<sup>SBJ</sup>, who is of both Native American and [Scottish]<sup>OBJ</sup> descent, delivered the sermon during Saturday evening’s workshop service [...]” We experiment both with the examples (a) generated and retrieved and (b) their summarized versions with the prompt in Figure 8.

## 5 Experimental Setup

We experiment primarily with two datasets (FS-TACRED and FS-FewRel) and four LLMs (two sizes of two LLM families) as described below.

**Datasets** Our primary evaluation benchmark is FS-TACRED (Sabo et al., 2021), a few-shot reformulation of TACRED (Zhang et al., 2017). We use the five test episode files (all episodes are 5-way 1-shot) released by FS-TACRED. Each file has 10,000 episodes and 3 queries per episode. We do not train any model on these episodes. Instead, we

use the development episodes as held-out data to tune hyperparameters (see Section A.1 for details).

To test whether our strategies transfer across benchmarks, we also experiment with FS-FewRel, a few-shot version of FewRel (Han et al., 2018) we created following the algorithm by Alam et al. (2024). Because the labels for the test split are not publicly available, we report results with the development split. Specifically, we work with the six relations whose subject and object entities belong to named entity types compatible with the relation types in FS-TACRED. Similar to FS-TACRED, we construct 5 sets of 10,000 5-way 1-shot episodes, including a `no_relation` label, and each episode has three query sentences.

The majority label in both FS-TACRED and FS-FewRel is `no_relation` (97% and 95%). This is by design, as realistic few-shot relation extraction requires models to account for the most common case: none of the five relations hold between the entities in the query sentence.

We chose not to report results on FS-NYT (Alam et al., 2024) because it contains primarily relations between well-known named entities and locations (e.g., *place of birth* of famous people), which were very likely seen during the LLM pretraining. As a result, the vast parametric knowledge in pre-trained LLMs is sufficient to obtain much better results than with FS-TACRED and FS-FewRel with a prompt that provides the subject and object entities *and no additional context from the query sentence*. We refer the reader to Table 4 for the results to check for parametric knowledge within the three datasets.

**LLMs and Implementation Details** We design the pipeline to rely on a single LLM across all components (i.e., NER filtering, LLM-based generation and paraphrasing of examples, hybrid example selection, etc.), ensuring consistency, and simplifying deployment and transferability across different model families. Specifically, we evaluate the full system with four LLMs available in HuggingFace (Wolf et al., 2020): two sizes of two LLM families (Qwen3-4B, Qwen3-14B, Gemma3-4B, and Gemma3-12B). Further details on decoding and inference are provided in Section A.2, while hyperparameter tuning and retrieval efficiency are described in Section A.1.

**Baselines and Evaluation Metric** In addition to previous work on FS-TACRED (5-way, 1-shot setting), we consider a 1-shot ICL baseline that

considers the original support sentence as the only example. We compare this baseline against 1+4- and 1+9- shot ICL approach, where one example is the original support sentence and the other four or nine are obtained with the strategies described in Section 4. Following previous work, we report the average Precision, Recall, and F1 scores, along with their standard deviations and statistical significance results. Following previous work, we calculate metrics excluding `no_relation`.

## 6 Results

Tables 1 and 5 present the main results on FS-TACRED and FS-FewRel with Qwen3-4B and Qwen3-14B. The appendices present results with additional strategies to obtain additional examples (other clusterings, summarizing, etc.) on the same LLMs and datasets (Tables 6 and 7) as well as Gemma3-4B and Gemma3-12B (Tables 8 and 9). We draw several observations from these results.

### Structured semantic information helps select better examples for ICL

In three out of the four scenarios investigated using hybrid systems (i.e., two LLMs evaluated on two datasets) using examples retrieved with our semantic rule representations yield better overall results than all alternatives (1-shot baseline, using LLMs to generate or paraphrase, retrieving with SBERT representations). We conjecture that this is due to the fact that the examples retrieved with our rules (a) are realistic, as they are retrieved from a collection of real texts; (b) preserve the relation faithfulness better due to the underlying rules; and (c) are more diverse (see Section 7) than LLM-generated and LLM-paraphrased examples, which tend to stay closer to the original support sentence. An exception occurs with Qwen3-14B on FS-FewRel, where the LLM-generated examples slightly outperform our approach. This suggests that larger models sometimes incorporate more diversity in their generated examples than smaller ones. However, overall, structured semantic information yields better example selection for ICL in most settings.

### Retrieved examples complement LLM-generated examples

The hybrid strategy consistently ranks at or near the top (F1) with the smaller models on both datasets and one large model (Qwen3-14B) on FS-TACRED (Tables 1, 5, 8 and 9). This suggests that the two sources provide complementary benefits.

	#ex.	Qwen3-4B			Qwen3-14B		
		G+A	P	R	F1	P	R
Few-shot baseline	1+0	28.7 ± 0.99	17.8 ± 1.21	22.0 <sup>†</sup> ± 1.19	31.6 ± 0.96	40.3 ± 1.47	35.4 <sup>†</sup> ± 1.02
Including Additional examples							
Paraphrased with LLM	1+4	30.8 ± 1.53	08.6 ± 0.96	13.5 <sup>†</sup> ± 1.28	42.5 ± 1.17	24.7 ± 0.73	31.2 <sup>†</sup> ± 0.68
	1+9	31.4 ± 2.17	09.1 ± 0.83	14.1 <sup>†</sup> ± 1.16	41.0 ± 1.27	23.0 ± 0.66	29.5 <sup>†</sup> ± 0.82
Generated with LLM	1+4	26.8 ± 0.50	24.2 ± 1.34	25.4 <sup>†</sup> ± 0.92	37.3 ± 1.02	40.2 ± 1.73	<b>38.7*</b> ± 1.11
	1+9	27.8 ± 1.08	24.7 ± 1.14	<b>26.1<sup>†</sup></b> ± 0.84	38.6 ± 1.42	36.4 ± 1.55	37.5 <sup>†</sup> ± 1.10
Retrieved (closest) using SBERT representations	1+4	24.5 ± 1.47	15.4 ± 1.67	18.9 <sup>†</sup> ± 1.56	31.1 ± 1.14	36.0 ± 1.41	33.4 <sup>†</sup> ± 1.05
	1+9	24.9 ± 1.10	15.9 ± 0.88	19.4 <sup>†</sup> ± 0.83	30.4 ± 1.09	31.1 ± 0.87	30.7 <sup>†</sup> ± 0.76
+ hybrid	1+4	25.9 ± 0.58	23.4 ± 1.31	24.6 <sup>†</sup> ± 0.81	39.3 ± 0.70	39.4 ± 1.48	<b>39.3</b> ± 0.70
	1+9	26.6 ± 0.91	25.4 ± 0.79	<b>26.0<sup>†</sup></b> ± 0.57	39.3 ± 0.80	35.7 ± 1.40	37.4 <sup>†</sup> ± 0.86
Semantic rule representations	1+4	23.9 ± 0.48	23.3 ± 0.56	23.6 <sup>†</sup> ± 0.43	33.3 ± 1.15	38.6 ± 1.49	35.8 <sup>†</sup> ± 1.21
	1+9	22.9 ± 0.56	25.1 ± 0.68	23.9 <sup>†</sup> ± 0.50	33.0 ± 1.66	33.9 ± 2.08	33.4 <sup>†</sup> ± 1.76
+ hybrid	1+4	26.6 ± 0.87	26.7 ± 0.61	26.6* ± 0.54	39.4 ± 1.04	39.4 ± 1.27	<b>39.4</b> ± 0.89
	1+9	26.2 ± 0.60	28.9 ± 0.91	27.5 ± 0.48	40.0 ± 1.11	36.2 ± 1.12	38.0 <sup>†</sup> ± 0.82
and <i>k</i> -means clustering	1+4	22.8 ± 0.89	28.5 ± 1.21	25.3 <sup>†</sup> ± 0.96	29.5 ± 0.81	46.3 ± 1.18	36.0 <sup>†</sup> ± 0.83
	1+9	22.1 ± 0.75	30.6 ± 0.98	25.7 <sup>†</sup> ± 0.70	29.7 ± 0.89	45.1 ± 1.59	35.8 <sup>†</sup> ± 0.96
+ hybrid	1+4	25.7 ± 1.05	26.6 ± 0.68	26.1 <sup>†</sup> ± 0.70	37.9 ± 0.94	40.5 ± 1.64	39.1 ± 1.15
	1+9	26.4 ± 0.69	28.9 ± 1.88	<b>27.6</b> ± 1.20	38.7 ± 1.19	37.0 ± 1.36	37.8 <sup>†</sup> ± 0.84

Table 1: Results obtained with FS-TACRED using in-context learning in the 1-shot setting (i.e., one Gold example). Including Additional examples is always beneficial. Our hybrid approach to obtain additional examples (i.e., coupling LLM-generated examples with our semantic rule-based retrieval) yields the best results. Only the 4B model benefits from nine instead of four additional examples, and the hybrid approach is always beneficial. \* and <sup>†</sup> indicate the best method in the whole column is statistically significantly better at  $p < 0.05$  and  $p < 0.01$ , respectively, based on one-sided paired bootstrap test with replacement on F1 scores.

Table 10 provides insights into the selection process among LLM-generated and retrieved examples. On average, one example from the retrieved ones is picked to form the additional four-shot pool; and one to three examples from the retrieved ones are picked to form the nine-shot pool. This suggests the two approaches provide complementary strengths: retrieved examples bring realistic, relation-faithful contexts, while LLM-generated ones are simpler with strong semantic coverage around the support sentence. Combining these methods with hybrid strategy results in larger gains than using either source alone, underlining that hybrid selection is the most reliable way to boost ICL performance for few-shot relation extraction.

**Semantic rule representations yield better results than SBERT and paraphrasing with LLMs** Our semantic rule-based retrieval is consistently stronger than LLM paraphrasing and SBERT-based retrieval. In both FS-TACRED and FS-FewRel, paraphrased examples and SBERT-based retrieved examples lead to lower F1s, particularly for smaller models. On the other hand, semantic rule representation-based retrieved examples maintain better overall F1 (Tables 1, 5, 9, and 8). We ob-

serve that paraphrasing tends to result in additional examples that remain too close to the original support sentence, limiting diversity and reducing their effectiveness. In contrast, SBERT-based retrieval often selects examples that are topically similar but do not accurately preserve the target relation. Overall, semantic rule representations allow us to retrieve more diverse and relation-faithful instances, leading to better generalization in ICL.

**Semantic rule representations work better with smaller LLMs** The gains from including examples retrieved with our semantic rule representations are clear for the smaller Qwen3-4B and Gemma3-4B models, and modest for Qwen3-14B (Tables 1, 5, 9, and 8). The larger Gemma3-12B model shows no improvement with examples retrieved with any retrieval strategy. This suggests that structured example selection is most helpful when model capacity is limited, whereas the larger models exhibit more inconsistent behavior. Some models benefit more from internal example generation than retrieval, whereas others fail to benefit from either approach.

	#ex.	FS-TACRED		
		G+A	P	R
MNAV (Sabo et al., 2021)	1+0	-	-	12.4 ± 1.01
OdinSynth (Vacareanu et al., 2022a)	1+0	23.5 ± 1.46	11.5 ± 1.02	15.4 ± 1.21
CKPT (Lv et al., 2022)	1+0	-	-	15.1 ± 1.12
Anchor+gen. rules (Singh and Blanco, 2024)	1+0	19.6 ± 0.63	31.9 ± 1.04	24.2 ± 0.72
SoftRules (Vacareanu et al., 2024)	1+0	33.5 ± 1.47	19.7 ± 1.14	24.8 ± 1.22
Qwen3-4B, hybrid (ours)	1+9	26.4 ± 0.69	28.9 ± 1.88	<b>27.6 ± 1.20</b>

Table 2: Comparison of our best approach with Qwen3-4B on FS-TACRED (Table 1) and previous work. All methods only use one Gold support example. Even with the smaller (4B) Qwen, in-context learning with the gold example and nine Additional examples retrieved via our hybrid approach yields the new state of the art. The same approach using Qwen3-14B yields stronger results (F1: 37.8, Table 1).

### ICL with additional examples retrieved with semantic rule representations outperforms previous work

Table 2 compares previous work and our results with Qwen3-4B on FS-TACRED (hybrid approach, bottom row in Table 1). These prior methods are dedicated few-shot relation extraction approaches and do not rely on in-context learning; at inference time, they use one gold shot and no additional examples. Additionally, their approaches are based on fine-tuning relatively smaller models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and CodeT5 (Wang et al., 2023b).

Using one gold shot plus nine retrieved examples in our hybrid approach outperforms all prior work by a substantial margin (F1: 27.6 vs. 24.8). The improvements are even larger if we switch the LLM with Qwen3-14B (F1: 37.8, Table 1). These results show that adding genuine examples retrieved from unannotated corpora consistently outperforms prior work without requiring additional gold examples or model fine-tuning.

## 7 Qualitative Analyses

To shed more light on the full system and the benefits of retrieving examples with semantic rule representations, we carried out multiple qualitative analyses. We summarize these analyses below.

### Structured semantic retrieval yields more informative examples

Table 3 illustrates the additional examples obtained with LLMs and retrieved from corpora. LLM-generated examples are fluent but tend to be generic, mention well-known entities, and closely follow the same wording as the support sentence. On the other hand, paraphrasing mostly repeats the support sentence, offering limited differences beyond swapping a handful of words. SBERT-based retrieved examples are real-

istic but often only topically similar; additionally, they often miss the relation at hand. In contrast, due to its semantic underpinning, the examples retrieved with semantic rule representations include varied contexts that better express the target relation, making them more useful for ICL.

### Diverse examples do not always yield better results

As shown in Table 11, some methods (e.g., selected strategies under semantic rule-based retrieval) result in very diverse example sets (low token % overlap and cosine similarity), yet the most diverse examples are not always the top performers (Tables 1 and 5). Although diversified examples improve over the baseline, extremely diversified examples can drift away from the relation at hand, which negatively impacts performance.

### Additional examples are not always beneficial

For most settings, adding additional examples improves over the few-shot baseline, but the large model (Gemma3-12B) is an exception (Table 8 and 9). Indeed, on both datasets, none of the strategies (LLM-based, retrieval-based, or hybrid) improves over its 1-shot baseline performance, and several even degrade it. This indicates that larger models can be sensitive to prompt augmentation: they may already have enough task understanding that extra examples offer little benefit or can even be detrimental.

### Balanced similarity-diversity trade-off works best

Table 11 shows a clear trade-off between staying close to the (gold) support example and maintaining diversity among additional examples. LLM paraphrases are very similar to the gold support, having high token overlap and cosine similarity, which indicate low diversity. On the other hand, the retrieved example sets using  $k$ -means and its

variants maximize diversity but also can drift away from the support sentence and the relation. The best performance comes from the middle ground, i.e., the hybrid selection, which retains enough similarity to preserve the relation (through the generated examples) while introducing diversity (through the more realistic retrieved examples).

**Identifying multiple relations at once with a single ICL hurts performance** Table 12 shows that using a single prompt for multi-class relation classification yields much lower F1 than prompting each relation separately with a binary decision prompt (i.e., does this relation hold or not?). In the multi-class setting, especially for the smaller models, we observe errors such as predicting misspelled relation labels or hallucinated relations. We therefore use five binary prompts in our main experiments and did not pursue the single-prompt option further.

**NER-filtering as preprocessing consistently improves results** Table 13 presents results with and without the NER filtering step. Across datasets and models, the NER filtering consistently improves the F1 scores. Furthermore, removing this preprocessing step substantially increases computational cost, as more relations must be prompted per episode. Overall, NER filtering improves precision without a noticeable drop in recall.

**Computational Efficiency and Runtime Analysis** Our best approach, the hybrid method, increases the wall-clock runtime by only about 0.5% per episode compared to the LLM-generated baseline. In one episode, retrieval and clustering are performed for five relation instances, resulting in a total cost of 6.9s ( $1.38s \times 5$ ); this cost is independent of the underlying LLM model size, as it relies solely on embedding-based operations. Similarly, LLM-based example generation using Qwen3-4B is invoked five times, for a total of 8.95s ( $1.79s \times 5$ ). In contrast, the additional example selection step used in the hybrid approach is applied only once per episode, generates only a small list of 4 or 9 IDs, and adds just 0.049s of overhead, leading to negligible computational cost overall.

## 8 Conclusions

1-shot relation extraction is a challenging problem that is however critical in domains where response time matters, e.g., intelligence or medical. We study how to improve in-context learning in this

setting by selecting additional examples that effectively transform a 1-shot setup into a 5- or 10-shot setting without supervision or model training. Specifically, we propose an example selection framework that retrieves relation-faithful examples using structured syntactic-semantic representations, promotes diversity through clustering, and complements retrieval with LLM-generated examples. Together, these strategies yield a robust hybrid selection method that balances similarity to the gold example with diversity across examples.

Our hybrid example selection strategy transfers well across benchmarks (FS-TACRED and FS-FewRel) and LLMs (Qwen3-4B, Qwen3-14B, Gemma3-4B). Overall, our analyses show that explicitly balancing (a) similarity to the gold example via LLM-generated examples and (b) diversity via syntactic-semantic retrieval is key to effective example selection for in-context learning.

## Limitations

While our proposed approach is effective, some practical limitations remain. Not all relations naturally align across corpora; therefore, our experiments focus on the subset of FewRel relations that align with TACRED. In addition, different LLMs can respond differently to the same example set, suggesting that example selection could be further tailored to individual models for optimal performance. Finally, although semantic rule representations provide a principled way to retrieve examples, occasional inaccuracies in these rules may introduce less relevant demonstrations. Addressing these aspects offers directions for future work.

## Acknowledgements

We thank Robert Vacareanu for sharing data resources and for providing reference code and documentation that clarified the processing of the UMBC Index.

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

## References

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

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). *CoRR*, abs/2005.14165.
- Yunmo Chen, Tongfei Chen, Harsh Jhamtani, Patrick Xia, Richard Shin, Jason Eisner, and Benjamin Van Durme. 2024a. [Learning to retrieve iteratively for in-context learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7156–7168, Miami, Florida, USA. Association for Computational Linguistics.
- Yunmo Chen, Tongfei Chen, Harsh Jhamtani, Patrick Xia, Richard Shin, Jason Eisner, and Benjamin Van Durme. 2024b. [Learning to retrieve iteratively for in-context learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7156–7168, Miami, Florida, USA. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. [The automatic content extraction \(ACE\) program – tasks, data, and evaluation](#). In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. [The faiss library](#). *Preprint*, arXiv:2401.08281.
- Junbo Fei, Weixin Zeng, Xiang Zhao, Xuanyi Li, and Weidong Xiao. 2022. [Few-shot relational triple extraction with perspective transfer network](#). In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22*, page 488–498, New York, NY, USA. Association for Computing Machinery.
- Lushan Han, Abhay L. Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. [UMBC\\_EBIQUITY-CORE: Semantic textual similarity systems](#). In *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 44–52, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. [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.
- Jun Hirai, Sriram Raghavan, Hector Garcia-Molina, and Andreas Paepcke. 2000. [Webbase: a repository of web pages](#). *Comput. Netw.*, 33(1):277–293.
- Mengting Hu, Jianfeng Wu, Ming Jiang, Yalan Xie, Zhunheng Wang, Rui Ying, Xiaoyi Liu, Ruixuan Xu, Hang Gao, and Renhong Cheng. 2025. [Towards robust few-shot relation classification: Incorporating relation description with agreement](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8335–8349, Suzhou, China. Association for Computational Linguistics.
- Janak Kapuriya, Manit Kaushik, Debasis Ganguly, and Sumit Bhatia. 2025. [Exploring the role of diversity in example selection for in-context learning](#). In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25*, page 2962–2966, New York, NY, USA. Association for Computing Machinery.
- Guozheng Li, Peng Wang, and Wenjun Ke. 2023. [Revisiting large language models as zero-shot relation extractors](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6877–6892, Singapore. Association for Computational Linguistics.
- Xiaonan Li and Xipeng Qiu. 2023. [Finding support examples for in-context learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6219–6235, Singapore. Association for Computational Linguistics.
- Zexuan Li, Hongliang Dai, and Piji Li. 2025. [Generating diverse training samples for relation extraction with large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 713–726, Vienna, Austria. Association for Computational Linguistics.
- Zhiming Li and Yuchen Lyu. 2024. [GRADUAL: Granularity-aware dual prototype learning for better few-shot relation extraction](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13566–13577, Bangkok, Thailand. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.

- Da Luo, Yanglei Gan, Rui Hou, Run Lin, Qiao Liu, Yuxiang Cai, and Wannian Gao. 2024. [Synergistic anchored contrastive pre-training for few-shot relation extraction](#). *Preprint*, arXiv:2312.12021.
- Bo Lv, Li Jin, Yanan Zhang, Hao Wang, Xiaoyu Li, and Zhi Guo. 2022. [Commonsense knowledge-aware prompt tuning for few-shot nota relation classification](#). *Applied Sciences*, 12(4).
- Xilai Ma, Jing Li, and Min Zhang. 2023. [Chain of thought with explicit evidence reasoning for few-shot relation extraction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2334–2352, Singapore. Association for Computational Linguistics.
- Elaine Marsh. 1998. [TIPSTER information extraction evaluation: The MUC-7 workshop](#). In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*, pages 233–234, Baltimore, Maryland, USA. Association for Computational Linguistics.
- Shiao Meng, Xuming Hu, Aiwei Liu, Shuang Li, Fukun Ma, Yawen Yang, and Lijie Wen. 2023. [RAPL: A relation-aware prototype learning approach for few-shot document-level relation extraction](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5208–5226, Singapore. Association for Computational Linguistics.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13*, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.
- Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. [Distant supervision for relation extraction without labeled data](#). In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.
- Makoto Miwa and Mohit Bansal. 2016. [End-to-end relation extraction using LSTMs on sequences and tree structures](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany. Association for Computational Linguistics.
- Ying Mo, Jiahao Liu, Jian Yang, Qifan Wang, Shun Zhang, Jingang Wang, and Zhoujun Li. 2024. [C-ICL: Contrastive in-context learning for information extraction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10099–10114, Miami, Florida, USA. Association for Computational Linguistics.
- Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. 2024. [In-context learning with iterative demonstration selection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7441–7455, Miami, Florida, USA. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). *Preprint*, arXiv:1908.10084.
- Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Ofer Sabo, Yanai Elazar, Yoav Goldberg, and Ido Dagan. 2021. [Revisiting few-shot relation classification: Evaluation data and classification schemes](#). *Transactions of the Association for Computational Linguistics*, 9:691–706.
- Alexander Scarlatos and Andrew Lan. 2024. [Reticl: Sequential retrieval of in-context examples with reinforcement learning](#). *Preprint*, arXiv:2305.14502.
- Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. 2022. [A decade of knowledge graphs in natural language processing: A survey](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 601–614, Online only. Association for Computational Linguistics.
- Mayank Singh and Eduardo Blanco. 2024. [Learning to generate rules for realistic few-shot relation classification: An encoder-decoder approach](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5907–5921, Miami, Florida, USA. Association for Computational Linguistics.
- Yuan Sui, Yufei He, Nian Liu, Xiaoxin He, Kun Wang, and Bryan Hooi. 2025. [FiDeLiS: Faithful reasoning in large language models for knowledge graph question answering](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8315–8330, Vienna, Austria. Association for Computational Linguistics.
- Robert Vacareanu, Fahmida Alam, Md Asiful Islam, Haris Riaz, and Mihai Surdeanu. 2024. [Best of both worlds: A pliable and generalizable neuro-symbolic approach for relation classification](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2576–2594, Mexico City, Mexico. Association for Computational Linguistics.
- Robert Vacareanu, George C.G. Barbosa, Enrique Noriega-Atala, Gus Hahn-Powell, Rebecca Sharp,

- Marco A. Valenzuela-Escárcega, and Mihai Surdeanu. 2022a. [A human-machine interface for few-shot rule synthesis for information extraction](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 64–70, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.
- Robert Vacareanu, Marco A. Valenzuela-Escárcega, George Caique Gouveia Barbosa, Rebecca Sharp, Gustave Hahn-Powell, and Mihai Surdeanu. 2022b. [From examples to rules: Neural guided rule synthesis for information extraction](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6180–6189, Marseille, France. European Language Resources Association.
- Marco A. Valenzuela-Escárcega, Gus Hahn-Powell, and Dane Bell. 2020. [Odinson: A fast rule-based information extraction framework](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2183–2191, Marseille, France. European Language Resources Association.
- Robin Wagner, Emanuel Kitzelmann, and Ingo Boersch. 2025. [Mitigating hallucination by integrating knowledge graphs into LLM inference – a systematic literature review](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 795–805, Vienna, Austria. Association for Computational Linguistics.
- Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. [GPT-RE: In-context learning for relation extraction using large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547, Singapore. Association for Computational Linguistics.
- Liang Wang, Nan Yang, and Furu Wei. 2024. [Learning to retrieve in-context examples for large language models](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767, St. Julian’s, Malta. Association for Computational Linguistics.
- Qing Wang, Kang Zhou, Qiao Qiao, Yuepei Li, and Qi Li. 2023a. [Improving unsupervised relation extraction by augmenting diverse sentence pairs](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12136–12147, Singapore. Association for Computational Linguistics.
- Yue Wang, Hung Le, Akhilesh Gotmare, Nghi Bui, Junnan Li, and Steven Hoi. 2023b. [CodeT5+: Open code large language models for code understanding and generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1069–1088, Singapore. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Jianfeng Wu, Mengting Hu, Yike Wu, Bingzhe Wu, Yalan Xie, Mingming Liu, and Renhong Cheng. 2023. [Density-aware prototypical network for few-shot relation classification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2477–2489, Singapore. Association for Computational Linguistics.
- Wenyang Xiao, Haoyu Zhao, and Lingxiao Huang. 2025. [The role of diversity in in-context learning for large language models](#). *Preprint*, arXiv:2505.19426.
- Derong Xu, Xinhang Li, Ziheng Zhang, Zhenxi Lin, Zhihong Zhu, Zhi Zheng, Xian Wu, Xiangyu Zhao, Tong Xu, and Enhong Chen. 2025. [Harnessing large language models for knowledge graph question answering via adaptive multi-aspect retrieval-augmentation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25570–25578.
- Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. [How to unleash the power of large language models for few-shot relation extraction?](#) In *Proceedings of the Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 190–200, Toronto, Canada (Hybrid). Association for Computational Linguistics.
- Liang Zhang, Yang Zhang, Ziyao Lu, Fandong Meng, Jie Zhou, and Jinsong Su. 2025. [A self-denoising model for robust few-shot relation extraction](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26782–26797, Vienna, Austria. Association for Computational Linguistics.
- Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. [Position-aware attention and supervised data improve slot filling](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.
- Yifan Zheng, Wenjun Ke, Qi Liu, Yuting Yang, Ruizhuo Zhao, Dacheng Feng, Jianwei Zhang, and Zhi Fang. 2024. [Making llms as fine-grained relation extraction data augmentor](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6660–6668. International Joint Conferences on Artificial Intelligence Organization. Main Track.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. [Exploring various knowledge in relation extraction](#). In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 427–434, Ann Arbor, Michigan. Association for Computational Linguistics.

## A Appendix

### A.1 Hyperparameter Tuning and Retrieval Efficiency

We use the FS-TACRED development episodes as held-out data to tune hyperparameters. In practice, the only tunable parameter in our framework is the similarity threshold  $\tau$ , which controls the pool of candidate examples selected for clustering. We tune  $\tau$  on FS-TACRED development episodes and fix it to  $\tau = 0.6$  in all experiments. The same threshold value is applied when evaluating on FS-FewRel, and no additional hyperparameters are tuned on the FS-FewRel development set. All other components, including the clustering algorithms and cluster selection strategies, use fixed settings as described in Section 4.

To enable efficient retrieval at scale, cosine similarity search over precomputed sentence and rule embeddings is implemented using FAISS (Douze et al., 2025), allowing all candidate similarities to be computed offline and reused during inference. This design keeps inference efficient and avoids introducing additional tunable parameters.

### A.2 Decoding and Inference Details

The relation inference and named entity recognition tasks are formulated as binary classifications using the logits of the first generated token, restricted to “yes” and “no”. A prediction of “yes” is made when  $P(\text{yes}) \geq P(\text{no})$ , and “no” otherwise. For all generation-related components, we use the default decoding configurations of the underlying language models. For Qwen models, we use `do_sample=True`, `temperature=0.6`, `top_p=0.95`, and `top_k=20`, following the model defaults. For Gemma models, we use `do_sample=True`, `top_p=0.95`, `top_k=64`, and the default temperature setting (`=1.0`). We do not truncate input prompts, while generated outputs are capped at a maximum of 1,000 tokens, which is sufficient for all prompts and tasks considered in this work. These settings follow standard model usage and avoid introducing additional decoding hyperparameters that could confound comparisons across strategies.

---

You are given below a Relation name, a Description of the relation between brackets, N Support sentences exemplifying the relation, and a Query sentence.

A relation connects the Subject and the Object entities. The Subject and the Object entities are indicated with the subject and object tags respectively. You need to decide whether the relation holds between the Subject and the Object entities in the Query sentence.

Relation name: "#RELATION#" (#RELATION\_DESCRIPTION#)

Support Sentence 1: #SUPPORT\_SENTENCE\_1#

Support Sentence 2: #SUPPORT\_SENTENCE\_2#

...

Support Sentence N: #SUPPORT\_SENTENCE\_N#

Query Sentence: #QUERY\_SENTENCE#

If the relation holds between the Subject and Object entities in the Query sentence, say "yes," otherwise say "no." Just output "yes" or "no," and nothing else.

---

Figure 2: Prompt template used for relation extraction using the binary approach ( $N \in \{1, 5, 10\}$  in our experiments). Given (a) a relation name and description, (b) a set of support sentences that exemplify the relation, and (c) a query sentence, the model is asked to decide whether the relation holds between the subject and object entities in the query sentence.

---

You are given below a sentence, an entity contained within the sentence, and an entity type:

Sentence: #SENTENCE#

Entity: #ENTITY#

Entity Type: #ENTITY\_TYPE#

Your task is to decide whether the Entity in the context of the Sentence either:

1. belongs to the entity type "#ENTITY\_TYPE#"

or,

2. is a mention that points to an entity that belongs to the entity type "#ENTITY\_TYPE#" (such as a pronoun or other co-referring expression)

Only answer "yes" or "no," nothing else.

---

Figure 3: Prompt template used for validating compatibility between the (expected) named entity types of a relation type and the entities in a query sentence (Section 3). The entity types considered in this work are PERSON, LOCATION, ORGANIZATION, DATE, CITY, COUNTRY, STATE, and PROVINCE).

---

You are given below five Relation names, the Description of the relations between brackets, N Support sentences exemplifying each relation, and a Query sentence.

A relation connects the Subject and the Object entities. The Subject and the Object entities are indicated with the subject and object tags respectively. You need to decide whether the relation holds between the Subject and the Object entities in the Query sentence.

Relation name: "#RELATION1#" (#RELATION\_DESCRIPTION#)  
Support Sentence 1: #SUPPORT\_SENTENCE\_1#  
Support Sentence 2: #SUPPORT\_SENTENCE\_2#  
...  
Support Sentence N: #SUPPORT\_SENTENCE\_N#

Relation name: "#RELATION2#" (#RELATION\_DESCRIPTION#)  
Support Sentence 1: #SUPPORT\_SENTENCE\_1#  
Support Sentence 2: #SUPPORT\_SENTENCE\_2#  
...  
Support Sentence N: #SUPPORT\_SENTENCE\_N#

...

Relation name: "#RELATION5#" (#RELATION\_DESCRIPTION#)  
Support Sentence 1: #SUPPORT\_SENTENCE\_1#  
Support Sentence 2: #SUPPORT\_SENTENCE\_2#  
...  
Support Sentence N: #SUPPORT\_SENTENCE\_N#

If the relation holds between the Subject and Object entities in the Query sentence, say "yes," otherwise say "no." Just output "yes" or "no," and nothing else.

---

Figure 4: Prompt template used for relation extraction with the multi-class approach, which we evaluate only for the 1-shot baseline ( $N = 1$ ), with five candidate relations per episode. Given (a) five relation names and their descriptions, (b) a set of support sentences that exemplify each relation, and (c) a query sentence, the model is asked to decide which relation, if any, holds between the subject and object entities in the query.

---

You are given below a Relation name, a Description of the relation, and a support sentence exemplifying the relation.

A relation connects two entities: the Subject and the Object entities in the sentence. The Subject and the Object entities are indicated with the <subject>.</subject> and <object>.</object> tags respectively.

Relation name: "#RELATION#"   
Relation description: "#RELATION\_DESCRIPTION#"   
Support Sentence: #SUPPORT\_SENTENCE#

Your task is to generate N paraphrases of the support sentence that hold the same relation between the same Subject and Object entities. In each paraphrase, you must include the subject and object tags to identify the Subject and the Object entities.

Output in the following format:  
01: your 1st paraphrased sentence  
02: your 2nd paraphrased sentence  
.....  
N: your Nth paraphrased sentence

---

Figure 5: Prompt template used for obtaining additional examples by paraphrasing with an LLM the original (gold) support sentence. Given a relation name, its description, and a support sentence, the model is instructed to generate multiple paraphrases that preserve the relation.

Method	Additional Examples
Paraphrased with LLM	<ol style="list-style-type: none"> <li>1. [New Fabris]<sup>SBJ</sup> company director [Pierre Reau]<sup>OBJ</sup> stated that the company can currently only offer between 10,000 and 15,000 euros in redundancy to workers with 20 or more years of experience.</li> <li>2. [New Fabris]<sup>SBJ</sup> director [Pierre Reau]<sup>OBJ</sup> mentioned that the firm is currently only able to pay between 10,000 and 15,000 euros in redundancy to employees with 20 or more years of experience.</li> <li>3. [New Fabris]<sup>SBJ</sup> company director [Pierre Reau]<sup>OBJ</sup> explained that the company can only currently pay between 10,000 and 15,000 euros in redundancy to workers with 20 or more years of service.</li> <li>4. [New Fabris]<sup>SBJ</sup> director [Pierre Reau]<sup>OBJ</sup> said that the company is currently only able to provide between 10,000 and 15,000 euros in redundancy to employees with 20 or more years of experience.</li> </ol>
Generated with LLM	<ol style="list-style-type: none"> <li>1. [GlobalTech]<sup>SBJ</sup> CEO [Emma Lee]<sup>OBJ</sup> highlighted that the company's top employees are driving innovation and growth.</li> <li>2. [GreenEnergy]<sup>SBJ</sup> executive [James Carter]<sup>OBJ</sup> mentioned that the firm's top members are leading sustainable development initiatives.</li> <li>3. [HealthNova]<sup>SBJ</sup> director [Dr. Laura Martinez]<sup>OBJ</sup> stated that the organization's top employees are focused on patient care and research.</li> <li>4. [ArtisanWorks]<sup>SBJ</sup> founder [Robert Kim]<sup>OBJ</sup> noted that the company's top members are responsible for creative direction and product design.</li> </ol>
Retrieved SBERT	<ol style="list-style-type: none"> <li>1. According to [Microtune]<sup>SBJ</sup>, many of the laid off employees have been hired by [TFS]<sup>OBJ</sup>, which has agreed to address and build demand for radio frequency subsystems.</li> <li>2. [Derek Twigg]<sup>OBJ</sup>: In line with the Government's policy on controlling public sector pay, the Royal Fleet Auxiliary ( RFA ) pay submission has been subject to close scrutiny by the [Ministry of Defence and HM Treasury]<sup>SBJ</sup>.</li> <li>3. The new Board will work with the [FSA]<sup>SBJ</sup> to merge six existing compensation schemes into a single one stop shop, as proposed under the Financial Services and Markets [Bill]<sup>OBJ</sup>, currently before Parliament.</li> <li>4. "We are expecting the worst after August 14, when there will be retrenchment and the golden handshake will be offered to bank employees ahead of privatization," said [Habibuddin Junaidi]<sup>OBJ</sup>, president of [Habib Bank General Workers Front]<sup>SBJ</sup> Pakistan.</li> </ol>
Retrieved with semantic rule representations	<ol style="list-style-type: none"> <li>1. Among the representative works of the Bai people are Transit Star Catalogue for Time Determination by the [Ming Dynasty]<sup>SBJ</sup> scholar [Zhou Silian]<sup>OBJ</sup>, Collection of Secret Prescriptions by Chen Dongtian and Tested Prescriptions by Li Xingwei.</li> <li>2. [Sony Pictures]<sup>SBJ</sup> Classics co-presidents [Michael Barker]<sup>OBJ</sup> and Tom Bernard dropped in on a small dinner at local hotspot Prego, celebrating Alice Wu's "Saving Face" last week, and quickly closed a deal for the Discovery section film.</li> <li>3. "Options under consideration include regular and expanded sampling for hazardous substances such as anthrax, opening packages for inspection, and high-tech alternatives, including digitizing the mail," [House]<sup>SBJ</sup> Administration Chairman [Bob Ney]<sup>OBJ</sup> ( R-Ohio ) wrote in a recent letter to his colleagues.</li> <li>4. [Elon University]<sup>SBJ</sup> freshman [Elizabeth Easterly]<sup>OBJ</sup> presented research on body dissatisfaction, self-esteem and exercise motivators among college-age women.</li> </ol>
Retrieved with semantic rule representations and <i>k</i> -means clustering	<ol style="list-style-type: none"> <li>1. For further information, contact [Mike Smart]<sup>OBJ</sup>, the [Bureau]<sup>SBJ</sup>'s Senior Policy Advisor ( you can address him in German if you wish, as well as in English, French, and Spanish, and what not else ).</li> <li>2. "IPv6 is an emerging technology that will increase in significance as the demand for IP addresses grows," says [Lawrence Orans]<sup>OBJ</sup>, senior analyst, [Gartner, Inc.]<sup>SBJ</sup> "We anticipate that this demand will be strongest in the Asia-Pacific region, which owns far fewer registered IP addresses than does North America."</li> <li>3. After graduating from Strathclyde, I obtained a PhD at the University of Glasgow and continued as a researcher in the lab of Dr [Jim Brewer]<sup>OBJ</sup> ( himself a [Strathclyde Immunology]<sup>SBJ</sup> graduate ) looking at ways to improve the efficacy of vaccines.</li> <li>4. [Lieberman]<sup>OBJ</sup>, a three-term senator and the [Democratic]<sup>SBJ</sup> vice presidential candidate in 2000, lost by a 52 to 48 percent margin to Ned Lamont, an heir to the Lamont family fortune and multi-millionaire businessman, who ran as an opponent of the war in Iraq.</li> </ol>

Table 3: Examples of additional support examples obtained with different methods (Section 4) given the following original support sentence: *[New Fabris]<sup>SBJ</sup> company director [Pierre Reau]<sup>OBJ</sup> said that currently the firm would only able to pay between 10,000 and 15,000 euros in redundancy to workers with 20 or more years experience, and some junior workes would get just 3,000 euros* (relation: org:top\_members/employees). Additional examples obtained with the LLM (generated from scratch or by paraphrasing) tend to be near duplicates and semantically nearly identical to the original support example. On the other hand, retrieval-based methods provide additional examples with more diverse contexts. In particular, examples retrieved using semantic rule representations and *k*-means clustering show greater lexical and structural diversity while preserving the target relation.

---

You are given below a Relation name, the Description of the relation, and a support sentence exemplifying the relation.

A relation connects two entities: the Subject and the Object entities in the sentence. The Subject and the Object entities are indicated with the <subject>.</subject> and <object>.</object> tags respectively.

Relation name: "#RELATION#"
Relation description: "#RELATION\_DESCRIPTION#"
Support Sentence: #SUPPORT\_SENTENCE#

Your task is to generate N completely different new examples that hold the same relation. You must follow these guidelines:

1. In each example, include subject and object tags to identify the Subject and the Object entities.
2. To increase diversity, use different words, phrases, and sentence structures across different examples.

Output in the following format:

01: your 1st example sentence
02: your 2nd example sentence
...
N: your Nth example sentence

---

Figure 6: Prompt template used for obtaining additional examples by generating them with an LLM. Given a relation name, its description, and a support sentence, the model is instructed to generate diverse examples that hold the relation.

---

You are given below a Relation name, a Description of the relation between brackets, and N Support sentences exemplifying the relation.

A relation connects the Subject and the Object entities. The Subject and the Object entities are marked within the subject and object tags respectively.

Relation name: "#RELATION#" (#RELATION\_DESCRIPTION#)
Support Sentence 1: #SUPPORT\_SENTENCE\_1#
Support Sentence 2: #SUPPORT\_SENTENCE\_2#
...
Support Sentence N: #SUPPORT\_SENTENCE\_N#

Your task is to pick N/2 support sentences that maximize diversity. In other words, you should pick support sentences that use different words, phrases, and sentence structures.

Output your best picks as a Python-style list of the N/2 IDs of the support sentences (e.g., [1, 4, 6, 7]).

Only output the list and nothing else.

---

Figure 7: Prompt template used in our hybrid approach to select additional examples from the ones (a) generated with an LLM and (b) retrieved with our semantic rule representations and *k*-means clustering. In our experiments, (a) *N* = 8 or *N* = 18 and (b) we shuffle the *N* support sentences in the prompt to avoid positional biases.

---

You are given a context sentence containing a Subject and an Object entity.

The Subject and Object entities are marked using <subject> and <object> tags, respectively.

Your task is to summarize the relation expressed between the Subject and the Object entities in the context.

Context: #SUPPORT\_SENTENCE#

You must retain the <subject> and <object> tags in the summarized output.

Only output the summarized relation between the Subject and the Object entities, and nothing else.

---

Figure 8: Prompt template used for summarizing support sentences. Given a support sentence exemplifying a relation, the model is instructed to produce a compact natural language description of the relation between the subject and object entities, while retaining the subject and object tags.

Dataset	#ex.	Qwen3-4B			Gemma3-4B		
		G+A	P	R	F1	P	R
FS-TACRED	1+0	14.6 ± 1.05	8.4 ± 0.90	10.6 ± 1.00	4.4 ± 0.32	4.3 ± 0.29	4.4 ± 0.30
FS-FewRel	1+0	52.0 ± 2.31	16.5 ± 1.36	25.0 ± 1.79	49.3 ± 7.91	1.8 ± 0.18	3.4 ± 0.34
FS-NYT29	1+0	41.9 ± 0.31	64.1 ± 0.69	50.7 ± 0.41	56.3 ± 1.37	21.1 ± 0.82	30.7 ± 1.07

Table 4: Results obtained with a 1-shot prompt only providing the subject and object entities and nothing else from the query sentence (Figure 9). This experiment was designed to test the parametric knowledge of the LLMs with respect to the test instances in FS-TACRED, FS-FewRel, and FS-NYT29. The results are relatively low with FS-TACRED and FS-FewRel but surprisingly high with FS-NYT29. We conclude that the query sentences in FS-NYT29 contain many instances for which parametric knowledge—even in relatively small LLMs—is enough to make the correct inference. Thus, we do not consider FS-NYT29 in our main experiments.

---

You are given below a Relation name, a Description of the relation in brackets, a Support sentence exemplifying the relation between the Subject and the Object entities, and a Query (a subject and an object entity).

A relation connects the Subject and the Object entities. The Subject and the Object entities are given within the subject and object tags respectively. You need to decide whether the relation holds between the Subject and the Object entities in the given Query.

Relation name: “#RELATION#” (#RELATION\_DESCRIPTION#)

Support Sentence 1: #SUPPORT\_SENTENCE\_1#

Support Sentence 2: #SUPPORT\_SENTENCE\_2#

...

Support Sentence N: #SUPPORT\_SENTENCE\_N#

Query Subject: #SUBJECT#

Query Object: #OBJECT#

If the relation between the Subject and the Object entities in the Query matches the given Relation say “yes,” otherwise “no.”

---

Figure 9: Prompt template for 1-shot relation extraction only providing the subject and object entities and nothing else from the query sentence. This prompt was designed to test the parametric knowledge of the LLMs with respect to the test instances in FS-TACRED, FS-FewRel, and FS-NYT29.

	#ex.	Qwen3-4B			Qwen3-14B		
		G+A	P	R	F1	P	R
Few-shot baseline	1+0	41.7 ± 1.37	26.4 ± 1.17	32.4 <sup>†</sup> ± 1.14	43.5 ± 0.81	42.3 ± 0.73	42.9 <sup>†</sup> ± 0.52
Including Additional examples							
Paraphrased with LLM	1+4	44.3 ± 1.91	13.4 ± 0.78	20.5 <sup>†</sup> ± 1.05	42.6 ± 1.17	48.8 ± 0.75	<b>45.5</b> ± 0.75
	1+9	40.2 ± 2.37	13.3 ± 0.97	20.0 <sup>†</sup> ± 1.36	45.6 ± 0.72	43.9 ± 0.51	44.7 <sup>†</sup> ± 0.09
Generated with LLM	1+4	43.8 ± 0.95	29.4 ± 0.75	<b>35.2</b> <sup>†</sup> ± 0.75	42.7 ± 1.20	48.8 ± 0.76	<b>45.5</b> ± 0.78
	1+9	42.0 ± 0.96	28.4 ± 0.68	33.9 <sup>†</sup> ± 0.79	45.8 ± 0.92	44.2 ± 0.37	45.0* ± 0.31
Retrieved (closest) using SBERT representations	1+4	43.3 ± 0.94	22.4 ± 0.48	29.5 <sup>†</sup> ± 0.57	43.7 ± 0.95	43.0 ± 1.86	43.4 <sup>†</sup> ± 1.13
	1+9	42.4 ± 1.10	21.2 ± 0.45	28.2 <sup>†</sup> ± 0.47	44.1 ± 1.04	39.8 ± 1.52	41.8 <sup>†</sup> ± 1.08
+ hybrid	1+4	44.1 ± 1.19	29.4 ± 0.68	<b>35.3</b> <sup>†</sup> ± 0.86	44.0 ± 0.93	43.1 ± 1.77	<b>43.5</b> <sup>†</sup> ± 1.13
	1+9	43.0 ± 0.73	29.2 ± 0.51	34.7 <sup>†</sup> ± 0.50	44.7 ± 1.18	39.9 ± 1.53	42.1 <sup>†</sup> ± 1.00
Semantic rule representations	1+4	40.6 ± 1.44	33.0 ± 1.30	36.4 ± 1.37	43.3 ± 0.66	45.6 ± 1.63	<b>44.4</b> <sup>†</sup> ± 0.98
	1+9	38.1 ± 1.12	33.0 ± 0.57	35.3 <sup>†</sup> ± 0.78	42.7 ± 0.66	42.5 ± 1.63	42.6 <sup>†</sup> ± 0.98
+ hybrid	1+4	44.1 ± 1.50	31.6 ± 1.20	<b>36.8</b> ± 1.25	43.2 ± 0.61	45.4 ± 1.49	44.2 <sup>†</sup> ± 0.86
	1+9	41.9 ± 1.36	31.6 ± 0.90	36.0 ± 1.06	42.8 ± 1.22	43.0 ± 1.72	42.9 <sup>†</sup> ± 1.19
and <i>k</i> -means clustering	1+4	39.1 ± 1.66	33.0 ± 0.98	35.8* ± 1.22	40.3 ± 0.97	47.1 ± 1.71	43.4 <sup>†</sup> ± 1.03
	1+9	35.6 ± 1.23	32.4 ± 0.58	33.9 <sup>†</sup> ± 0.85	39.9 ± 0.99	46.3 ± 1.27	42.9 <sup>†</sup> ± 0.88
+ hybrid	1+4	43.9 ± 1.49	31.6 ± 0.99	36.7 ± 1.09	40.3 ± 1.14	46.7 ± 1.86	43.2 <sup>†</sup> ± 1.21
	1+9	43.0 ± 1.46	32.2 ± 1.29	<b>36.8</b> ± 1.35	39.9 ± 0.82	45.9 ± 0.81	42.7 <sup>†</sup> ± 0.64

Table 5: Results obtained with FS-FewRel using in-context learning in the 1-shot setting (i.e., one Gold example). Our hybrid approach to obtain additional examples (i.e., coupling LLM-generated examples with our semantic rule-based retrieval) yields the best results with Qwen3-4B (F1: +1.6 compared to generating additional examples with an LLM), but we only observe benefits on Recall with Qwen3-14B. \* and <sup>†</sup> indicate the best method in the whole column is statistically significantly better at  $p < 0.05$  and  $p < 0.01$ , respectively, based on one-sided paired bootstrap test with replacement on F1 scores.

	#ex.	Qwen3-4B			Qwen3-14B		
		G+A	P	R	F1	P	R
Few-shot baseline	1+0	28.7 ± 0.99	17.8 ± 1.21	22.0 <sup>†</sup> ± 1.19	31.6 ± 0.96	40.3 ± 1.47	35.4 <sup>†</sup> ± 1.02
Including Additional examples							
Paraphrased with LLM	1+4	30.8 ± 1.53	08.6 ± 0.96	13.5 <sup>†</sup> ± 1.28	42.5 ± 1.17	24.7 ± 0.73	31.2 <sup>†</sup> ± 0.68
	1+9	31.4 ± 2.17	09.1 ± 0.83	14.1 <sup>†</sup> ± 1.16	41.0 ± 1.27	23.0 ± 0.66	29.5 <sup>†</sup> ± 0.82
Generated with LLM	1+4	26.8 ± 0.50	24.2 ± 1.34	25.4 <sup>†</sup> ± 0.92	37.3 ± 1.02	40.2 ± 1.73	<b>38.7*</b> ± 1.11
	1+9	27.8 ± 1.08	24.7 ± 1.14	<b>26.1<sup>†</sup></b> ± 0.84	38.6 ± 1.42	36.4 ± 1.55	37.5 <sup>†</sup> ± 1.10
Retrieved (closest) using							
SBERT representations	1+4	24.5 ± 1.47	15.4 ± 1.67	18.9 <sup>†</sup> ± 1.56	31.1 ± 1.14	36.0 ± 1.41	33.4 <sup>†</sup> ± 1.05
	1+9	24.9 ± 1.10	15.9 ± 0.88	19.4 <sup>†</sup> ± 0.83	30.4 ± 1.09	31.1 ± 0.87	30.7 <sup>†</sup> ± 0.76
+ hybrid	1+4	25.9 ± 0.58	23.4 ± 1.31	24.6 <sup>†</sup> ± 0.81	39.3 ± 0.70	39.4 ± 1.48	<b>39.3</b> ± 0.70
	1+9	26.6 ± 0.91	25.4 ± 0.79	<b>26.0<sup>†</sup></b> ± 0.57	39.3 ± 0.80	35.7 ± 1.40	37.4 <sup>†</sup> ± 0.86
Semantic rule representations	1+4	23.9 ± 0.48	23.3 ± 0.56	23.6 <sup>†</sup> ± 0.43	33.3 ± 1.15	38.6 ± 1.49	35.8 <sup>†</sup> ± 1.21
	1+9	22.9 ± 0.56	25.1 ± 0.68	23.9 <sup>†</sup> ± 0.50	33.0 ± 1.66	33.9 ± 2.08	33.4 <sup>†</sup> ± 1.76
+ hybrid	1+4	26.6 ± 0.87	26.7 ± 0.61	26.6 ± 0.54	39.4 ± 1.04	39.4 ± 1.27	<b>39.4</b> ± 0.89
	1+9	26.2 ± 0.60	28.9 ± 0.91	<b>27.5</b> ± 0.48	40.0 ± 1.11	36.2 ± 1.12	38.0 <sup>†</sup> ± 0.82
and <i>k</i> -means clustering							
w/ random cluster selection	1+4	22.8 ± 0.89	28.5 ± 1.21	25.3 <sup>†</sup> ± 0.96	29.5 ± 0.81	46.3 ± 1.18	36.0 <sup>†</sup> ± 0.83
	1+9	22.1 ± 0.75	30.6 ± 0.98	25.7 <sup>†</sup> ± 0.70	29.7 ± 0.89	45.1 ± 1.59	35.8 <sup>†</sup> ± 0.96
+ hybrid	1+4	25.7 ± 1.05	26.6 ± 0.68	26.1 <sup>†</sup> ± 0.70	37.9 ± 0.94	40.5 ± 1.64	39.1 ± 1.15
	1+9	26.4 ± 0.69	28.9 ± 1.88	<b>27.6</b> ± 1.20	38.7 ± 1.19	37.0 ± 1.36	37.8 <sup>†</sup> ± 0.84
w/ closest cluster selection	1+4	22.1 ± 0.66	27.7 ± 1.32	24.6 <sup>†</sup> ± 0.90	30.1 ± 0.79	43.2 ± 1.82	35.5 <sup>†</sup> ± 1.05
	1+9	21.2 ± 0.39	28.9 ± 1.37	24.5 <sup>†</sup> ± 0.69	31.0 ± 0.96	41.5 ± 2.06	35.5 <sup>†</sup> ± 1.04
+ hybrid	1+4	25.3 ± 0.59	26.2 ± 0.82	25.8 <sup>†</sup> ± 0.49	38.2 ± 0.81	40.5 ± 1.44	<b>39.3</b> ± 0.75
	1+9	25.5 ± 0.62	27.8 ± 1.30	26.6* ± 0.83	39.7 ± 1.22	37.5 ± 1.49	38.6* ± 0.97
and <i>k</i> -means++ clustering							
w/ furthest cluster selection	1+4	22.7 ± 0.93	28.1 ± 1.72	25.1 <sup>†</sup> ± 1.24	28.4 ± 0.80	46.4 ± 1.58	35.2 <sup>†</sup> ± 0.98
	1+9	21.6 ± 0.54	31.6 ± 1.11	25.7 <sup>†</sup> ± 0.73	29.1 ± 0.92	46.0 ± 1.78	35.7 <sup>†</sup> ± 1.09
+ hybrid	1+4	25.0 ± 0.65	26.1 ± 1.55	25.5 <sup>†</sup> ± 0.97	37.6 ± 0.84	40.2 ± 1.57	38.8 ± 0.89
	1+9	25.8 ± 0.38	28.9 ± 1.39	<b>27.3</b> ± 0.79	39.1 ± 1.31	37.7 ± 1.67	38.4 <sup>†</sup> ± 1.19
w/ closest cluster selection	1+4	22.3 ± 0.50	26.3 ± 1.05	24.1 <sup>†</sup> ± 0.70	31.3 ± 1.12	42.7 ± 1.75	36.1 <sup>†</sup> ± 1.11
	1+9	21.7 ± 0.52	29.6 ± 0.72	25.0 <sup>†</sup> ± 0.36	31.4 ± 1.20	41.8 ± 2.37	35.8 <sup>†</sup> ± 1.46
+ hybrid	1+4	25.4 ± 0.82	26.3 ± 1.42	25.8 <sup>†</sup> ± 1.00	38.2 ± 1.17	40.2 ± 1.45	<b>39.2</b> ± 1.04
	1+9	26.1 ± 0.46	27.9 ± 1.54	27.0 ± 0.95	39.8 ± 1.39	37.2 ± 1.71	38.5* ± 1.20
Retrieved (closest) using							
Semantic rule representations							
and <i>k</i> -means, <i>summarization</i>							
w/ random cluster selection	1+4	21.1 ± 0.77	33.9 ± 1.00	26.0 <sup>†</sup> ± 0.84	27.8 ± 0.60	44.3 ± 1.63	34.1 <sup>†</sup> ± 0.90
	1+9	20.8 ± 0.84	41.1 ± 1.25	<b>27.6</b> ± 1.01	28.4 ± 1.13	41.0 ± 1.96	33.6 <sup>†</sup> ± 1.36
+ hybrid	1+4	22.8 ± 0.29	29.1 ± 0.91	25.5 <sup>†</sup> ± 0.38	32.8 ± 0.96	41.3 ± 1.46	<b>36.6<sup>†</sup></b> ± 1.08
	1+9	23.6 ± 0.69	33.2 ± 2.10	<b>27.6</b> ± 1.18	35.1 ± 1.45	35.6 ± 1.54	35.4 <sup>†</sup> ± 1.32
w/ closest cluster selection	1+4	21.3 ± 0.74	32.6 ± 1.58	25.8 <sup>†</sup> ± 0.96	28.6 ± 0.86	40.7 ± 1.53	33.6 <sup>†</sup> ± 0.93
	1+9	20.2 ± 0.71	41.6 ± 1.61	27.2 ± 0.91	29.8 ± 1.02	39.0 ± 1.36	33.8 <sup>†</sup> ± 1.00
+ hybrid	1+4	23.5 ± 0.56	29.3 ± 1.45	26.1 <sup>†</sup> ± 0.85	32.7 ± 1.70	40.7 ± 1.33	36.2 <sup>†</sup> ± 1.47
	1+9	22.7 ± 0.40	33.1 ± 1.69	27.0 ± 0.77	34.8 ± 1.22	35.6 ± 1.55	35.2 <sup>†</sup> ± 1.06
and <i>k</i> -means++, <i>summarization</i>							
w/ furthest cluster selection	1+4	22.1 ± 0.57	35.2 ± 1.25	27.1 ± 0.74	28.0 ± 0.90	44.4 ± 2.00	34.4 <sup>†</sup> ± 1.20
	1+9	20.0 ± 0.68	42.4 ± 1.38	<b>27.2</b> ± 0.88	29.4 ± 1.46	42.0 ± 1.85	34.6 <sup>†</sup> ± 1.51
+ hybrid	1+4	23.2 ± 0.84	29.2 ± 1.31	25.9 <sup>†</sup> ± 0.90	33.0 ± 1.39	41.7 ± 1.27	<b>36.8<sup>†</sup></b> ± 1.16
	1+9	23.2 ± 0.38	32.9 ± 1.95	<b>27.2</b> ± 0.89	35.4 ± 1.40	36.2 ± 1.23	35.8 <sup>†</sup> ± 1.01
w/ closest cluster selection	1+4	21.0 ± 1.27	32.1 ± 1.95	25.4 <sup>†</sup> ± 1.51	29.5 ± 1.01	41.1 ± 1.58	34.3 <sup>†</sup> ± 0.99
	1+9	19.4 ± 0.68	40.3 ± 1.53	26.2 <sup>†</sup> ± 0.91	29.3 ± 0.76	37.8 ± 1.48	33.0 <sup>†</sup> ± 0.80
+ hybrid	1+4	22.4 ± 0.69	28.9 ± 1.96	25.2 <sup>†</sup> ± 1.12	32.5 ± 0.98	40.9 ± 1.60	36.2 <sup>†</sup> ± 1.05
	1+9	22.5 ± 0.47	32.2 ± 1.78	26.5* ± 0.81	35.6 ± 1.13	35.4 ± 1.41	35.5 <sup>†</sup> ± 1.07

Table 6: Full results obtained with FS-TACRED using Qwen3-4B and Qwen3-14B. This table complements Table 1 by providing results with alternative cluster selection strategies and summarizing examples. We observe the same trends as in Table 1: our hybrid approach combining additional examples generated with an LLM and retrieved with semantic rule representations outperforms the alternatives. \* and † indicate the best method in the whole column is statistically significantly better at  $p < 0.05$  and  $p < 0.01$ , respectively, based on one-sided paired bootstrap test with replacement on F1 scores.

	#ex.	Qwen3-4B			Qwen3-14B		
		G+A	P	R	F1	P	R
Few-shot baseline	1+0	41.7 ± 1.37	26.4 ± 1.17	32.4 <sup>†</sup> ± 1.14	43.5 ± 0.81	42.3 ± 0.73	42.9 <sup>†</sup> ± 0.52
Including Additional examples							
Paraphrased with LLM	1+4	44.3 ± 1.91	13.4 ± 0.78	20.5 <sup>†</sup> ± 1.05	42.6 ± 1.17	48.8 ± 0.75	<b>45.5</b> ± 0.75
	1+9	40.2 ± 2.37	13.3 ± 0.97	20.0 <sup>†</sup> ± 1.36	45.6 ± 0.72	43.9 ± 0.51	44.7 <sup>†</sup> ± 0.09
Generated with LLM	1+4	43.8 ± 0.95	29.4 ± 0.75	<b>35.2</b> <sup>†</sup> ± 0.75	42.7 ± 1.20	48.8 ± 0.76	<b>45.5</b> ± 0.78
	1+9	42.0 ± 0.96	28.4 ± 0.68	33.9 <sup>†</sup> ± 0.79	45.8 ± 0.92	44.2 ± 0.37	45.0 ± 0.31
Retrieved (closest) using SBERT representations							
	1+4	43.3 ± 0.94	22.4 ± 0.48	29.5 <sup>†</sup> ± 0.57	43.7 ± 0.95	43.0 ± 1.86	43.4 <sup>†</sup> ± 1.13
	1+9	42.4 ± 1.10	21.2 ± 0.45	28.2 <sup>†</sup> ± 0.47	44.1 ± 1.04	39.8 ± 1.52	41.8 <sup>†</sup> ± 1.08
+ hybrid	1+4	44.1 ± 1.19	29.4 ± 0.68	<b>35.3</b> <sup>†</sup> ± 0.86	44.0 ± 0.93	43.1 ± 1.77	<b>43.5</b> <sup>†</sup> ± 1.13
	1+9	43.0 ± 0.73	29.2 ± 0.51	34.7 <sup>†</sup> ± 0.50	44.7 ± 1.18	39.9 ± 1.53	42.1 <sup>†</sup> ± 1.00
Semantic rule representations							
	1+4	40.6 ± 1.44	33.0 ± 1.30	36.4 <sup>†</sup> ± 1.37	43.3 ± 0.66	45.6 ± 1.63	<b>44.4</b> <sup>†</sup> ± 0.98
	1+9	38.1 ± 1.12	33.0 ± 0.57	35.3 <sup>†</sup> ± 0.78	42.7 ± 0.66	42.5 ± 1.63	42.6 <sup>†</sup> ± 0.98
+ hybrid	1+4	44.1 ± 1.50	31.6 ± 1.20	<b>36.8</b> <sup>†</sup> ± 1.25	43.2 ± 0.61	45.4 ± 1.49	44.2 <sup>†</sup> ± 0.86
	1+9	41.9 ± 1.36	31.6 ± 0.90	36.0 <sup>†</sup> ± 1.06	42.8 ± 1.22	43.0 ± 1.72	42.9 <sup>†</sup> ± 1.19
and <i>k</i> -means clustering							
w/ random cluster selection	1+4	39.1 ± 1.66	33.0 ± 0.98	35.8 <sup>†</sup> ± 1.22	40.3 ± 0.97	47.1 ± 1.71	43.4 <sup>†</sup> ± 1.03
	1+9	35.6 ± 1.23	32.4 ± 0.58	33.9 <sup>†</sup> ± 0.85	39.9 ± 0.99	46.3 ± 1.27	42.9 <sup>†</sup> ± 0.88
+ hybrid	1+4	43.9 ± 1.49	31.6 ± 0.99	36.7 <sup>†</sup> ± 1.09	40.3 ± 1.14	46.7 ± 1.86	43.2 <sup>†</sup> ± 1.21
	1+9	43.0 ± 1.46	32.2 ± 1.29	<b>36.8</b> <sup>†</sup> ± 1.35	39.9 ± 0.82	45.9 ± 0.81	42.7 <sup>†</sup> ± 0.64
w/ closest cluster selection	1+4	38.1 ± 1.09	34.1 ± 0.61	36.0 <sup>†</sup> ± 0.81	41.1 ± 0.94	46.5 ± 1.71	43.6 <sup>†</sup> ± 1.12
	1+9	36.1 ± 1.18	33.7 ± 0.71	34.9 <sup>†</sup> ± 0.91	40.4 ± 0.89	45.4 ± 1.68	42.7 <sup>†</sup> ± 1.15
+ hybrid	1+4	43.7 ± 1.28	31.8 ± 0.72	<b>36.8</b> <sup>†</sup> ± 0.84	41.3 ± 1.05	46.4 ± 1.59	<b>43.7</b> <sup>†</sup> ± 1.13
	1+9	42.8 ± 1.06	32.1 ± 0.81	36.6 <sup>†</sup> ± 0.81	40.1 ± 1.29	44.8 ± 1.62	42.3 <sup>†</sup> ± 1.29
and <i>k</i> -means++ clustering							
w/ furthest cluster selection	1+4	40.0 ± 1.20	33.7 ± 1.32	36.6 <sup>†</sup> ± 1.16	39.6 ± 0.85	47.9 ± 1.72	43.3 <sup>†</sup> ± 0.99
	1+9	37.2 ± 0.98	33.5 ± 0.64	35.3 <sup>†</sup> ± 0.74	39.6 ± 0.68	46.1 ± 1.20	42.6 <sup>†</sup> ± 0.63
+ hybrid	1+4	43.4 ± 1.49	31.1 ± 0.68	36.2 <sup>†</sup> ± 0.91	39.8 ± 0.49	47.7 ± 1.48	43.4 <sup>†</sup> ± 0.68
	1+9	42.5 ± 1.35	31.3 ± 0.94	36.1 <sup>†</sup> ± 1.00	39.2 ± 0.86	45.9 ± 1.04	42.3 <sup>†</sup> ± 0.80
w/ closest cluster selection	1+4	39.1 ± 1.16	34.0 ± 0.89	36.4 <sup>†</sup> ± 0.96	40.9 ± 1.15	46.3 ± 1.36	43.5 <sup>†</sup> ± 1.06
	1+9	36.2 ± 1.20	33.1 ± 0.77	34.6 <sup>†</sup> ± 0.94	40.1 ± 1.15	44.6 ± 1.33	42.2 <sup>†</sup> ± 1.06
+ hybrid	1+4	43.5 ± 1.16	31.9 ± 0.65	<b>36.8</b> <sup>†</sup> ± 0.68	41.1 ± 1.12	46.6 ± 1.27	<b>43.7</b> <sup>†</sup> ± 0.97
	1+9	42.7 ± 0.88	31.7 ± 0.57	36.3 <sup>†</sup> ± 0.54	40.2 ± 1.15	44.7 ± 1.04	42.4 <sup>†</sup> ± 0.95
Retrieved (closest) using Semantic rule representations and <i>k</i> -means, <i>summarization</i>							
w/ random cluster selection	1+4	36.5 ± 0.97	36.0 ± 0.78	36.2 <sup>†</sup> ± 0.82	39.5 ± 0.89	44.7 ± 1.24	42.0 <sup>†</sup> ± 0.85
	1+9	33.1 ± 0.56	39.5 ± 0.58	36.0 <sup>†</sup> ± 0.46	38.4 ± 1.34	42.5 ± 0.85	40.3 <sup>†</sup> ± 1.05
+ hybrid	1+4	41.2 ± 0.72	35.8 ± 0.68	38.3 ± 0.64	39.9 ± 0.76	45.7 ± 1.23	<b>42.6</b> <sup>†</sup> ± 0.66
	1+9	38.8 ± 0.94	36.8 ± 0.77	37.7 <sup>*</sup> ± 0.78	39.0 ± 1.36	43.5 ± 0.91	41.1 <sup>†</sup> ± 1.07
w/ closest cluster selection	1+4	36.1 ± 0.84	36.3 ± 0.90	36.2 <sup>†</sup> ± 0.75	39.7 ± 0.90	44.4 ± 1.07	41.9 <sup>†</sup> ± 0.80
	1+9	33.1 ± 1.24	39.9 ± 1.06	36.2 <sup>†</sup> ± 1.12	39.4 ± 0.60	42.4 ± 0.74	40.8 <sup>†</sup> ± 0.42
+ hybrid	1+4	41.2 ± 0.77	36.0 ± 0.74	<b>38.4</b> ± 0.66	40.1 ± 0.83	44.6 ± 1.33	42.2 <sup>†</sup> ± 0.83
	1+9	39.2 ± 1.05	37.3 ± 1.17	38.2 ± 1.03	39.6 ± 0.95	42.8 ± 1.66	41.1 <sup>†</sup> ± 1.14
and <i>k</i> -means++, <i>summarization</i>							
w/ furthest cluster selection	1+4	37.8 ± 1.19	38.0 ± 1.23	37.9 ± 1.09	39.3 ± 0.74	45.8 ± 1.03	42.3 <sup>†</sup> ± 0.59
	1+9	33.2 ± 1.31	38.4 ± 1.20	35.6 <sup>†</sup> ± 1.18	38.5 ± 1.12	43.2 ± 0.79	40.7 <sup>†</sup> ± 0.91
+ hybrid	1+4	41.1 ± 1.40	35.8 ± 1.76	<b>38.3</b> ± 1.50	39.0 ± 1.05	45.4 ± 0.39	41.9 <sup>†</sup> ± 0.69
	1+9	39.3 ± 1.10	36.7 ± 0.95	37.9 ± 0.91	38.5 ± 1.06	43.3 ± 1.08	40.7 <sup>†</sup> ± 0.96
w/ closest cluster selection	1+4	37.3 ± 1.16	37.9 ± 1.19	37.6 <sup>†</sup> ± 1.03	40.6 ± 1.29	44.5 ± 1.11	<b>42.4</b> <sup>†</sup> ± 1.05
	1+9	33.0 ± 1.02	39.8 ± 0.93	36.1 <sup>†</sup> ± 0.86	39.6 ± 1.14	42.6 ± 0.82	41.0 <sup>†</sup> ± 0.89
+ hybrid	1+4	40.9 ± 0.92	35.9 ± 0.77	38.2 ± 0.72	40.3 ± 1.18	44.6 ± 1.28	42.3 <sup>†</sup> ± 1.09
	1+9	39.1 ± 1.11	37.1 ± 0.83	38.1 ± 0.83	39.3 ± 1.09	42.5 ± 0.65	40.8 <sup>†</sup> ± 0.78

Table 7: Full results obtained with FS-FewRel using Qwen3-4B and Qwen3-14B. This table complements Table 5 by providing results with alternative cluster selection strategies and summarizing examples. We observe the same trends as in Table 5: our hybrid approach combining additional examples generated with an LLM and retrieved with semantic rule representations outperforms the alternatives with Qwen3-4B. \* and † indicate the best method in the whole column is statistically significantly better at  $p < 0.05$  and  $p < 0.01$ , respectively, based on one-sided paired bootstrap test with replacement on F1 scores.

	#ex.	Gemma3-4B			Gemma3-12B		
		G+A	P	R	F1	P	R
Few-shot baseline	1+0	11.9 ± 0.60	18.5 ± 0.54	14.5 ± 0.58	25.3 ± 0.41	51.3 ± 1.69	33.9 ± 0.62
Including Additional examples							
Paraphrased with LLM	1+4	08.9 ± 0.49	15.7 ± 1.02	11.3 <sup>†</sup> ± 0.64	28.6 ± 1.71	33.1 ± 1.81	30.7 <sup>†</sup> ± 1.68
	1+9	07.9 ± 0.59	14.0 ± 1.39	10.1 <sup>†</sup> ± 0.83	29.0 ± 1.58	30.0 ± 2.06	29.5 <sup>†</sup> ± 1.77
Generated with LLM	1+4	08.9 ± 0.30	38.1 ± 1.43	<b>14.5<sup>†</sup></b> ± 0.48	21.4 ± 0.35	62.6 ± 2.13	<b>31.9<sup>†</sup></b> ± 0.59
	1+9	09.4 ± 0.41	31.5 ± 1.86	<b>14.5<sup>†</sup></b> ± 0.65	20.8 ± 0.22	64.1 ± 1.78	31.4 <sup>†</sup> ± 0.43
Retrieved (closest) using							
SBERT representations	1+4	10.0 ± 0.53	27.8 ± 1.28	14.7 ± 0.72	23.5 ± 0.69	49.1 ± 1.89	<b>31.8<sup>†</sup></b> ± 0.96
	1+9	10.1 ± 0.49	24.3 ± 0.97	14.3 <sup>†</sup> ± 0.63	23.3 ± 0.69	43.1 ± 1.23	30.2 <sup>†</sup> ± 0.85
+ hybrid	1+4	09.4 ± 0.33	38.0 ± 1.47	<b>15.1</b> ± 0.53	21.3 ± 0.37	61.9 ± 1.70	31.7 <sup>†</sup> ± 0.48
	1+9	09.7 ± 0.34	32.1 ± 1.51	14.9 ± 0.55	21.2 ± 0.33	62.0 ± 1.78	31.6 <sup>†</sup> ± 0.49
Semantic rule representations	1+4	08.9 ± 0.20	39.0 ± 1.15	14.5* ± 0.31	20.2 ± 0.44	60.7 ± 2.42	30.3 <sup>†</sup> ± 0.79
	1+9	08.6 ± 0.21	37.6 ± 1.53	14.0 <sup>†</sup> ± 0.38	19.6 ± 0.27	57.7 ± 1.59	29.3 <sup>†</sup> ± 0.48
+ hybrid	1+4	08.9 ± 0.39	40.6 ± 1.45	14.5* ± 0.60	21.2 ± 0.31	63.4 ± 2.26	<b>31.8<sup>†</sup></b> ± 0.58
	1+9	09.3 ± 0.15	36.8 ± 1.49	<b>14.8</b> ± 0.30	21.0 ± 0.49	64.4 ± 2.61	31.6 <sup>†</sup> ± 0.81
and <i>k</i> -means clustering							
w/ random cluster selection	1+4	08.8 ± 0.35	42.8 ± 1.50	14.5* ± 0.56	19.1 ± 0.30	62.2 ± 2.01	29.2 <sup>†</sup> ± 0.55
	1+9	08.8 ± 0.24	41.2 ± 1.42	14.5* ± 0.40	18.8 ± 0.37	61.4 ± 1.47	28.8 <sup>†</sup> ± 0.57
+ hybrid	1+4	08.8 ± 0.43	41.9 ± 1.90	14.6* ± 0.69	20.7 ± 0.30	62.7 ± 1.82	31.1 <sup>†</sup> ± 0.49
	1+9	09.1 ± 0.45	36.2 ± 1.69	14.6* ± 0.69	21.0 ± 0.27	65.2 ± 2.11	<b>31.7<sup>†</sup></b> ± 0.50
w/ closest cluster selection	1+4	08.5 ± 0.34	42.5 ± 1.35	14.2 <sup>†</sup> ± 0.54	20.1 ± 0.34	61.5 ± 1.81	30.3 <sup>†</sup> ± 0.56
	1+9	08.8 ± 0.20	42.0 ± 1.23	14.6* ± 0.32	19.2 ± 0.26	60.0 ± 1.08	29.1 <sup>†</sup> ± 0.38
+ hybrid	1+4	08.8 ± 0.15	41.0 ± 0.78	14.4* ± 0.21	20.9 ± 0.30	63.6 ± 1.63	31.4 <sup>†</sup> ± 0.37
	1+9	09.5 ± 0.34	37.3 ± 1.39	<b>15.1</b> ± 0.53	20.7 ± 0.29	64.6 ± 1.98	31.4 <sup>†</sup> ± 0.45
and <i>k</i> -means++ clustering							
w/ furthest cluster selection	1+4	08.8 ± 0.40	42.7 ± 1.26	14.6* ± 0.61	19.7 ± 0.30	62.5 ± 1.94	30.0 <sup>†</sup> ± 0.55
	1+9	08.6 ± 0.14	39.7 ± 1.50	14.1 <sup>†</sup> ± 0.26	18.8 ± 0.34	61.2 ± 0.90	28.7 <sup>†</sup> ± 0.42
+ hybrid	1+4	08.9 ± 0.48	41.7 ± 1.72	<b>14.7</b> ± 0.74	21.0 ± 0.23	63.0 ± 1.65	31.5 <sup>†</sup> ± 0.32
	1+9	09.1 ± 0.34	36.8 ± 1.92	14.6* ± 0.56	21.0 ± 0.38	64.6 ± 2.05	<b>31.7<sup>†</sup></b> ± 0.53
w/ closest cluster selection	1+4	08.7 ± 0.24	41.5 ± 1.63	14.4 <sup>†</sup> ± 0.41	19.9 ± 0.21	61.0 ± 1.79	30.0 <sup>†</sup> ± 0.41
	1+9	08.6 ± 0.35	40.0 ± 1.43	14.1 <sup>†</sup> ± 0.56	19.8 ± 0.23	60.5 ± 1.32	29.9 <sup>†</sup> ± 0.37
+ hybrid	1+4	08.9 ± 0.31	42.0 ± 1.68	<b>14.7* ± 0.50</b>	20.9 ± 0.29	63.0 ± 2.06	31.3 <sup>†</sup> ± 0.48
	1+9	09.0 ± 0.27	37.1 ± 1.30	14.5 <sup>†</sup> ± 0.43	20.8 ± 0.26	64.4 ± 1.70	31.4 <sup>†</sup> ± 0.35
Retrieved (closest) using							
Semantic rule representations							
and <i>k</i> -means, <i>summarization</i>							
w/ random cluster selection	1+4	08.0 ± 0.41	39.4 ± 2.00	13.3 <sup>†</sup> ± 0.67	18.7 ± 0.39	64.1 ± 1.25	28.9 <sup>†</sup> ± 0.53
	1+9	09.1 ± 0.44	35.3 ± 2.39	<b>14.5*</b> ± 0.76	18.9 ± 0.47	64.8 ± 1.12	29.3 <sup>†</sup> ± 0.63
+ hybrid	1+4	07.7 ± 0.17	40.0 ± 1.18	12.9 <sup>†</sup> ± 0.29	20.7 ± 0.59	61.1 ± 1.25	<b>30.9<sup>†</sup></b> ± 0.71
	1+9	08.5 ± 0.32	36.6 ± 2.09	13.8 <sup>†</sup> ± 0.55	19.4 ± 0.36	62.9 ± 1.20	29.7 <sup>†</sup> ± 0.44
w/ closest cluster selection	1+4	08.3 ± 0.25	40.8 ± 1.00	13.8 <sup>†</sup> ± 0.39	19.3 ± 0.21	63.4 ± 1.08	29.5 <sup>†</sup> ± 0.22
	1+9	08.7 ± 0.27	36.7 ± 1.12	14.1 <sup>†</sup> ± 0.43	19.8 ± 0.41	63.9 ± 0.70	30.2 <sup>†</sup> ± 0.50
+ hybrid	1+4	07.8 ± 0.20	41.6 ± 1.48	13.1 <sup>†</sup> ± 0.34	20.4 ± 0.47	61.0 ± 1.82	30.6 <sup>†</sup> ± 0.65
	1+9	08.6 ± 0.35	37.8 ± 1.52	14.0 <sup>†</sup> ± 0.55	19.2 ± 0.33	62.5 ± 1.23	29.4 <sup>†</sup> ± 0.45
and <i>k</i> -means++, <i>summarization</i>							
w/ furthest cluster selection	1+4	08.1 ± 0.26	38.0 ± 0.83	13.4 <sup>†</sup> ± 0.40	19.1 ± 0.33	64.2 ± 1.06	29.4 <sup>†</sup> ± 0.36
	1+9	09.1 ± 0.44	35.2 ± 1.61	<b>14.5*</b> ± 0.66	19.1 ± 0.26	64.9 ± 1.09	29.5 <sup>†</sup> ± 0.29
+ hybrid	1+4	07.7 ± 0.16	40.3 ± 0.78	12.9 <sup>†</sup> ± 0.23	21.0 ± 0.40	61.4 ± 1.06	<b>31.3<sup>†</sup></b> ± 0.38
	1+9	08.4 ± 0.29	36.6 ± 1.97	13.7 <sup>†</sup> ± 0.50	19.7 ± 0.29	63.3 ± 1.42	30.0 <sup>†</sup> ± 0.38
w/ closest cluster selection	1+4	08.1 ± 0.21	38.6 ± 1.36	13.4 <sup>†</sup> ± 0.36	19.4 ± 0.17	62.9 ± 1.50	29.6 <sup>†</sup> ± 0.22
	1+9	08.9 ± 0.38	36.0 ± 1.81	14.3 <sup>†</sup> ± 0.63	19.2 ± 0.35	63.0 ± 1.04	29.4 <sup>†</sup> ± 0.43
+ hybrid	1+4	07.8 ± 0.20	40.2 ± 1.64	13.0 <sup>†</sup> ± 0.34	20.5 ± 0.46	60.7 ± 1.43	30.6 <sup>†</sup> ± 0.59
	1+9	08.7 ± 0.19	38.4 ± 1.37	14.2 <sup>†</sup> ± 0.31	19.2 ± 0.46	62.1 ± 1.76	29.3 <sup>†</sup> ± 0.65

Table 8: Full results obtained with FS-TACRED using Gemma3-4B and Gemma3-12B. This table complements Table 1 and 6 by providing results with two additional LLMs. Incorporating additional examples results in minor improvements with Gemma3-4B and inconsistent gains with Gemma3-12B. \* and † indicate the best method in the whole column is statistically significantly better at  $p < 0.05$  and  $p < 0.01$ , respectively, based on one-sided paired bootstrap test with replacement on F1 scores.

	#ex.	Gemma3-4B			Gemma3-12B		
		G+A	P	R	F1	P	R
Few-shot baseline	1+0	35.1 ± 0.93	22.8 <sup>†</sup> ± 0.98	27.6 ± 0.96	36.7 ± 0.82	60.3 ± 1.58	45.6 ± 0.81
Including Additional examples							
Paraphrased with LLM	1+4	32.4 ± 1.65	21.7 ± 0.92	26.0 <sup>†</sup> ± 1.12	36.5 ± 0.67	56.8 ± 1.52	44.4 <sup>†</sup> ± 0.70
	1+9	30.1 ± 0.89	17.7 ± 0.60	22.3 <sup>†</sup> ± 0.64	36.2 ± 0.67	59.5 ± 1.58	<b>45.0<sup>†</sup></b> ± 0.69
Generated with LLM	1+4	32.0 ± 0.84	43.9 ± 1.06	<b>37.0</b> ± 0.60	36.5 ± 0.67	56.8 ± 1.52	44.4 <sup>†</sup> ± 0.70
	1+9	35.6 ± 1.21	37.6 ± 1.70	36.5* ± 1.20	36.2 ± 0.74	59.5 ± 1.48	<b>45.0<sup>†</sup></b> ± 0.69
Retrieved (closest) using							
SBERT representations	1+4	35.6 ± 1.76	35.3 ± 1.85	35.4 <sup>†</sup> ± 1.69	34.6 ± 0.85	57.7 ± 1.39	43.2 <sup>†</sup> ± 0.87
	1+9	37.3 ± 1.63	30.1 ± 1.54	33.3 <sup>†</sup> ± 1.35	36.0 ± 0.89	53.5 ± 1.58	43.0 <sup>†</sup> ± 0.97
+ hybrid	1+4	33.2 ± 0.70	42.7 ± 1.34	<b>37.4</b> ± 0.65	34.6 ± 1.04	58.2 ± 1.61	<b>43.4<sup>†</sup></b> ± 1.03
	1+9	37.8 ± 1.13	36.7 ± 1.30	37.2 ± 0.84	36.0 ± 0.78	54.6 ± 1.06	43.3 <sup>†</sup> ± 0.64
Semantic rule representations	1+4	27.8 ± 0.86	45.7 ± 1.83	34.6 <sup>†</sup> ± 0.95	31.5 ± 0.86	60.3 ± 1.61	41.4 <sup>†</sup> ± 0.95
	1+9	28.4 ± 0.55	40.3 ± 1.57	33.3 <sup>†</sup> ± 0.80	31.4 ± 1.04	56.8 ± 1.41	40.4 <sup>†</sup> ± 1.05
+ hybrid	1+4	31.3 ± 0.78	45.3 ± 1.99	<b>37.0</b> ± 1.07	31.7 ± 0.96	60.3 ± 1.70	<b>41.5<sup>†</sup></b> ± 1.05
	1+9	33.2 ± 0.70	39.5 ± 1.80	36.1 <sup>†</sup> ± 0.83	31.3 ± 1.15	56.8 ± 1.80	40.3 <sup>†</sup> ± 1.28
and <i>k</i> -means clustering							
w/ random cluster selection	1+4	27.6 ± 1.06	44.0 ± 2.07	33.9 <sup>†</sup> ± 1.27	31.3 ± 0.83	61.3 ± 1.75	41.4 <sup>†</sup> ± 0.97
	1+9	26.6 ± 0.52	37.1 ± 1.12	31.0 <sup>†</sup> ± 0.66	30.4 ± 0.84	57.4 ± 1.68	39.7 <sup>†</sup> ± 0.87
+ hybrid	1+4	31.1 ± 0.57	45.0 ± 1.55	<b>36.8*</b> ± 0.68	31.5 ± 0.84	61.3 ± 1.62	<b>41.6<sup>†</sup></b> ± 0.92
	1+9	33.6 ± 0.69	39.0 ± 1.78	36.1 <sup>†</sup> ± 1.00	30.5 ± 0.86	58.6 ± 1.46	40.1 <sup>†</sup> ± 0.89
w/ closest cluster selection	1+4	27.3 ± 0.94	43.6 ± 1.43	33.6 <sup>†</sup> ± 0.98	31.0 ± 0.94	61.4 ± 2.28	41.2 <sup>†</sup> ± 1.20
	1+9	27.5 ± 0.86	39.6 ± 0.96	32.5 <sup>†</sup> ± 0.84	30.1 ± 0.93	57.7 ± 1.87	39.6 <sup>†</sup> ± 1.12
+ hybrid	1+4	30.8 ± 0.27	45.0 ± 1.28	36.6* ± 0.42	31.2 ± 0.73	61.9 ± 2.13	41.5 <sup>†</sup> ± 0.93
	1+9	33.0 ± 0.67	38.6 ± 1.78	35.5 <sup>†</sup> ± 1.03	30.6 ± 1.10	59.0 ± 1.90	40.3 <sup>†</sup> ± 1.25
and <i>k</i> -means++ clustering							
w/ furthest cluster selection	1+4	28.1 ± 0.55	45.5 ± 1.49	34.8 <sup>†</sup> ± 0.65	31.2 ± 0.55	60.5 ± 1.68	41.1 <sup>†</sup> ± 0.69
	1+9	25.3 ± 0.89	33.4 ± 1.02	28.8 <sup>†</sup> ± 0.76	30.4 ± 0.81	57.6 ± 1.60	39.7 <sup>†</sup> ± 0.88
+ hybrid	1+4	31.1 ± 0.48	44.8 ± 1.74	36.7* ± 0.58	31.7 ± 0.60	60.9 ± 1.58	<b>41.7<sup>†</sup></b> ± 0.71
	1+9	33.5 ± 0.72	38.2 ± 1.75	35.6 <sup>†</sup> ± 1.10	30.4 ± 0.76	58.6 ± 1.79	40.1 <sup>†</sup> ± 0.85
w/ closest cluster selection	1+4	28.6 ± 0.80	47.1 ± 1.78	35.6 <sup>†</sup> ± 0.94	31.3 ± 0.89	61.3 ± 2.06	41.4 <sup>†</sup> ± 1.00
	1+9	27.1 ± 0.66	38.6 ± 1.68	31.8 <sup>†</sup> ± 0.89	30.3 ± 1.05	57.9 ± 1.73	39.8 <sup>†</sup> ± 1.17
+ hybrid	1+4	31.1 ± 1.00	45.4 ± 1.36	<b>36.9</b> ± 0.94	31.4 ± 0.79	61.5 ± 1.82	41.6 <sup>†</sup> ± 0.87
	1+9	33.8 ± 0.71	39.4 ± 1.52	36.4 <sup>†</sup> ± 0.82	30.2 ± 1.12	58.1 ± 1.80	39.7 <sup>†</sup> ± 1.23
Retrieved (closest) using							
Semantic rule representations							
and <i>k</i> -means, <i>summarization</i>							
w/ random cluster selection	1+4	27.6 ± 1.18	47.9 ± 1.79	35.0 <sup>†</sup> ± 1.18	31.6 ± 1.04	62.1 ± 1.30	41.9 <sup>†</sup> ± 1.08
	1+9	28.0 ± 0.65	41.2 ± 0.95	33.3 <sup>†</sup> ± 0.49	31.3 ± 0.95	60.2 ± 1.63	41.2 <sup>†</sup> ± 1.03
+ hybrid	1+4	28.7 ± 0.30	48.2 ± 1.64	<b>36.0<sup>†</sup></b> ± 0.42	31.7 ± 0.82	61.9 ± 1.04	41.9 <sup>†</sup> ± 0.77
	1+9	30.0 ± 0.47	44.0 ± 1.75	35.6 <sup>†</sup> ± 0.65	31.0 ± 0.85	60.6 ± 1.56	41.0 <sup>†</sup> ± 0.90
w/ closest cluster selection	1+4	26.8 ± 0.74	47.5 ± 1.78	34.2 <sup>†</sup> ± 0.83	31.7 ± 0.74	62.2 ± 1.45	<b>42.0<sup>†</sup></b> ± 0.76
	1+9	27.7 ± 0.49	42.7 ± 1.02	33.6 <sup>†</sup> ± 0.40	31.7 ± 0.82	60.7 ± 1.60	41.6 <sup>†</sup> ± 0.81
+ hybrid	1+4	28.5 ± 0.30	48.2 ± 1.62	35.8 <sup>†</sup> ± 0.51	31.5 ± 0.84	61.9 ± 1.61	41.8 <sup>†</sup> ± 0.89
	1+9	29.7 ± 0.53	44.2 ± 1.35	35.5 <sup>†</sup> ± 0.59	31.1 ± 0.90	60.5 ± 1.43	41.0 <sup>†</sup> ± 0.90
and <i>k</i> -means++, <i>summarization</i>							
w/ furthest cluster selection	1+4	28.0 ± 0.71	45.7 ± 1.47	34.7 <sup>†</sup> ± 0.71	31.4 ± 0.71	62.0 ± 1.58	41.7 <sup>†</sup> ± 0.79
	1+9	28.0 ± 0.84	40.1 ± 1.52	33.0 <sup>†</sup> ± 0.91	31.3 ± 0.86	59.7 ± 1.36	41.1 <sup>†</sup> ± 0.85
+ hybrid	1+4	28.9 ± 0.37	47.4 ± 1.77	35.9 <sup>†</sup> ± 0.70	31.5 ± 0.86	61.9 ± 1.70	41.8 <sup>†</sup> ± 0.90
	1+9	30.5 ± 0.60	43.7 ± 1.40	35.9 <sup>†</sup> ± 0.66	30.9 ± 0.87	60.0 ± 1.49	40.7 <sup>†</sup> ± 0.81
w/ closest cluster selection	1+4	27.8 ± 0.67	48.5 ± 1.39	35.3 <sup>†</sup> ± 0.65	31.7 ± 1.02	62.1 ± 1.92	42.0 <sup>†</sup> ± 1.17
	1+9	28.7 ± 0.63	44.5 ± 1.38	34.9 <sup>†</sup> ± 0.74	31.9 ± 0.97	60.2 ± 1.51	41.7 <sup>†</sup> ± 0.99
+ hybrid	1+4	28.9 ± 0.62	48.4 ± 1.92	<b>36.2<sup>†</sup></b> ± 0.91	31.9 ± 0.84	61.9 ± 1.55	<b>42.1<sup>†</sup></b> ± 0.87
	1+9	30.3 ± 0.61	43.9 ± 1.40	35.9 <sup>†</sup> ± 0.71	31.2 ± 0.98	60.3 ± 1.54	41.1 <sup>†</sup> ± 1.03

Table 9: Full results obtained with FS-FewRel using Gemma3-4B and Gemma3-12B. This table complements Table 5 and 7 by providing results with two additional LLMs. Incorporating additional examples results in minor improvements with Gemma3-4B, and inconsistent gains with Gemma3-12B—the 1-shot baseline is the best-performing system. \* and † indicate the best method in the whole column is statistically significantly better at  $p < 0.05$  and  $p < 0.01$ , respectively, based on one-sided paired bootstrap test with replacement on F1 scores.

	# Addtl. examples	FS-TACRED		FS-FewRel	
		# Generated	# Retrieved	# Generated	# Retrieved
SBERT representations	4	2.94	1.06	3.03	0.97
	9	7.10	2.35	7.16	1.84
Semantic rule representations	4	2.77	1.23	2.88	1.12
	9	6.65	2.35	6.66	2.34
and $k$ -means clustering	4	3.05	0.95	3.08	0.92
	9	7.18	1.82	7.11	1.89

Table 10: Analysis of our hybrid approach using Qwen3-4B on FS-TACRED and FS-FewRel (best performing methods in Tables 1 and 5). Regardless of whether we aim for 4 or 9 additional examples and the dataset, the majority (roughly 2/3) of additional examples are selected from the ones generated with an LLM. The remaining one third, however, are selected from our retrieval approach grounded using semantic rule representations and clustering. Recall that the hybrid approach (as opposed to using only examples generated with an LLM) yields better results (Tables 1 and 5). We hypothesize that this is due to the higher diversity in the retrieved examples (Section 7).

Method	#ex	Gold. vs. Addtl.		Among Addtl.		
		G+A	% overlap	cosine	% overlap	cosine
Paraphrased with LLM	1+4		41.0	0.827	54.1	0.867
	1+9		41.3	0.827	59.1	0.880
Generated with LLM	1+4		7.6	0.275	12.9	0.379
	1+9		8.0	0.272	17.0	0.398
Retrieved (closest) using						
SBERT representations	1+4		8.3	0.361	08.9	0.339
	1+9		8.6	0.345	08.2	0.311
Semantic rule representations	1+4		12.8	0.257	12.3	0.251
	1+9		11.8	0.245	11.0	0.239
+ hybrid	1+4		9.4	0.267	09.9	0.303
	1+9		9.2	0.269	12.6	0.326
and $k$ -means clustering	1+4		7.3	0.171	6.8	0.158
	1+9		7.4	0.170	6.9	0.160
+ hybrid	1+4		6.8	0.278	9.8	0.308
	1+9		7.2	0.247	12.8	0.367

Table 11: Diversity analysis of all the additional examples obtained with several methods on FS-TACRED. We report the average diversity metrics between (a) the gold support sentence (i.e., original example) and the additional examples (columns 3–4) and (b) among the additional examples (columns 5–6). The diversity metrics are the percentage of tokens overlap and the cosine similarity of the SBERT representations. Lower values indicate higher diversity with both metrics. Paraphrasing with an LLM results in additional examples with high token overlap and cosine similarities, indicating low diversity. Generating with an LLM alleviates the issue, but retrieving with semantic rule representations results in additional examples with the lowest token overlaps and cosine similarity. Note that combining both of these examples (hybrid) is more beneficial for in-context learning (Table 6), thus the relation at hand does hold despite the higher diversity.

Dataset	#ex.	Qwen3-4B			Gemma3-4B		
		G+A	P	R	F1	P	R
FS-TACRED							
Binary	1+0	28.7 ± 0.99	17.8 ± 1.21	<b>22.0</b> ± 1.19	11.9 ± 0.60	18.5 ± 0.54	<b>14.5</b> ± 0.58
Multi-class	1+0	05.0 ± 0.15	64.4 ± 1.11	09.3 ± 0.26	05.5 ± 0.18	60.5 ± 1.38	10.2 ± 0.31
FS-FewRel							
Binary	1+0	41.7 ± 1.37	26.4 ± 1.17	<b>32.4</b> ± 1.14	35.1 ± 0.93	22.8 ± 0.98	27.6 ± 0.96
Multi-class	1+0	14.5 ± 0.59	53.8 ± 1.81	22.9 ± 0.86	23.9 ± 0.47	59.3 ± 1.87	<b>34.0</b> ± 0.62

Table 12: Results with the 1-shot setting using the binary (for each of the five relations in an episode, does it hold between the entities at hand?) and multi-class approaches (which of these five relations holds between the entities at hand, if any?). See descriptions of both approaches in Section 3. The binary approach yields stronger results for all corpora and LLMs except Gemma3-4B with FS-FewRel, so we report all our main results with the binary approach.

	#ex. G+A	Qwen3-4B			Gemma3-4B		
		P	R	F1	P	R	F1
FS-TACRED	1+0	28.7 ± 0.99	17.8 ± 1.21	<b>22.0</b> ± 1.19	11.9 ± 0.60	18.5 ± 0.54	<b>14.5</b> ± 0.58
without NER filter	1+0	21.7 ± 0.65	18.9 ± 1.15	20.2 ± 0.93	07.6 ± 0.39	18.4 ± 0.56	10.8 ± 0.47
FS-FewRel	1+0	41.7 ± 1.37	26.4 ± 1.17	<b>32.4</b> ± 1.14	35.1 ± 0.93	22.8 ± 0.98	<b>27.6</b> ± 0.96
without NER filter	1+0	29.4 ± 1.14	28.1 ± 1.14	28.7 ± 0.98	20.9 ± 0.31	23.5 ± 0.92	22.1 ± 0.56

Table 13: Ablation study on the effect of NER-based filtering in the baseline 1-shot setting on FS-TACRED and FS-FewRel using Qwen3-4B and Gemma3-4B (first rows in Tables 1 and 5). Recall that the deterministic filtering simply discards a relation at inference time unless the named entities of the subject and object named entities have the expected types. Removing the filter substantially decreases precision while minimally decreasing recall (resulting in lower F1 across corpora and LLMs), so we keep the filter in our main experiments.