Modeling Complex Semantics Relations with Contrastively Fine-Tuned Relational Encoders

Naïm Es-sebbani[1,2] and **Esteban Marquer**[1] and **Zied Bouraoui**[1]
[1] CRIL, Univ. Artois & CNRS, France [2] GRAMMATICA UR 4521, Univ. Artois, France
{essebbani,marquer,bouraoui}@cril.fr

Abstract

Modeling relationships between concepts and entities is essential for many applications. While Large Language Models (LLMs) capture relational and commonsense knowledge effectively, they are computationally expensive and often underperform in tasks requiring efficient relational encoding, such as relation induction, extraction, and information retrieval. Despite advancements in learning relational embeddings, existing methods often fail to capture nuanced representations and the rich semantics needed for high-quality embeddings. In this work, we propose different relational encoders designed to capture diverse relational aspects and semantic properties of entity pairs. Although several datasets exist for training such encoders, they often rely on structured knowledge bases or predefined schemas, which primarily encode simple and static relations. To overcome this limitation, we also introduce a novel dataset generation method leveraging LLMs to create a diverse spectrum of relationships. Our experiments demonstrate the effectiveness of our proposed encoders and the benefits of our generated dataset.

1 Introduction

Understanding semantic relationships between entities is fundamental to advanced natural language understanding and reasoning. For example, analogy-making requires precise comprehension of the relations between entities (Gentner and Markman, 1997; Turney, 2012; Kumar and Schockaert, 2023; Srikumar and Roth, 2013). The presence of distractors that share only surface-level similarities with a query relation $A : A'$ has been shown to cause performance drops in automatic systems attempting to identify analogous relationships $B : B'$ (Bitton et al., 2023).

Modeling relational knowledge is essential for tasks such as relation induction (Bouraoui et al., 2018; Sun et al., 2022a), relation extraction (Zhong and Chen, 2021; Zhou and Chen, 2022), and question answering (Yasunaga et al., 2021; Sun et al., 2022c; Jiang et al., 2022). Knowledge Graphs (KGs), which represent relations as triplets (h, r, t) between a head entity h and a tail entity t, offer a structured means of encoding relational information. Likewise, Large Language Models (LLMs) have shown strong capabilities in capturing commonsense and relational knowledge (Bouraoui et al., 2020). However, both KGs and LLMs face limitations. KGs are constrained by predefined schemas, which restrict the expressiveness and granularity of relations they can represent (Yasunaga et al., 2021; Sun et al., 2022c; Jiang et al., 2022; Wang et al., 2018). On the other hand, while LLMs are powerful, their high computational cost makes them impractical for large-scale applications involving extensive corpora such as news or scientific texts. In contrast, smaller language models such as BERT offer a more scalable and efficient alternative for relation extraction (Ushio et al., 2023). In this context, relational embeddings provide a compact and cost-effective means of encoding entity relationships without requiring explicit relation labels or rigid schemas. These embeddings leverage the internal knowledge of LMs to capture relational semantics between head and tail entities (Baldini Soares et al., 2019; Hao et al., 2023; Cohen et al., 2023). Recent approaches that employ relational embeddings have achieved strong performance, particularly in relation extraction (Zhong and Chen, 2021; Hao et al., 2023; Cohen et al., 2023; Jiang et al., 2022). Nevertheless, designing efficient and semantically rich relation embeddings remains an open challenge (Ushio et al., 2023).

Although current strategies have shown encouraging outcomes, they may still be less than ideal. First, most relational encoding models, to the best of our knowledge, are fine-tuned on triplets predominantly extracted from KGs. They inherit structural limitations inherent to KG schema. Specif-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20259–20288
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

ically, KGs typically encode simple relations, as complex relationships are conventionally decomposed into sequences of simpler ones. For instance, the relationship between "Paris" and "France" in KGs might only encode simple attributes like "capital of," omitting richer semantic contexts such as "tourist hub" or "cultural center." The reliance on a predefined and closed set of relations constrains the variety and expressiveness of relational contexts, reducing their ability to generalize beyond the training schema. Second, current models often fail to capture the full diversity of semantic properties and the complexity of inter-entity relationships. Their limited capacity for representing nuanced and multifaceted relations hinders performance in tasks that require fine-grained relational reasoning (Kumar and Schockaert, 2023).

To address these concerns, we first introduce a new dataset of over $80,000$ triplets that leverages LLMs to capture a wide spectrum of relational knowledge. Our dataset generation process is flexible and can be extended to incorporate new categories of relations, ensuring adaptability to diverse tasks. When used to train relation encoders, this dataset consistently improves performance across existing models. In addition to the dataset, we propose three contrastive relational encoders designed to capture diverse aspects of relational semantics. Each encoder is fine-tuned using carefully crafted prompts, either to model relational semantics directly or to independently encode the head and tail entities before combining them into a compact relationship representation. To evaluate the effectiveness of these models, we focus on analogy questions and lexical relation classification tasks. The results show that our proposed relational encoders outperform state-of-the-art models in capturing and representing relational semantics.

2 Related Work

Learning Relational Encoders Since their introduction, pre-trained language models (LMs) have been extensively studied to evaluate their capacity for capturing commonsense knowledge (Forbes et al., 2019; Zhou et al., 2020; Roberts et al., 2020) and their potential for modeling relational knowledge (Petroni et al., 2019; Bouraoui et al., 2020; Sun et al., 2022b). To model relational knowledge, (Petroni et al., 2019) proposed a BERT-based model that utilizes manually defined prompt sentences, where the tail entity of the relation to predict is masked. Predictions for the masked token are used for link prediction, completing knowledge graph triples. Similarly, (Bosselut et al., 2019) introduced a fine-tuned GPT-based model for commonsense knowledge graph completion tasks, such as ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019). Several studies have since moved beyond hand-crafted templates, exploring automated methods for template generation. For instance, (Bouraoui et al., 2020) mined text corpora to identify effective templates for specific relations, using these templates with a BERT encoder to evaluate whether a word pair, instantiated within the template, forms a natural sentence. Other works, such as (Jiang et al., 2020; Haviv et al., 2021), developed paraphrasing-based strategies to enhance lexical diversity in templates. AutoPrompt (Shin et al., 2020; Liu et al., 2019) introduced a gradient-based search strategy to identify trigger tokens for creating prompts automatically. Additionally, (Jiang et al., 2020) combined mining-based, paraphrasing-based, and ensemble methods to generate diverse discrete textual prompts for relational prediction.

Despite recent progress, both manually crafted and discrete textual prompts have notable limitations. Hand-written prompts often fail to fully leverage the capabilities of LMs in encoding factual and relational knowledge. Approaches such as RelBERT (Ushio et al., 2023) aim to address this by encoding word pairs into relational embeddings using predefined or automatically generated prompts (e.g., “Paris is the [MASK] of France”) and fine-tuning LMs on datasets like SemEval 2012 Task 2. RelBERT employs loss functions such as triplet loss to optimize relational similarity and constructs relation embeddings by averaging contextualized LM outputs. However, this method may miss important semantic nuances of the relation between entities, due to inherent biases in the prompts and the limited richness of the training datasets.

Relational Reasoning with LLMs. LLMs face challenges in relational reasoning and information extraction, particularly due to difficulties in domain adaptation, implicit relation identification, and the need for specification-heavy instructions. Fine-tuning has proven essential to mitigate these limitations. Peng et al. (2023) highlighted that LLMs often struggle with tasks requiring detailed guidelines. Although LLMs are adept at capturing commonsense knowledge, this strength does not always extend to tasks such as relation extraction, induction, or analogy reasoning, especially in domain-specific or highly constrained contexts. To address such issues, Engelbach et al. (2023) tackled hallucinations in extraction tasks using fine-tuned question-answering models, while Shi and Luo (2024) proposed methods to enhance logical consistency and precision. Despite their broad knowledge, LLMs still fall short in delivering high-quality relational embeddings, which remain critical for robust performance in these tasks (Bitton et al., 2023). Nonetheless, LLMs offer a valuable tool for generating diverse relational triplets that can compensate for the limitations of static, schema-bound datasets. By employing carefully designed prompts, we ensure our dataset encompasses a broad spectrum of relational types, resulting in a rich and nuanced resource for training relational encoders.

3 Deriving Relational Knowledge

Several datasets exist for training relational encoders, and we will revisit them in detail in Section 3.1. However, these datasets often have notable limitations, as they are predominantly derived from structured knowledge bases or predefined schemas that focus on encoding simple and static relationships. Consequently, they struggle to capture the complexity and richness of semantic relationships, particularly those grounded in real-world contexts or commonsense reasoning. To address these shortcomings, we propose a new approach in Section 3.2 that leverages the capabilities of LLMs to capture diverse commonsense knowledge.

3.1 Existing Datasets

Four primary sources are commonly used for training relational encoders, as outlined below. These datasets were meticulously curated in the work of (Ushio et al., 2023) and combined to train the RelBERT model[1].

RelSim This dataset is derived from SemEval 2012 Task 2 and includes crowdsourced annotations on 79 fine-grained semantic relations organized into 10 major categories: *Class Inclusion*, *Part-Whole*, *Similar*, *Contrast*, *Attribute*, *Non-Attribute*, *Case Relation*, *Cause-Purpose*, *Space-Time*, and *Representation*.

ConceptNet This dataset originates from a filtered version of ConceptNet (Li et al., 2016) where certain triples with low confidence scores are excluded during the filtering process to ensure quality.

NELL-One NELL-One (Xiong et al., 2018) is a curated knowledge graph tailored for "one-shot" relational learning and is a cleaned version of the original NELL dataset (Mitchell et al., 2018).

T-REX Constructed from Wikipedia and Wikidata, T-REX (Elsahar et al., 2018), the dataset is reduced to 839 distinct relations, creating a more focused resource for relational learning.

In the following, we will refer to these datasets as Relational Knowledge Bases (RelKB)

3.2 LLM Generated Data

We leverage LLMs as a source of commonsense knowledge to create a semantically diverse dataset for training relational encoders. Using `Llama-3.1-8B-Instruct`[2], we generate triplets (head, relation, tail) that capture a wide range of semantic relationships, including causality, spatial relations. This allows to enrich and complement the existing datasets reported above, offering more varied examples of relations.

Relation categories and prompts To generate more detailed and accurate triplets, some of the RelSim relations were first refined or split into corresponding subcategories. For instance, we distinguished *Spatial* and *Temporal* relations from the original *Space-Time* category in RelSim. Similarly, *Hypernym* relations were complemented by *Hyponymy* to account for the LLMs sensitivity to the directionality of relations in prompts. Moreover, *Meronymy* and *Synonymy* relations were expanded into *Member-Collection* and *Similarity*, respectively, using more descriptive prompts that avoid reliance on the terms “meronym” and “synonym.” This adjustment aims to capture a broader range of related concepts, including triplets that do not strictly align with traditional definitions of Meronymy and Synonymy. Additionally, the *Entailment* relation was introduced to reflect a more logic-driven notion compared to *Causal* relations, enabling the generation of distinct triplets. We also added *Commonsense*, *Functional*, *Collocation*, and *Troponymy* (manner of an action) relations, which were not included in the original RelSim dataset.

[1] All datasets are available at https://huggingface.co/datasets/relbert/

[2] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Finally, we excluded *Non-Attribute* relations, as these describe the absence of attributes and may not be well-suited for an open-world setting relying on the commonsense knowledge encoded in the LLM. As illustration, we use such prompt for leveraging commonsense relation: "*I am interested in knowing knowledge triplets that describe commonsense relation. For example : * glass, falls, breaks, * day, follows, night, * fish, live in, water, * wheel, part of, car, * knife, used for, cutting. Generate a bullet list of 100 different examples.*" For other relation types, we use the same template, replacing the examples and "*commonsense relation*" by examples and a description fitting the kind of considered relation. The set of the relations we consider and corresponding few-shot examples for the prompts of the LLM generated data are provided Tab. 6 in the Appendix.

As shown in the example above, we used five carefully selected examples to guide the LLM's generation process. This assumes that a small set of highly informative examples can significantly improve the diversity and accuracy of generated triplets. To further enhance diversity and avoid linear relational patterns common in sequence-based generative tasks, we formatted the output as a bullet list. This structure encouraged the LLM to treat each triplet as an independent instance, rather than as sequentially related to others. The small-scale generation strategy proved effective in maintaining the quality of the relational data. By limiting the number of triplets generated per prompt, we minimized the risk of reducing semantic precision.

SemRelLM Dataset. To construct our SemRelLM dataset, we iteratively queried the language model until we obtained at least 5,000 valid triplets for each targeted relation type. Statistics for the generated triplets across relation categories are provided in Tab. 1. We retained only successfully parsed triplets, without applying additional filtering. A qualitative analysis of the resulting data is presented in App. A.

4 Contrastive Relation Encoders

In this section, we describe our relational encoder models. We assume that we are given a set of triples $\{(h_1, r_1, t_1), ..., (h_n, r_n, t_n)\}$ as training data where $h =_i$ and t_i are respectively head and tail entities and r_i the relation. The model needs to learn a relation embedding of each relation between a given pair (h, t) of head and tail entities.

Relation category	Unique triplet tokens	Unique triplet lemmas	Unique relation tokens	Unique relation lemmas
Antonymy	5003	4861	44	44
Attribute	5015	4885	5	4
Commonsense	5019	3902	1303	527
Functional	5026	4604	732	661
Collocation	5029	4779	1	1
Troponymy	5014	4883	2	2
Causal	5017	4358	486	317
Entailment	5006	4178	1	1
Spatial	5000	4892	108	106
Temporal	5003	4762	155	133
Hypernymy	5019	4971	218	205
Hyponymy	5001	4935	114	98
Meronymy	5015	4347	17	5
Member-collection	5019	4986	12	10
Synonymy	5014	4961	626	620
Similarity	5039	4852	1	1
All	80239	74576	3021	1811

Table 1: Number of unique triplets and relations in the LLM generated data. Tokens correspond to the output of the LLM, while lemmas are obtained using SpaCy to lemmatize and remove stop words. Relation categories that form complementary pairs are grouped together.

In the following, we propose three encoders aiming to capture rich semantics of relations, illustrated in App. Fig. 4.

4.1 UniPrompt Encoder

Contrastive learning has proven effective in learning meaningful concept embeddings (Li et al., 2023; Kteich et al., 2024). We adapt this idea to relational vectors, aiming to bring vectors that encode similar relationships between head and tail entities closer together, while pushing apart vectors that represent different types of relations. We propose UniPRE, a contrastive learning method for relation encoder that fine-tune LM attention weights to obtain better separation between relation. We use the InfoNCE objective function (Sohn, 2016) with cosine similarity, defined as follows:

$$-\sum_{r} \sum_{a,a^+ \in Pos_r} \log \frac{\exp\left(\frac{\cos\left(\mathsf{f}(a),\mathsf{f}(a^+)\right)}{\tau}\right)}{\sum_{a^- \in Neg_r} \exp\left(\frac{\cos(\mathsf{f}(a),\mathsf{f}(a^-))}{\tau}\right)} \tag{1}$$

where $\mathsf{f}(\cdot)$ is the relation encoder, such that $\mathsf{f}(a)$ is the embedding of relation a, Pos_r is the set of positive head-tail pairs $\{(h_1, t_1), \ldots\}$ associated with relation r, meaning that $a = (h_a, t_a)$ and $a^+ = (h_{a^+}, t_{a^+})$ share the relation r. Neg_r is the set of negative head-tail pairs for relation r, meaning there is no valid triplet where the head and tail are related by r. τ is a temperature parameter that controls the sharpness of the similarity distribution.

A common approach for training the encoder $\mathsf{f}(\cdot)$ is to use the prompt "The relationship between [HEAD] and [TAIL] is [MASK]," where the embedding of the `[MASK]` token serves as the relation representation. As shown in Tab. 10, the choice of prompt has a significant impact on relational encoding performance across multiple analogy datasets. These prompts are crafted to capture different aspects of relational meaning between [HEAD] and [TAIL] pairs, varying in specificity and contextual framing. Among them, the prompt "One property of [HEAD] is to be the [MASK] of [TAIL]" yielded the highest average performance. This suggests that embedding relational properties within a structured and interpretable context enhances the model's ability to capture diverse characteristics of a relation.

However, the performance of a UniPRE is highly domain-dependent. For instance, prompts optimized for part-whole relations perform well on tasks involving compositional relationships but fail to generalize to other types of relations, such as temporal relations. This limitation arises because each prompt inherently biases the encoder toward a specific relational meaning, leaving other relational dimensions underrepresented or poorly encoded.

4.2 MultiPrompts Encoder

Another strategy we follow is to integrate multiple LM encoders, each using a distinct prompt tailored to capture specific relational domains, such as temporal relations. By aggregating embeddings from these diverse prompts, the MultiPRE model generates richer and more balanced relational representations, overcoming the domain-specific biases of UniPRE approaches. Let f_i denote the i-th LM encoder with its own prompt, and (h, t) represent a head-tail pair. To aggregate embeddings, we apply a learnable weighting mechanism, scaling each embedding by a weight w_i. The final embedding, is obtained by concatenating the weighted embeddings, followed by normalization:

$$\mathbf{embedding}(h,t) = \text{Concat}\left(w_1\mathsf{f}_1(h,t),\ldots,w_N\mathsf{f}_N(h,t)\right) \quad (2)$$

where N is the number of LM encoders. We apply L2 normalization on $\mathbf{embedding}(h,t)$. We use the same loss Eq. (1) where f is replaced by the final embedding from this model. The prompt combinations used are presented as prompt 5 and 6 in App. Tab. 10. This approach allows to capture a rich, multi-dimensional representation of relationships by leveraging the diverse relational perspectives offered by multiple encoders. By combining these embeddings through learnable weights, the model dynamically optimizes its understanding of semantic relationships.

4.3 Semantic Properties Encoder

While the multi-prompt encoder effectively captures diverse relational aspects, it primarily focuses on the relationship itself, abstracting away the intrinsic properties of the entities. For instance, it performs well on pairs like "coffee" and "black," where the relation emerges from their combination. However, in cases like "fast" and "food," understanding the relationship requires modeling the semantic attributes of each entity. We propose the Semantic Properties Encoder (SemPRE), a bi-encoder architecture with separate LM encoders for the HEAD and TAIL entities. Using prompts such as "[HEAD] means [MASK]." and "[TAIL] means [MASK]." as suggested in (Kteich et al., 2024), it extracts rich semantic embeddings that capture the core meanings of each entity. These embeddings are then combined using the Hadamard product:

$$\mathbf{embedding}(h,t) = \mathsf{f}_{\text{HEAD}}(h) \odot \mathsf{f}_{\text{TAIL}}(t)$$

The model is also trained using an InfoNCE loss.

5 Experiments and Results

We evaluate our relation encoders through intrinsic tasks, including analogy questions (Sec. 5.2) and lexical relation classification (Sec. 5.3), to assess their ability to capture relational semantics and generalize to unseen relations. The experiments compare our models with SOTA baselines and analyze scalability across different BERT family models (e.g., BERT, RoBERTa, DeBERTa). Additionally, we also conduct experiments with LLMs, such as LLama3.3-70B, to compare performance with our encoders[3]. The performance of LLMs are discussed in App. C.

[3] Our datasets and implementation are available at `https://github.com/essebbaninaim/encoder-acl`,and hyperparameters are detailed in App. D.2

Model	U2	U4	BATS	Google	SCAN	NELL	T-REX	CN	Average
RelBERT-large	67.0	63.0	80.0	**95.0**	**27.0**	65.0	63.0	47.0	63.37
RelKB(RelBERT finetuning data)									
UniPRE-large	61.84	60.65	79.66	92.8	19.74	**77.17**	68.85	47.9	63.58
MultiPRE-large	61.84	67.36	82.32	93.6	20.3	76.33	70.49	**53.52**	65.72
SemPRE-large	31.58	32.64	49.58	53.6	11.51	74.67	56.28	26.85	42.09
SemRelLM(LLM-generated data)									
UniPRE-large	**73.68**	**75.0**	79.1	92.0	20.36	56.17	69.95	37.92	63.02
MultiPRE-large	70.61	74.31	80.77	92.2	23.21	62.67	64.48	37.16	63.18
SemPRE-large	37.72	40.51	49.58	54.8	15.53	54.0	16.39	23.32	36.48
RelKB+SemRelLM									
UniPRE-large	72.81	74.54	82.66	93.6	20.3	73.83	**71.04**	40.69	**66.18**
MultiPRE-large	71.05	73.15	**83.77**	94.6	25.56	69.83	65.57	39.51	65.38
SemPRE-large	38.6	40.05	57.53	62.2	16.96	69.83	38.25	24.83	43.53

Fixed best results not being the ones in bold for some cases

Table 2: Accuracy (in %) on analogy questions. RelBERT results are from (Ushio et al., 2023).

5.1 Experimental Setup

Training data and details We utilized three datasets: *(i)* RelKB used for training RelBERT (Ushio et al., 2021) *(ii)* our newly generated dataset SemRelLM, and *(iii)* a combined dataset RelKB+SemRelLM. We fine-tuned three LMs encoder variants –BERT (Devlin et al., 2019), RoBERTa (Wolf et al., 2020), and DeBERTa (He et al., 2021)– in both small and large versions to evaluate the robustness of our encoders across architectures and sizes. In our main experiments, we report results using DeBERTa as it produces better results. Detailed comparisons across all models and configurations are provided in App. D.4 Tabs. 12 to 23 and App. D.5 Tabs. 25 to 36. We study in Sec. 5.4 the impact of the finetuning dataset.

Baseline models As a baseline, we use the RelBERT models from (Ushio et al., 2023). In particular, we consider two variants of the model, RelBERT-base and RelBERT-large, respectively with RoBERTa-base and RoBERTa-large (Wolf et al., 2020) as they present the best configurations.

While not entirely comparable as they do not provide relational embeddings, we also provide the performance of the latest (at the time of writing) publicly available LLMs, namely GPT-4o and `Llama-3.3-70B-Instruct`[4]. For GPT-4o, we consider only 100 samples for each subset of the benchmarks, due to cost limitations. As such, GPT-4o results should be interpreted with caution.

5.2 Analogy Questions

The first downstream task we consider is analogy questions dataset[5]. The task involves predicting which pair of entities, among a set of candidates, has the most similar relationship to a given query pair, framed as a multiple-choice problem. For instance, for the query pair *strong*:*stronger*, the candidate pairs are *fortunate*:*fortunately*, *tight*:*young*, *tall*:*taller*, *newer*:*larger*. In this example, the expected answer is *tall*:*taller*. In this setting, we encode both the query pair and the candidate pairs using our relational encoders. The candidate pair with the highest cosine similarity to the query pair is selected as the correct answer. This approach is unsupervised, relying solely on the quality of the relation embeddings.

Tab. 2 reports the results on analogy questions for large encoders (results for base encoders are in Tab. 11) for eight datasets (U2, U4, BATS, Google, SCAN, NELL, T-REX and ConceptNet), with accuracy summarized using an average score. RelBERT-large demonstrates strong performance across datasets, achieving an average accuracy of 63.37%, with particularly high scores on Google and NELL.

The proposed relational encoders UniPRE, MultiPRE, and SemPRE show varied performance de-

[4] The prompts utilized are detailed in App. B Figs. 2 and 3.

[5] all the datasets are available at: `https://huggingface.co/datasets/relbert/analogy_questions`

Model	BLESS	CogALexV	EVALution	K&H+N	ROOT09	Average
RelBERT-large	81.25	72.14	53.45	88.59	75.67	74.22
RelKB (RelBERT finetuning data)						
UniPRE-large	88.14	78.86	67.24	93.48	83.67	82.28
MultiPRE-large	90.54	86.82	64.94	94.37	83	83.93
SemPRE-large	80.93	69.4	41.38	92.74	69.67	70.82
SemRelLM(LLM-generated data)						
UniPRE-large	88.46	78.86	64.94	93.41	89.33	83
MultiPRE-large	**92.63**	85.07	67.24	94.37	89.67	85.8
SemPRE-large	79.33	69.9	47.13	93.11	78.0	73.49
RelKB+SemRelLM						
UniPRE-large	88.94	82.09	63.79	93.26	89.0	83.42
MultiPRE-large	91.67	**90.8**	**67.82**	**94.74**	**91.0**	**87.21**
SemPRE-large	80.13	72.14	46.55	93.26	74.0	73.22

Table 3: Micro F1-score (in %) on lexical relation classification. RelBERT results are from (Ushio et al., 2023)

pending on the training data used. When trained on RelKB data, MultiPRE-large achieves the highest average accuracy (65.72%), excelling on datasets such as U4 and ConceptNet. UniPRE-large performs competitively with an average score of 63.58%, particularly strong in NELL and T-REX. However, SemPRE-large lags behind, with an average accuracy of 42.09%, due to its focus on entity-level semantic properties rather than relational nuances. Training on SemRelLM data yields similar trends, with UniPRE-large and MultiPRE-large achieving comparable results (63.02% and 63.18%, respectively), while SemPRE-large underperforms again (36.48%). When combining RelKB and SemRelLM data, UniPRE-large achieves the highest average score (66.18%), with notable improvements in T-REX and NELL, while MultiPRE-large performs slightly lower overall (65.38%).

These results demonstrate the effectiveness of the proposed models, particularly MultiPRE-large and UniPRE-large, in capturing diverse relational semantics. MultiPRE-large leverages multiple relational prompts effectively, while UniPRE-large performs exceptionally well when trained on combined datasets. SemPRE-large, designed to model entity-specific properties, struggles in comparison, highlighting its limitations for analogy questions tasks.

Overall, our model outperforms RelBERT-large, particularly excelling on challenging datasets like T-REX and NELL. For Google, we achieve near SOTA performance (ours: 94.6%; RelBERT: 95%), and on SCAN, the difference is less than 2%. SCAN is somewhat unique, as it focuses solely on two relation types: science and metaphor. The advantage of RelBERT on these datasets can likely be attributed to its better alignment with simpler relational structures.

5.3 Lexical Relation Classification

A second task we consider to evaluate our approach is the lexical relation classification task[6]. In this task, the goal is to classify word pairs into predefined relation categories. We trained a multi-layer perceptron (MLP) with one hidden layer of size 150 and a learning rate of 0.0001. The input to the MLP is the relation embedding of the word pair, while the relational encoder remains frozen during training to evaluate its performance independently. We also evaluate the performance of LLMs in both 0-shot and 5-shot settings. For an example of the prompt used, refer to App. Fig. 2.

Tab. 3 reports the results for large models (results for base model are in Tab. 24) in terme of F1-scores across five datasets (BLESS, CogALexV, EVALution, K&H+N, and ROOT09). RelBERT-large achieves an average score of 74.22%, performing well on datasets like BLESS and K&H+N but underperforming on EVALution highlighting its limitations in handling complex relational semantics.

The proposed encoders demonstrate significant

[6] The dataset is available at: https://huggingface.co/datasets/relbert/lexical_relation_classification

improvements, particularly MultiPRE-large, which achieves the highest scores across all training setups. When trained on RelKB data, MultiPRE-large achieves an average score of 83.93%, excelling in datasets such as CogALexV and K&H+N. UniPRE-large also performs well with an average of 82.28%, while SemPRE-large lags behind at 70.82%, struggling particularly with datasets like EVALution. Training on SemRelLM data further improves performance, with MultiPRE-large achieving an average score of 85.8% and notable gains on ROOT09 and BLESS. Combining RelKB and SemRelLM data yields the best results, with MultiPRE-large achieving a top average score of 87.21%, outperforming UniPRE-large (83.42%) and SemPRE-large (73.22%). MultiPRE-large excels across all datasets, particularly CogALexV, EVALution, and ROOT09, demonstrating its ability to leverage diverse relational prompts and high-quality training data.

Overall, MultiPRE-large consistently outperforms other models, achieving state-of-the-art results, particularly when trained on the combined RelKB and SemRelLM data. This highlights the effectiveness of leveraging diverse relational prompts and high-quality datasets for robust relational encoding.

5.4 Main Factors of Model Performance

Training datasets Considering the average performance for UniPRE and MultiPRE, fine-tuning with the SemRelLM dataset outperforms RelKB for lexical relation classification, particularly for complex tasks like ROOT09 and BLESS which require nuanced relational reasoning not fully captured by RelKB. For analogy questions, the performance of the two datasets is comparable. For both tasks, the best results are achieved by combining SemRelLM and RelKB, highlighting their complementary nature, as SemRelLM enriches relational diversity. Its larger size allows it to cover a broader spectrum, supporting better generalization. Additionally, SemRelLM addresses commonsense relations not fully represented in RelKB, which makes it an effective complement. This combination leads to significant performance improvements, up to +3% on analogy tasks and +4% on lexical relation classification.

Pre-trained model checkpoint Overall, for UniPRE and MultiPRE, fine-tuning DeBERTa outperforms RoBERTa and BERT for both base and large variants. However, for the SemPRE, BERT performs best, followed by DeBERTa and RoBERTa, likely due to BERT's Wikipedia-based training, which better captures entity-dependent relational information. Detailed performance comparisons are provided in App. Tabs. 18 to 23 and 31 to 36 for UniPRE and MultiPRE, and Tabs. 12 to 17 and 25 to 30 for SemPRE.

LLM models As mentioned in Sec. 5.1, we performed experiments with LLaMA and GPT-4o repported in reported in App. Tabs. 8 and 9 and App. C for analogy questions and lexical relation classification respectively. The experimental settings for these LLMs are not aligned with those of our models or the RelBERT baseline, making the results incomparable, and the results for GPT-4o should be taken with caution, as its evaluation was limited to only 100 examples due to cost constraints.

Overall, LLMs outperform our models on simpler analogy tasks (see App. Tab. 8) involving well-known relations like "capital of" or straightforward word pairs such as "simple" and "difficult." However, their performance declines in more challenging settings like the NELL and CN datasets due to the "lost in the middle" phenomenon. In contrast, our models remain robust and consistent, benefiting from compact and focused relational embeddings. In lexical relation classification, LLMs demonstrate strong performance (see App. Tab. 9), highlighting their capacity to understand relational semantics. However, our encoders outperform them, likely due to the additional training of an MLP on top of the embeddings, which fine-tunes the models for task-specific nuances that LLMs may miss in zero-shot or few-shot settings.

6 Conclusion

This paper presents a novel approach to semantic relation encoding, leveraging diverse prompts and embedding fusion to enhance relational representations. Explicit, contextually rich prompts proved crucial for capturing nuanced relationships, while challenges such as factual accuracy and asymmetric relations highlight areas for improvement. Future work will explore automated prompt generation to refine relational encoding further. Moreover, LLMs do not naively produce embeddings, which are essential for applications like retrieval, classification, clustering, and other operations that require robust, semantically meaningful vector representations.

Limitations

Our analysis in this paper focused exclusively on the English language and the fully supervised setting for relational encoders. While we conducted experiments with LLMs, tasks such as analogy questions and lexical classification were limited to open-source LLMs due to the cost of using proprietary models like GPT-4o. For GPT-4o, we performed a limited experiment with only 100 samples to evaluate its capabilities. Our work also raises the question of whether dedicated relation embeddings are still necessary in the era of LLMs. Relation embeddings, however, offer distinct advantages, such as being more effective for modeling relational similarity and providing efficient, task-specific representations that LLMs may not inherently optimize for.

Acknowledgments

This work was supported by ANR-22-CE23-0002 ERIANA and was granted access to the HPC resources of IDRIS under the allocation 2025-A0171013338.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Yonatan Bitton, Ron Yosef, Eliyahu Strugo, Dafna Shahaf, Roy Schwartz, and Gabriel Stanovsky. 2023. Vasr: Visual analogies of situation recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):241–249.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *ACL*, pages 4762–4779. Association for Computational Linguistics.

Zied Bouraoui, José Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7456–7463. AAAI Press.

Zied Bouraoui, Shoaib Jameel, and Steven Schockaert. 2018. Relation induction in word embeddings revisited. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1627–1637, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023. Crawling the internal knowledge-base of language models. In *EACL*, pages 1856–1869. 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. *Preprint*, arXiv:1810.04805.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Matthias Engelbach, Dennis Klau, Felix Scheerer, Jens Drawehn, and Maximilien Kintz. 2023. Fine-tuning and aligning question answering models for complex information extraction tasks. *arXiv preprint*.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? In *CogSci*, pages 1753–1759. cognitivesciencesociety.org.

Dedre Gentner and Arthur B. Markman. 1997. Structure mapping in analogy and similarity. *American Psychologist*, 52:45–56.

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

Adi Haviv, Jonathan Berant, and Amir Globerson. 2021. Bertese: Learning to speak to BERT. In *EACL*, pages 3618–3623. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Jinhao Jiang, Kun Zhou, Ji-Rong Wen, and Xin Zhao. 2022. *great truths are always simple* : a rather simple knowledge encoder for enhancing the commonsense reasoning capacity of pre-trained models. In *NAACL*, pages 1730–1741. Association for Computational Linguistics.

Pengcheng Jiang, Jiacheng Lin, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2024. Genres: Rethinking evaluation for generative relation extraction in the era of large language models. In *NAACL*, pages 2820–2837.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438.

Hanane Kteich, Na Li, Usashi Chatterjee, Zied Bouraoui, and Steven Schockaert. 2024. Modelling commonsense commonalities with multi-facet concept embeddings. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 1467–1480. Association for Computational Linguistics.

Nitesh Kumar and Steven Schockaert. 2023. Solving hard analogy questions with relation embedding chains. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6224–6236. Association for Computational Linguistics.

Na Li, Hanane Kteich, Zied Bouraoui, and Steven Schockaert. 2023. Distilling semantic concept embeddings from contrastively fine-tuned language models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 216–226. ACM.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1445–1455, Berlin, Germany. 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.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2018. Never-ending learning. *Commun. ACM*, 61(5):103–115.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023. When does in-context learning fall short and why? A study on specification-heavy tasks. *CoRR*, abs/2311.08993.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *EMNLP*, pages 2463–2473. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *EMNLP*, pages 5418–5426. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *AAAI-IAAI-EAAI*, pages 3027–3035. AAAI Press.

Zhengpeng Shi and Haoran Luo. 2024. Cre-llm: A domain-specific chinese relation extraction framework with fine-tuned large language model. *arXiv preprint*.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, pages 4222–4235. Association for Computational Linguistics.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451. AAAI Press.

Vivek Srikumar and Dan Roth. 2013. Modeling semantic relations expressed by prepositions. *Transactions of the Association for Computational Linguistics*, 1:231–242.

Lu Sun, Yongliang Shen, and Weiming Lu. 2022a. Minimally-supervised relation induction from pretrained language model. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1776–1786, Seattle, United States. Association for Computational Linguistics.

Lu Sun, Yongliang Shen, and Weiming Lu. 2022b. Minimally-supervised relation induction from pretrained language model. In *NAACL*, pages 1776–1786. Association for Computational Linguistics.

Yueqing Sun, Qi Shi, Le Qi, and Yu Zhang. 2022c. JointLK: Joint reasoning with language models and knowledge graphs for commonsense question answering. In *NAACL*, pages 5049–5060. Association for Computational Linguistics.

Peter D. Turney. 2012. Domain and function: A dual-space model of semantic relations and compositions. *J. Artif. Intell. Res.*, 44:533–585.

Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert. 2021. Distilling relation embeddings from pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9044–9062, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert. 2023. Relbert: Embedding relations with language models. *Preprint*, arXiv:2310.00299.

Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. Dkn: Deep knowledge-aware network for news recommendation. In *WWWC*, WWW '18, page 1835–1844. International World Wide Web Conferences Steering Committee.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, and Clement Delangue et al. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP*, pages 38–45. Association for Computational Linguistics.

Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2018. One-shot relational learning for knowledge graphs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1980–1990, Brussels, Belgium. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *NAACL*, pages 535–546. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 2: Short Papers, Online only, November 20-23, 2022*, pages 161–168. Association for Computational Linguistics.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pretrained language models. In *AAAI-IAAI-EAAI*, pages 9733–9740. AAAI Press.

A Analysis of the quality of LLM generated data

A.1 Qualitative analysis

General observations A qualitative analysis of the triplets generated using the propmts and examples in Appendix Tabs. 6 and 7 reveals that triplets that do not follow the expected (head, relation, tail) structure are present. Examples of such triplets include (fast, rapid, fleet) for *Synonymy*, as it is a sequence of 3 entities with no relation, and (shuffle, slide, is a way to) for *Troponymy*, as the relation and the tail have been swapped. Other triplets that do not follow the expected structure of a triplet have been discarded in the LLM output parsing step. Finally, we identified the presence of triplets generated with inconsistent separation between head, relation, and tail (e.g.: *Hyponymy* triplet (ferrari is a brand of, high, performance car) instead of (ferrari, is a brand of, high performance car)).

Aside from incorrect triplets, the LLM generates triplets that are meaningful and varied in terms of the entities involved and the relations used. As can be seen in Tab. 1, after lemmatizing using SpaCy[7] and excluding stop words, the relation and triplet count does not drop by a significant amount. This is an illustration of the rich lexicon of relations generated for most relations categories, with 90% of *Commensense* relations used in at most 6 triplets. For *Attribute, Collection, Entailment, Meronymy, Member-collection, Similarity, and Troponymy*, where the number of used relation lemmas is small compared to the other relation categories, the LLM adheres closely to the relations provided in the few-shot examples.

Complementary relation categories introduced Compared to (Ushio et al., 2023) we consider 4 new categories of relations and separate 5 of the original categories into pairs of complementary categories. These new categories enable a wider variety of relation to be generated and account for the dependency of the LLM to prompt formulation and the few-shot examples chosen.

Temporal and *Spatial* relations contain both general temporal (e.g.: gives way to, before, after, etc.) and spatial (e.g.: in, on, beside, above, etc.) relations, and relations that are not exclusively temporal or spatial but contain a spatio-temporal aspect (e.g.: (plant, grows, flower), (earth, orbits, sun)). some of the relations can be seen as factual (e.g.: (earth, orbits, sun)) while other express likely relations (e.g.: (cat, in, basket)). The latter is particularly interesting as such triplets are not usually found in KGs, and correspond to commonsense knowledge extracted from the LLM used. Similarly, triplet using the *Causal* prompt cover explicit causal relations (e.g.: leads to, regulates, prevents, etc.) and relations with a causal aspect (e.g.: destroys, that means to cause the destruction), as well as strict (e.g.: destroys, produce, etc.) or non-strict (e.g.: support, promotes, etc.) causes. While not expected when designing the prompt, the prompt for *Functional* relations also resulted in relations with a causal aspect (e.g.: regulates, pollinates, filters, etc.) that correspond to a global and high-level view of what the function of an entity can be.

Some of the added categories provide more specific or higher quality triplets. For instance, the prompt for *Entailment* expands the natural causes described in *Causal* relations with logical (e.g.: (to be married, entails, having a wedding)) and commonsense (e.g.: (to sleep, entails, to be tired)) entailment, including definitions (e.g.: (to be helpful, entails, assist others)) that can be seen as bidirectional entailment. The prompt for *Synonymy* results in about 2 out of 5 triplets being incorrectly formed sequences of 3 entities without relations that are lists of synonyms (ex: (fast, rapid, fleet) or (cold, icy, glacial)). Comparatively, the more explicit and factual description used for *Similarity* results in no such triplets. Finally, while *Meronymy* triplets relate components and what they are a part of (e.g.: (petal, is a part of, flower)), roughly half of *Member-collection* relations specify the generic term use to designate a member (head) of a larger set (tail), which is a literal interpretation of the description in the prompt.

A.2 Quantitative analysis setup

To evaluate the extent of the above-mentioned phenomena, we performed a human evaluation by submitting a sample of the generated triplets to an evaluator among the authors. Inspired by (Jiang et al., 2024), we adapted their evaluation metrics into 8 multiple-choice questions to fit human evaluation and the range of relations generated. For each item, only one answer is possible. The exact description of the items and the possible answers, as well as the examples given in the instructions, are listed hereafter. For each triplet, all the information used during generation is provided to the evaluator

[7] https://spacy.io/api/lemmatizer, using the en_core_web_sm model

as in example Fig. 1.

Grammaticality: Is the triplet (h, r, t) well formed (use the examples as a reference)? *If NO is selected, all other values are ignored, as the triplet is not valid.*

- **YES:** Triplet is of the form (h, r, t) with r a relation from h to t, the latter two being of a similar nature.
- **NO:** In "(fast, rapid, fleet)", "rapid" is not a relation so the triplet is agrammatical. "(honda cr, v, is a model)" is incorrectly separated into head, relation, and tail.

Topical Similarity: Does the relation fit the definition of the category used in the prompt? *If NO is selected, "Uniqueness / diversity (relation)" is not accounted for.*

- **YES:** For instance, relations "leads to, regulates, prevents" are exact match for the definition of causality.
- **MARGINALLY:** If we consider a relaxed interpretation of the definition, then yes. For instance relations "destroys" is a marginal match for causality, as it causes destruction.
- **NO:** For instance, "(honesty, is, essential for trust)" is not a valid triplet for causality, as "is" is not a causal relation.

Uniqueness / diversity (relation): Is the set of triplets different from the provided examples?

- **REFORMULATION:** Reuses a relation of an example relation (to a reformulation). For instance, "causes" is redundant with "is a cause of".
- **SYNONYM:** Relation with close meaning to an example relation, but a different root. For instance, "causes" is close to "influences".
- **RELATED:** Relation without close meaning to an example triplet. For instance, "causes" is related to "destroys".

Fatualness: Is the generated triplet true, or at least usually true?

- **ALWAYS TRUE:** There are no exception, for instance "(cat, is a, mammal)" USUALLY TRUE or COMMON-SENSE: There are some exceptions, but it can be commonly accepted, for instance "(cat, sleeps in, basket)".
- **RARELY TRUE or SUBJECTIVE:** There are some instances where it is true, but usually it is not, for instance "(pig, eats, duck foie gras)".
- **FALSE or HALLUCINATION:** Never true, for instance "(cat, is a, fish)", or nonsensical, for instance "(cat, sleeps in, dog)".

Granularity: Does the triplet express a broad relation or a very specific relation that belongs to the relation category?

- **FINE-GRAINED:** Can be relaxed into triplets that would better fit the relation category, for instance for causality "(speeding, is the main cause of, road traffic injuries)" could be relaxed into "(speeding, causes, road traffic injuries)" and still be adequate.
- **MEDIUM-GRAINED:** Can not be split into triplets that would better fit the relation category, for instance for causality "(storm, causes, floods)".
- **COARSE-GRAINED:** The triplet expresses a broad relation, that could be refined into one or multiple triplets that better fit the relation category, for instance for causality "(storm, influences, crops)" could be refined into "(storm, destroys, crops)".

Originality: Is the generated triplet informative and non-trivial?

- **TRIVIAL:** Triplets that contain no common-sense knowledge, for instance "(cat, eats, cat food)" where de facto "cat food" is food for "cats", or "(good day, opposite of, bad day)" that opposes good with bad with no necessary knowledge on "day".
- **INFORMATIVE:** Triplets that contain common-sense knowledge, for instance "(to be employed, entails, to have a salary)".
- **ORIGINAL or INVOLVED:** Triplets that contain specialist or involved knowledge, for instance "(blue light, more energetic than, red light)" or "(to be employed, entails, to pay taxes)".

Uniqueness / diversity (head): Is the set of triplets different enough from the provided examples?

- **REFORMULATION:** Reuses the head or tail of an example triplet (to a reformulation). For instance, "storm" is redundant with "stormy weather".

- **SYNONYM:** Has a close meaning to an example head or tail, but a different root. For instance, "storm" is synonymous to "tempest".

- **RELATED:** Is in the lexical field of an example head or tail. For instance, "storm" is related to "flood".

- **UNRELATED:** Is not in the lexical field of an example head or tail. For instance, "storm" is unrelated to "cat".

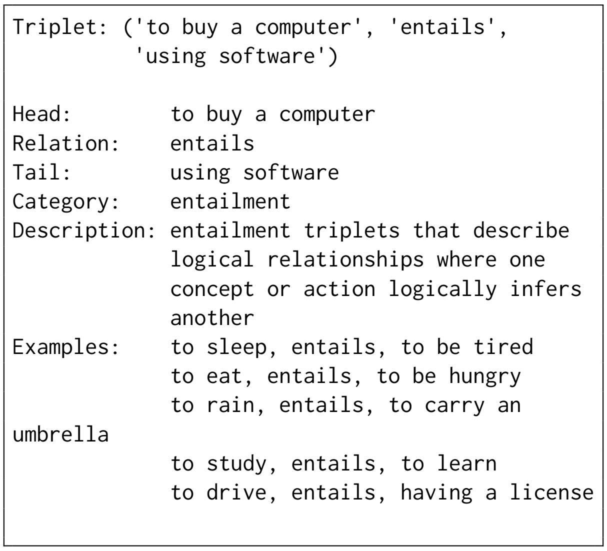
Uniqueness / diversity (tail): Is the set of triplets different enough from the provided examples?

- **REFORMULATION:** Reuses the head or tail of an example triplet (to a reformulation). For instance, "storm" is redundant with "stormy weather".

- **SYNONYM:** Has a close meaning to an example head or tail, but a different root. For instance, "storm" is synonymous to "tempest".

- **RELATED:** Is in the lexical field of an example head or tail. For instance, "storm" is related to "flood".

- **UNRELATED:** Is not in the lexical field of an example head or tail. For instance, "storm" is unrelated to "cat".



```
Triplet: ('to buy a computer', 'entails',
          'using software')

Head:         to buy a computer
Relation:     entails
Tail:         using software
Category:     entailment
Description: entailment triplets that describe
             logical relationships where one
             concept or action logically infers
             another
Examples:     to sleep, entails, to be tired
             to eat, entails, to be hungry
             to rain, entails, to carry an
umbrella
             to study, entails, to learn
             to drive, entails, having a license
```

Figure 1: Context provided for a triplet ('to buy a computer', 'entails', 'using software') generated for the category Entailment.

A.3 Quantitative analysis results

Two of the authors and a non-expert graduate student were recruited for the evaluation.

In a first pilot evaluation, performed by one of the authors, the first item (Grammaticality) was evaluated on 20 randomly selected triplets of each category, and the other items were evaluated on the first 10 selected triplets. In total, 320 triplets were evaluated for Grammaticality only and 160 triplet were evaluated for all items. The results of the pilot study are presented in Tab. 5.

In the second wave of evaluation, evaluators were presented with 10 randomly selected triplets of each category, the first 5 being shared across all annotators for inter annotator agreement measure. In total for all 3 annotators including the pilot study, 80 items were shared and 80 were specific to each annotator. The results are summarized in Tab. 4. Cohen's κ score of inter annotator agreement is as follows for each item:

- Grammaticality: 0.86

- Topical Similarity: 0.41

- Fatualness: 0.39

- Granularity: 0.26

- Originality: 0.16

- Uniqueness / diversity (relation): 0.29

- Uniqueness / diversity (head): 0.30

- Uniqueness / diversity (tail): 0.34

Reliability and overall quality of the triplets Of the 320 triplets considered in the pilot study, 11 were agrammatical (1 from Common-sense and the other 10 from Synonymy), meaning they do not follow the expected HEAD, RELATION, TAIL structure. Over the 160 triplets analyzed in details, 6 were agrammatical, so the questionnaire items excluding Grammaticality were studied over 154 triplets. Out of those 154 triplets, 5 were hallucinations as can be seen in the Fatualness item, which means they do not describe true information. In the full scale study, similar proportions were observed, with less than 4% of agrammatical triplets or hallucinations.

		Proportion of item (in %)
Item	Choice	
Grammaticality	YES	96.74
	NO	3.26
Topical Similarity	YES	92.73
	MARGINALLY	4.49
	NO	2.78
Fatualness	ALWAYS TRUE	54.70
	USUALLY T. or C.-S.	32.05
	RARELY T. or SUBJ.	9.40
	FALSE or HALLU.	3.85
Granularity	FINE-GRAINED	4.49
	MEDIUM-GRAINED	94.87
	COARSE-GRAINED	0.64
Originality	ORIGINAL or INV.	12.69
	INFORMATIVE	78.56
	TRIVIAL	8.75
Uniqueness / diversity (relation)	RELATED	16.08
	SYNONYM	34.36
	REFORMULATION	49.56
Uniqueness / diversity (head)	UNRELATED	82.69
	RELATED	8.55
	SYNONYM	3.63
	REFORMULATION	5.13
Uniqueness / diversity (tail)	UNRELATED	83.54
	RELATED	12.18
	SYNONYM	2.14
	REFORMULATION	2.14
Uniqueness / diversity (highest relatedness between head and tail per triplet)	UNRELATED	79.28
	RELATED	10.04
	SYNONYM	4.27
	REFORMULATION	6.41
Uniqueness / diversity (lowest relatedness between head and tail per triplet)	UNRELATED	86.97
	RELATED	10.68
	SYNONYM	1.50
	REFORMULATION	0.85

Table 4: Detailed result for each relation category, cumulated over all answers of the 3 annotators. The top items correspond to items in the questionnaire while the bottom ones are computed from the answers to “Uniqueness / diversity (head)” and “Uniqueness / diversity (tail)”. "USUALLY T. or C.-S." stands for "USUALLY TRUE or COMMON SENSE"; "RARELY T. or SUBJ." stands for "RARELY T. or SUBJ."; "FALSE or HALLU." stands for "FALSE or HALLU."; "ORIGINAL or INV." stands for "ORIGINAL or INV.".

If we consider the added value of the generated triplet measured with the Originality item, we can observe only 8.75% of trivial items such as (big opportunity, opposite of, small opportunity) or (dell inspiron, is a model of, dell). Interestingly, close to 12.7% of triplets can be considered as displaying non-trivial knowledge, such as (strut support, is a component of, aircraft), (to have a dream, entails, setting goals) or (finger print, similar to, signature). Most of these involved triplets belong to the Functional category, such as (historian, preserves, traditions), or the Attribute and Meronymy categories.

Quality of the relation In the pilot study, only 2 triplets had relations that did not match the relation category they were generated for (Topical Similarity: NO). Additionally, as can be seen in the Granularity item, 3 Causal triplets were overly precise (FINE-GRAINED), and only 2 triplets were not precise enough (COARSE-GRAINED). Overall, the vast majority of the generated relation were suitable in granularity and meaning for the category they belong to.

The originality of the generated triplets in terms of the relation can be observed with the Uniqueness / diversity (relation) item. Some relation categories where a variety of relations were used in the examples (such as Functional, Common-Sense, and Causal), display a variety of relations in the generated triplets, while for categories with a narrow set of valid formulations (like Entailment, Attribute, Collocation, etc.) the generate triplets closely match the provided examples.

Originality of the head and tail While the Originality item reveals how trivial or involved the knowledge expressed by a triplet is, it is not sufficient to confirm that the generated triplets are not just repetitions of the elements in the provided examples. To address this, the Uniqueness / diversity (head) and Uniqueness / diversity (tail) items allow us to consider the originality of the HEAD and TAIL of each triplet. Over all HEADs and TAILs, we observe above 80% of terms that are completely unrelated to the provided examples for the category, with less than 10% of REFORMULATIONs and SYNONYMs to terms present in the examples. This indicates that while the majority of triplets contain original elements, a significant amount of HEADs and TAILS are still close to the provided examples. On the plus side, when we consider both the HEAD and TAIL of the triplet, only 2.35% of triplets have both HEAD and TAIL as a SYNONYM or REFORMULATION (lowest relatedness, last group of rows of Tab. 4). Comparatively, 89.32% of triplets have both HEAD and TAIL at most in the same lexical field of the ones of the examples (highest relatedness, before-last group of rows of Tab. 4).

Overall, despite the lack of instruction like "do not reuse terms that appear in the examples for the HEAD or TAIL", the LLM tends to generate diverse HEADs and TAILs compared to the provided examples.

B LLMs Prompts

```
You are tasked with identifying the
most analogous pair of words based on
the given example.
The example pair is: lie and
prevaricate.

From the following list, choose the
most fitting pair:

1. betray and trust
2. philander and donate
3. waver and falter
4. deride and praise
5. corroborate and doubt

Only provide the number of the correct
answer.
```

Figure 2: Example of the prompt used for GPT-4o and `Llama-3.3-70B-Instruct` in 0-shot mode for the analogy question task.

```
You are tasked with identifying the
most fitting relationship of the given
example.
The example pair is: turtle and live.

From the following list, choose the
most fitting relationship:

1. hyper
2. coord
3. mero
4. random
5. attri
6. event

Only provide the number of the correct
answer.
```

Figure 3: Example of the prompt used for GPT-4o and `Llama-3.3-70B-Instruct` in 0-shot mode for the lexical relation classification task.

C LLMs performance

The performance of GPT-4o in both 0-shot and 5-shot settings is reported in Appendix Tab. 8 for analogy question and Appendix Tab. 9 for lexical relation classification. Note that the evaluation is limited to the first 100 items, which should be considered when interpreting the results.

For analogy questions, Llama3.3 and GPT-4o in 0-shot and 5-shot settings show strong performance in specific datasets like BATS and Google but has inconsistent results overall with average scores of 56.49% (0-shot) and 59.59% (5-shot) for Llama3.3. This contrasts with our proposed models and RelBERT, with for instance RelBERT-large demonstrating strong performance across datasets, achieving an average accuracy of 63.37%, with particularly high scores on Google and NELL.

Llama3.3 performs poorly in the 0-shot setting for lexical relation classification, with an average score of 43.0%, and struggles on datasets like K&H+N (13%) and CogALexV. In the 5-shot setting, its performance improves significantly, achieving an average of 70.4% and performing well on ROOT09 and CogALexV. However, it still falls short of task-specific models like MultiPRE-large, underscoring the advantages of specialized relational encoders. While Llama3.3 shows promise in few-shot, it remains less effective than relational encoders, further emphasizing the importance of

Item	Choice	entailment	hypernymy	attribute	hyponymy	functional	meronymy	temporal	common-sens	member-collection	similarity	synonymy	collocation	troponymy	antonymy	spatial	causal	**TOTAL**
Grammaticality	YES	20	20	20	20	20	20	20	19	20	20	10	20	20	20	20	20	**309**
	NO	0	0	0	0	0	0	0	1	0	0	10	0	0	0	0	0	**11**
Topical Similarity	YES	10	10	10	9	9	10	8	8	9	10	4	10	10	10	10	10	**147**
	MARGINALLY	0	0	0	0	1	0	1	1	1	0	1	0	0	0	0	0	**5**
	NO	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	**2**
Fatualness	ALWAYS TRUE	3	7	5	10	2	9	5	3	3	6	4	8	10	9	0	2	**86**
	USUALLY T. or C.-S.	7	0	2	0	7	1	3	4	7	3	1	2	0	0	3	8	**48**
	RARELY T. or SUBJ.	0	1	3	0	1	0	1	2	0	1	0	0	0	0	6	0	**15**
	FALSE or HALLU.	0	2	0	0	0	0	1	0	0	0	0	0	0	1	1	0	**5**
Granularity	FINE-GRAINED	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	**3**
	MEDIUM-GRAINED	10	9	10	10	9	10	10	9	10	10	5	10	10	10	10	7	**149**
	COARSE-GRAINED	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	**2**
Originality	ORIGINAL or INV.	5	1	5	0	8	4	1	2	2	3	0	1	1	0	2	4	**39**
	INFORMATIVE	5	9	5	9	2	5	9	7	8	7	5	8	9	8	8	6	**110**
	TRIVIAL	0	0	0	1	0	1	0	0	0	0	0	1	0	2	0	0	**5**
Uniqueness / diversity (relation)	RELATED	0	1	0	0	10	0	1	8	0	0	0	0	0	0	1	3	**24**
	SYNONYM	0	2	0	4	0	0	2	0	0	0	0	0	0	0	1	3	**12**
	REFORMULATION	10	7	10	5	0	10	6	1	10	10	5	10	10	10	8	4	**116**
Uniqueness / diversity (head)	UNRELATED	8	9	10	5	7	7	5	8	6	5	2	8	8	10	4	7	**109**
	RELATED	1	0	0	5	1	3	3	0	3	4	2	0	2	0	2	1	**27**
	SYNONYM	1	0	0	0	0	0	1	0	0	1	1	2	0	0	0	1	**7**
	REFORMULATION	0	1	0	0	2	0	1	1	1	0	0	0	0	0	4	1	**11**
Uniqueness / diversity (tail)	UNRELATED	9	9	9	5	9	6	7	8	5	6	3	8	8	10	4	6	**112**
	RELATED	0	1	1	5	1	3	2	1	4	3	1	1	2	0	5	3	**33**
	SYNONYM	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0	1	**5**
	REFORMULATION	1	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	**4**
Uniqueness / diversity (head + tail)	UNRELATED	17	18	19	10	16	13	12	16	11	11	5	16	16	20	8	13	**221**
	RELATED	1	1	1	10	2	6	5	1	7	7	3	1	4	0	7	4	**60**
	SYNONYM	1	0	0	0	0	0	2	0	1	2	2	2	0	0	0	2	**12**
	REFORMULATION	1	1	0	0	2	1	1	1	1	0	0	1	0	0	5	1	**15**
Uniqueness / diversity (highest relatedness between head and tail per triplet)	UNRELATED	8	9	9	5	7	6	5	8	5	5	2	8	8	10	3	5	**103**
	RELATED	1	0	1	5	1	3	3	0	4	3	2	0	2	0	3	2	**30**
	SYNONYM	0	0	0	0	0	0	1	0	0	2	1	1	0	0	0	2	**7**
	REFORMULATION	1	1	0	0	2	1	1	1	1	0	0	1	0	0	4	1	**14**
Uniqueness / diversity (lowest relatedness between head and tail per triplet)	UNRELATED	9	9	10	5	9	7	7	8	6	6	3	8	8	10	5	8	**118**
	RELATED	0	1	0	5	1	3	2	1	3	4	1	1	2	0	4	2	**30**
	SYNONYM	1	0	0	0	0	0	1	0	1	0	1	1	0	0	0	0	**5**
	REFORMULATION	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	**1**

Table 5: Detailed result of the pilot study, for each relation category. The top items correspond to items in the questionnaire while the bottom ones are computed from the answers to “Uniqueness / diversity (head)” and “Uniqueness / diversity (tail)”. "USUALLY T. or C.-S." stands for "USUALLY TRUE or COMMON SENSE"; "RARELY T. or SUBJ." stands for "RARELY TRUE or SUBJECTIVE"; "FALSE or HALLU." stands for "FALSE or HALLUCINATION"; "ORIGINAL or INV." stands for "ORIGINAL or INVOLVED".

Relation	Corresponding RelSim relation	Description
Commonsense	–	knowing knowledge triplets that describe commonsense relation
Functional	–	functional triplets that describe the function or role of an object or entity
Collocation	–	collocation triplets that describe words that frequently occur together
Troponymy	–	troponymy triplets that describe specific ways in which an action is performed
Antonymy	Antonym	antonymy triplets that describe opposite concepts or entities
Attribute	Attribute	attribute triplets that describe inherent properties or qualities of objects or entities
Causal	Cause-Purpose	causal triplets that describe cause and effect relationships
Entailment	Cause-Purpose	entailment triplets that describe logical relationships where one concept or action logically infers another
Spatial	Space-Time	spatial triplets that describe spatial relationship between objects or entities
Temporal	Space-Time	temporal triplets that describe the temporal relationship between events or states
Hyponymy	Hypernym	hyponymy triplets that describe the relationship where one term is a subtype or a specific instance of another
Hypernymy	Hypernym	hypernymy triplets that describe the relationship where one term encompasses a broader category to which the other belongs
Meronymy	Meronym	meronymy triplets that describe part-whole relationships
Member-collection	Meronym	member-collection triplets that describe the relationship between an individual item and the group to which it belongs
Synonymy	Synonym	synonymy triplets that describe similar or identical concepts
Similarity	Synonym	similarity triplets that describe the likeness or analogies between different concepts or entities

Table 6: Relations considered and corresponding description for the prompts of the LLM generated data.

Relation	Few-shot examples
Commonsense	(pen, writes with, ink), (chair, sits on, person), (fire, burns, wood), (house, lives in, family), (pencil, used for, drawing)
Functional	(engine, powers, machine), (nurse, cares, sick person), (beekeeper, monitors, honey production), (ocean, regulates, climate), (carpenter, builds, house), (researcher, discovers, new knowledge)
Collocation	(fire, and, flames), (wind, and, breeze), (coffee, and, cream), (tea, and, leaves), (sand, and, beach), (mountain, and, peak)
Troponymy	(hurry, is a way to, move), (gallop, is a way to, run), (trudge, is a way to, walk), (amble, is a way to, walk), (skip, is a way to, walk), (dash, is a way to, run)
Antonymy	(joy, opposite of, sorrow), (empty, opposite of, full), (silence, opposite of, noise), (morning, opposite of, night), (old, opposite of, young), (pure, opposite of, impure)
Attribute	(snow, has, whiteness), (silk, has, softness), (copper, has, conductivity), (butter, has, richness), (charcoal, has, carbon content), (perfume, has, fragrance)
Causal	(rain, brings, flowers), (learning, improves, memory), (eating, provides, energy), (sleep, helps, concentration), (reading, expands, vocabulary), (meditation, reduces, stress)
Entailment	(to read a book, entails, having eyes), (to write, entails, having a pen), (to bake a cake, entails, having flour), (to plant a garden, entails, having soil), (to learn a language, entails, studying vocabulary), (to be married, entails, having a wedding)
Spatial	(chair, beside, sofa), (kite, in, sky), (dog, on, beach), (phone, in, pocket), (book, on, desk), (man, on, bike)
Temporal	(childhood, before, adolescence), (learning, before, graduation), (voting, before, election), (winter, before, spring), (childhood, before, adulthood), (learning, before, promotion)
Hyponymy	(ferrari, is a brand of, car), (pineapple, is a type of, fruit), (guitar, is a type of, instrument), (laptop, is a type of, computer), (dolphin, is a type of, mammal), (trumpet, is a type of, instrument)
Hypernymy	(sport, categorizes, basketball), (animal, encompasses, cat), (music, includes, jazz), (food, categorizes, pizza), (building, covers, house), (poem, encompasses, sonnet)
Meronymy	(gear, is a component of, machine), (atom, is a component of, molecule), (fin, is a part of, fish), (bolt, is a component of, lock), (thread, is a part of, fabric), (blade, is a part of, windmill)
Member-collection	(bee, member of, swarm), (fish, part of, school), (tree, part of, forest), (book, part of, library), (student, member of, class)
Synonymy	(expensive, similar to, costly), (beautiful, similar to, lovely), (angry, similar to, irate), (busy, similar to, hectic), (soft, similar to, gentle), (sharp, similar to, pointed)
Similarity	(wings, similar to, arms of a butterfly), (bee, similar to, worker in a factory), (pen, similar to, brush), (voice, similar to, song), (wheels, similar to, pedals of a bicycle), (clouds, similar to, cotton balls)

Table 7: Relations considered and corresponding few-shot examples for the prompts of the LLM generated data.

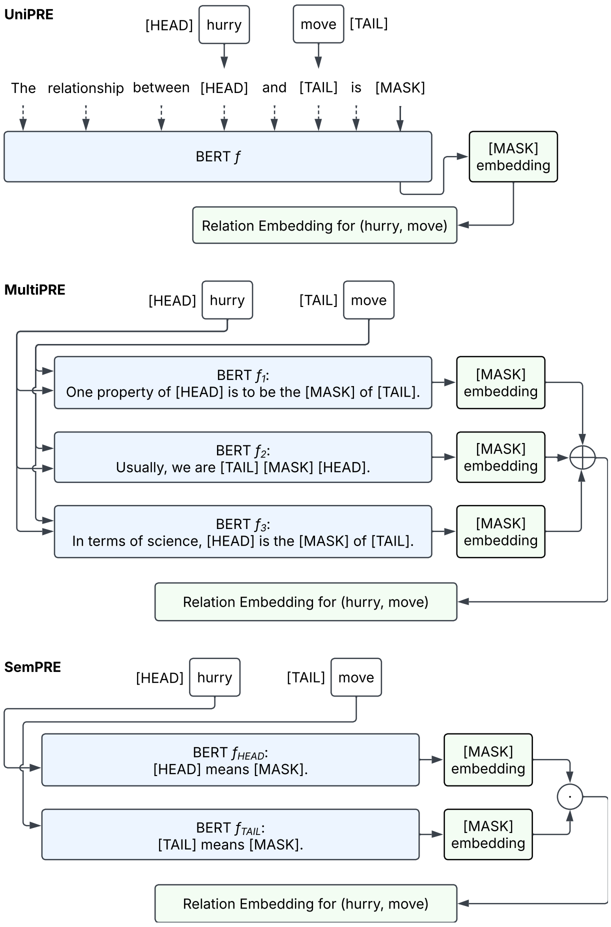

UniPRE
[HEAD]
hurry
move
[TAIL]
The relationship between [HEAD] and [TAIL] is [MASK]
BERT f
[MASK] embedding
Relation Embedding for (hurry, move)
MultiPRE
[HEAD]
hurry
[TAIL]
move
BERT f_1: One property of [HEAD] is to be the [MASK] of [TAIL].
[MASK] embedding
BERT f_2: Usually, we are [TAIL] [MASK] [HEAD].
[MASK] embedding
BERT f_3: In terms of science, [HEAD] is the [MASK] of [TAIL].
[MASK] embedding
Relation Embedding for (hurry, move)
SemPRE
[HEAD]
hurry
[TAIL]
move
BERT f_{HEAD}: [HEAD] means [MASK].
[MASK] embedding
BERT f_{TAIL}: [TAIL] means [MASK].
[MASK] embedding
Relation Embedding for (hurry, move)

Figure 4: Example of the proposed models for the triplet (hurry, is a way to, move) from the Troponymy category. UniPRE uses `prompt1` and MultiPRE uses `prompt5` from Tab. 10.

task-specific training for lexical relation classification.

D Details for our models

D.1 Structure

In Fig. 4 we illustrate the workings of the three models proposed in Sec. 4. $\bigoplus$ corresponds to concatenation, and $\odot$ corresponds to the Hadamar product.

D.2 Training hyperparameters

The training process for all encoders uses early stopping with a patience of 30, a batch size of 256, a learning rate of 1×10^{-5}, and a temperature parameter set to 0.03. These hyperparameters are chosen to ensure stability and convergence across diverse model variants and tasks.

D.3 Prompts

Tab. 10 show all different prompts used. Prompts 1 to 4 are designed for the Uni-Aspect Relation Encoder (UniPRE), while Prompts 5 and 6 are tailored for the Multi-Aspects Encoder (MultiPRE).

D.4 Analogy

Refer to Tab. 11 for a comparison between our DeBERTa-based encoder and RelBERT on the analogy question task.

D.4.1 Semantic Properties Encoder

Tabs. 12 to 17 detail the performance of the SemPRE model on Analogy Questions for DeBERTa, RoBERTa, and BERT base and large checkpoints, for each combination of fine-tuning datasets.

D.4.2 Single and Multi Aspect LM

Tabs. 18 to 23 detail the performance of the UniPRE and MultiPRE on Analogy Questions for DeBERTa, RoBERTa, and BERT base and large checkpoints, for each combination of fine-tuning datasets.

D.5 Lexical relation classification

Refer to Tab. 24 for a comparison between our DeBERTa-based encoder and RelBERT on the lexical relation classification task.

D.5.1 Semantic Properties Encoder

Tabs. 25 to 30 detail the performance of SemPRE on Lexical Relation Classification for DeBERTa, RoBERTa, and BERT base and large checkpoints, for each combination of fine-tuning datasets.

D.5.2 Single and Multi Aspect LM

Tabs. 31 to 36 detail the performance of UniPRE and MultiPRE on Lexical Relation Classification for DeBERTa, RoBERTa, and BERT base and large checkpoints, for each combination of fine-tuning datasets.

Model	U2	U4	BATS	Google	SCAN	NELL	T-REX	CN	Average
GPT-4o 0-shot	84	81	99	**100**	25	48	79	13	66.13
GPT-4o 5-shot	**87**	**86**	**100**	99	32	**89**	**79**	26	**74.75**
Llama3.3 0-shot	76	68	95	97	22	38	45	**29**	58.75
Llama3.3 5-shot	81	69	97	98	**36**	55	54	25	64.38

Table 8: Accuracy (in %) on analogy questions, was test on the 100 first items per datasets.

Model	BLESS	CogALexV	EVALution	K&H+N	ROOT09	Average
GPT-4o 0-shot	57	75	61	56	83	66.4
GPT-4o 5-shot	**79**	**91**	**70**	**84**	**85**	**81.8**
Llama3.3 0-shot	43	52	58	13	44	42.0
Llama3.3 5-shot	69	73	64	71	75	70.4

Table 9: Micro F1-score (in %) on lexical relation classification, was test on the 100 first items per datasets.

Name	Model	Template
`prompt1`	UniPRE	The relationship between [HEAD] and [TAIL] is <mask>.
`prompt2`	UniPRE	The word that best describes the relationship between [HEAD] and [TAIL] is <mask>.
`prompt3`	UniPRE	People often use the word <mask> to describe the relationship between [HEAD] and [TAIL].
`prompt4`	UniPRE	One property of [HEAD] is to be the <mask> of [TAIL].
`prompt5`	MultiPRE	- One property of [HEAD] is to be the <mask> of [TAIL]. - Usually, we are [TAIL] <mask> [HEAD].
`prompt6`	MultiPRE	- One property of [HEAD] is to be the <mask> of [TAIL]. - Usually, we are [TAIL] <mask> [HEAD]. - In terms of science, [HEAD] is the <mask> of [TAIL].

Table 10: Link between prompt name and prompt template

Model	U2	U4	BATS	Google	SCAN	NELL	T-REX	CN	Average
RelBERT-base	59.65	57.41	70.32	89.20	**25.93**	62.0	66.67	39.77	58.86
RelKB									
UniPRE-base	42.11	49.31	69.93	88.8	19.31	76.5	59.02	**41.19**	55.77
MultiPRE-base	53.51	56.48	72.21	92.4	19.43	**78.83**	62.84	35.07	58.85
SemPRE-base	34.65	32.64	45.53	47.4	10.4	71.5	32.24	20.47	36.85
SemRelLM									
UniPRE-base	63.16	67.59	72.15	89.0	20.61	59.67	63.93	34.06	58.77
MultiPRE-base	**68.42**	69.91	77.6	**94.4**	23.08	66.83	66.67	35.32	62.78
SemPRE-base	34.21	40.05	50.64	53.8	15.28	55.83	26.78	20.55	37.14
RelKB+SemRelLM									
UniPRE-base	63.6	67.13	75.32	91.2	21.23	74.17	64.48	35.32	61.56
MultiPRE-base	67.98	**71.06**	**78.04**	93.6	24.75	72.33	**68.85**	37.25	**64.23**
SemPRE-base	34.65	39.58	50.47	63.8	15.53	76.5	37.16	23.32	42.63

Table 11: Accuracy (in %) on analogy questions. For our models, we use DeBERTa. RelBERT models are reproduced following (Ushio et al., 2023).

U2	U4	BATS	GOOGLE	SCAN	NELL	T-REX	CN	Average
RelKB+SemRelLM								
34.65	39.58	50.47	**63.8**	**15.53**	**76.5**	**37.16**	**23.32**	**42.63**
RelKB								
34.65	32.64	45.53	47.4	10.4	71.5	32.24	20.47	36.85
SemRelLM								
34.21	**40.05**	**50.64**	53.8	15.28	55.83	26.78	20.55	37.14

Table 12: SemPRE performance on analogy with deberta-base.

U2	U4	BATS	GOOGLE	SCAN	NELL	T-REX	CN	Average
RelKB+SemRelLM								
28.95	31.48	25.74	32.8	5.94	26.50	14.75	7.63	21.72
RelKB								
26.75	**31.71**	**34.07**	**37.4**	**8.23**	**39.67**	**26.23**	**11.74**	**26.97**
SemRelLM								
28.95	30.56	26.18	28.8	6.19	24.00	12.57	8.47	20.71

Table 13: SemPRE performance on analogy with roberta-large.

U2	U4	BATS	GOOGLE	SCAN	NELL	T-REX	CN	Average
RelKB+SemRelLM								
38.60	40.05	**57.53**	**62.2**	**16.96**	69.83	38.25	24.83	**43.53**
RelKB								
31.58	32.64	49.58	53.6	11.51	**74.67**	**56.28**	**26.85**	42.09
SemRelLM								
37.72	**40.51**	49.58	54.8	15.53	54.00	16.39	23.32	36.48

Table 14: SemPRE performance on analogy with deberta-large.

U2	U4	BATS	GOOGLE	SCAN	NELL	T-REX	CN	Average
RelKB+SemRelLM								
24.12	**29.63**	**33.24**	**45.0**	**7.49**	45.83	18.03	**12.5**	26.98
RelKB								
30.7	28.94	26.29	37.8	4.02	**59.0**	**26.78**	8.39	**27.74**
SemRelLM								
6.32	28.01	27.52	31.8	6.75	23.83	6.01	9.06	19.91

Table 15: SemPRE performance on analogy with roberta-base

U2	U4	BATS	GOOGLE	SCAN	NELL	T-REX	CN	Average
RelKB+SemRelLM								
39.47	45.6	**48.25**	51.4	**26.73**	65.33	49.73	21.56	43.51
RelKB								
36.84	36.11	52.75	**54.6**	19.43	**69.83**	**61.75**	**25.08**	**44.55**
SemRelLM								
40.79	**45.83**	43.41	41.8	25.06	51.83	31.15	19.04	37.36

Table 16: SemPRE performance on analogy with bert-base

U2	U4	BATS	GOOGLE	SCAN	NELL	T-REX	CN	Average
RelKB+SemRelLM								
37.28	45.14	**52.03**	**57.8**	**28.03**	66.33	60.66	**23.07**	**46.29**
RelKB								
34.21	33.56	49.36	45.2	21.97	**71.0**	**61.75**	22.15	42.4
SemRelLM								
41.67	**47.92**	48.75	54.4	27.6	57.0	47.54	22.82	43.46

Table 17: SemPRE performance on analogy with bert-large

	U2	U4	BATS	GOOGLE	SCAN	NELL	T-REX	CN	Average
RelKB+SemRelLM									
prompt1	60.96	59.72	69.65	88.6	17.88	53.33	44.26	24.16	52.32
prompt2	63.16	62.27	71.76	90.0	19.55	52.5	50.27	21.73	53.91
prompt3	63.16	62.27	71.93	89.0	20.42	49.67	47.54	25.0	53.62
prompt4	63.6	67.13	75.32	91.2	21.23	74.17	64.48	35.32	61.56
prompt5	**69.74**	**71.06**	77.49	91.2	22.28	72.17	64.48	35.32	62.97
prompt6	67.98	**71.06**	**78.04**	93.6	**24.75**	72.33	**68.85**	37.25	**64.23**
RelKB									
prompt1	47.81	46.53	61.87	83.8	13.68	67.33	45.36	20.72	48.39
prompt2	53.51	52.08	69.98	89.8	14.67	67.5	50.82	17.95	52.04
prompt3	42.54	49.07	68.32	80.2	16.21	70.0	62.84	32.21	52.67
prompt4	42.11	49.31	69.93	88.8	19.31	76.5	59.02	41.19	55.77
prompt5	44.74	53.24	73.54	90.8	18.32	76.33	60.11	**38.51**	56.95
prompt6	53.51	56.48	72.21	92.4	19.43	**78.83**	62.84	35.07	58.85
SemRelLM									
prompt1	61.4	61.34	70.71	88.4	17.64	46.17	37.7	23.66	50.88
prompt2	62.72	63.89	69.09	87.4	17.33	47.17	40.44	26.01	51.76
prompt3	61.84	63.43	67.59	85.6	17.39	46.67	40.44	21.81	50.6
prompt4	63.16	67.59	72.15	89.0	20.61	59.67	63.93	34.06	58.77
prompt5	65.35	64.58	70.54	88.2	20.61	62.0	56.83	32.89	57.62
prompt6	68.42	69.91	77.6	**94.4**	23.08	66.83	66.67	35.32	62.78

Table 18: Accuracy (in %) for Uni-Aspect (prompts 1 to 4) and Multi-Aspect (prompts 5 and 6) Encoders on Analogy Tasks with deberta-base

	U2	**U4**	**BATS**	**GOOGLE**	**SCAN**	**NELL**	**T-REX**	**CN**	**Average**
RelKB+SemRelLM									
prompt1	57.89	65.05	65.76	78.4	19.37	71.5	70.49	32.47	57.62
prompt2	59.21	62.73	67.59	82.0	21.66	73.0	72.68	31.12	58.75
prompt3	59.21	63.66	70.87	86.0	22.52	73.5	65.03	32.05	59.1
prompt4	63.6	67.13	70.48	84.6	26.98	**80.0**	**76.5**	42.45	**63.97**
prompt5	64.04	65.51	66.93	81.4	24.57	78.83	67.76	42.11	61.39
prompt6	63.6	68.29	70.65	84.2	**28.03**	77.67	73.77	39.01	63.15
RelKB									
prompt1	45.61	52.31	70.26	84.2	21.04	74.0	71.04	42.53	57.62
prompt2	45.61	54.4	75.6	87.6	21.04	74.67	74.86	53.78	60.95
prompt3	48.25	52.55	67.87	83.4	21.41	73.5	63.93	33.47	55.55
prompt4	54.39	55.09	78.93	92.0	23.95	77.5	71.58	54.03	63.43
prompt5	50.44	55.56	76.99	90.4	22.83	75.33	68.85	**54.45**	61.86
prompt6	47.81	53.94	**79.6**	**93.0**	27.66	78.33	69.95	48.74	62.38
SemRelLM									
prompt1	60.09	63.43	61.37	79.0	20.67	63.5	71.58	31.29	56.37
prompt2	63.16	64.12	60.92	82.2	19.74	56.67	66.67	25.42	54.86
prompt3	59.21	62.5	62.09	80.0	18.19	53.67	65.57	27.43	53.58
prompt4	61.84	65.51	64.81	82.6	22.77	70.0	71.04	34.98	59.19
prompt5	**66.67**	**68.75**	62.03	80.6	22.65	66.0	74.32	37.42	59.8
prompt6	64.04	**68.75**	68.93	83.8	26.92	70.0	73.77	36.33	61.57

Table 19: Accuracy (in %) for Uni-Aspect (prompts 1 to 4) and Multi-Aspect (prompts 5 and 6) Encoders on Analogy Tasks with roberta-large

	U2	**U4**	**BATS**	**GOOGLE**	**SCAN**	**NELL**	**T-REX**	**CN**	**Average**
RelKB+SemRelLM									
prompt1	70.18	73.15	78.32	88.4	21.41	66.83	67.21	32.72	62.28
prompt2	66.67	68.52	82.49	93.2	20.48	62.33	58.47	29.78	60.24
prompt3	64.04	69.91	79.99	90.8	21.41	64.5	63.39	36.66	61.34
prompt4	72.81	74.54	82.66	93.6	20.3	73.83	71.04	40.69	66.18
prompt5	72.37	73.15	82.27	94.8	21.47	68.17	67.76	38.42	64.8
prompt6	71.05	73.15	83.77	94.6	25.56	69.83	65.57	39.51	65.38
RelKB									
prompt1	55.26	59.26	80.21	90.2	18.07	69.67	70.49	43.54	60.84
prompt2	60.09	61.11	80.82	94.2	18.19	74.5	69.4	38.76	62.13
prompt3	56.58	59.26	78.77	91.2	18.25	73.5	65.57	41.02	60.52
prompt4	61.84	60.65	79.66	92.8	19.74	77.17	68.85	47.9	63.58
prompt5	60.53	65.05	79.1	91.8	18.63	75.83	68.85	54.7	64.31
prompt6	61.84	67.36	82.32	93.6	20.3	76.33	70.49	53.52	65.72
SemRelLM									
prompt1	67.98	69.68	76.6	92.0	19.55	46.17	56.83	29.53	57.29
prompt2	71.49	72.22	80.1	93.4	21.29	56.17	52.46	30.54	59.71
prompt3	67.98	71.76	76.43	90.0	18.44	53.33	53.01	30.62	57.7
prompt4	73.68	75.0	79.1	92.0	20.36	56.17	69.95	37.92	63.02
prompt5	71.93	70.6	78.93	92.8	21.35	62.5	66.67	38.34	62.89
prompt6	70.61	74.31	80.77	92.2	23.21	62.67	64.48	37.16	63.18

Table 20: Accuracy (in %) for Uni-Aspect (prompts 1 to 4) and Multi-Aspect (prompts 5 and 6) Encoders on Analogy Tasks with deberta-large

	U2	**U4**	**BATS**	**GOOGLE**	**SCAN**	**NELL**	**T-REX**	**CN**	**Average**
RelKB+SemRelLM									
prompt1	59.65	62.27	63.15	82.8	24.94	74.83	65.57	33.98	58.4
prompt2	58.33	61.57	63.59	85.4	25.43	73.83	59.56	34.14	57.73
prompt3	55.7	60.65	67.09	86.0	24.81	72.17	59.56	31.96	57.24
prompt4	55.7	59.72	64.81	86.0	28.03	77.33	67.76	36.91	59.53
prompt5	59.65	63.89	71.98	92.8	30.69	79.67	66.12	37.67	62.81
prompt6	60.53	62.96	73.43	93.8	32.12	77.67	65.57	39.09	63.15
RelKB									
prompt1	45.18	46.3	70.59	86.0	20.73	74.83	70.49	47.23	57.67
prompt2	45.61	49.77	70.48	86.2	22.71	74.67	62.3	41.11	56.61
prompt3	39.04	46.3	68.2	84.0	21.6	74.17	60.11	39.09	54.06
prompt4	49.12	49.77	70.43	88.8	21.84	77.0	69.95	48.32	59.4
prompt5	47.37	54.4	72.04	88.0	26.42	76.17	66.67	38.84	58.74
prompt6	48.68	53.47	75.21	94.2	24.44	77.17	71.04	49.75	61.75
SemRelLM									
prompt1	59.65	62.04	59.92	82.2	23.82	67.83	64.48	32.05	56.5
prompt2	53.51	59.26	64.54	85.2	25.0	55.83	57.92	26.01	53.41
prompt3	53.95	56.94	62.92	85.2	24.32	59.83	64.48	29.78	54.68
prompt4	57.89	62.73	63.42	85.6	26.05	72.33	72.68	35.15	59.48
prompt5	59.65	64.35	68.59	91.0	29.46	76.83	66.12	36.83	61.6
prompt6	59.21	62.27	66.81	87.4	28.28	71.83	66.12	37.08	59.88

Table 21: Accuracy (in %) for Uni-Aspect (prompts 1 to 4) and Multi-Aspect (prompts 5 and 6) Encoders on Analogy Tasks with roberta-base

	U2	**U4**	**BATS**	**GOOGLE**	**SCAN**	**NELL**	**T-REX**	**CN**	**Average**
RelKB+SemRelLM									
prompt1	43.42	53.94	63.2	75.4	26.55	67.33	63.93	28.44	52.78
prompt2	44.74	49.31	58.92	72.6	27.48	62.67	61.2	23.66	50.07
prompt3	47.37	50.23	61.37	74.0	25.37	65.33	57.92	25.42	50.88
prompt4	50.0	53.94	70.09	81.8	28.4	73.67	72.68	34.4	58.12
prompt5	47.81	57.87	74.21	88.4	30.2	74.0	63.39	33.47	58.67
prompt6	45.61	48.15	74.1	87.2	32.36	72.67	68.31	30.45	57.36
RelKB									
prompt1	39.91	39.81	60.87	67.6	17.95	68.83	59.56	20.97	46.94
prompt2	37.28	39.58	62.76	64.6	20.85	69.67	68.31	22.73	48.22
prompt3	39.04	39.12	62.76	71.8	18.19	72.0	64.48	26.59	49.25
prompt4	42.54	42.13	73.6	85.8	23.45	74.33	74.32	40.52	57.09
prompt5	44.3	42.59	74.82	87.0	26.49	75.67	67.76	35.23	56.73
prompt6	45.61	44.68	73.87	88.0	28.34	75.83	72.68	34.56	57.95
SemRelLM									
prompt1	46.05	53.24	59.37	69.0	24.75	63.67	56.83	25.5	49.8
prompt2	42.98	49.07	54.25	63.2	27.6	57.33	50.82	19.8	45.63
prompt3	42.54	50.23	50.69	58.6	23.27	45.0	44.81	15.27	41.3
prompt4	47.81	53.7	66.43	79.2	28.84	65.17	66.12	33.39	55.08
prompt5	49.12	57.64	71.76	81.8	29.7	65.83	63.93	33.31	56.64
prompt6	47.81	56.71	73.1	87.4	32.12	66.83	64.48	31.8	57.53

Table 22: Accuracy (in %) for Uni-Aspect (prompts 1 to 4) and Multi-Aspect (prompts 5 and 6) Encoders on Analogy Tasks with bert-base

	U2	U4	BATS	GOOGLE	SCAN	NELL	T-REX	CN	Average
					RelKB+SemRelLM				
prompt1	46.93	51.62	62.59	75.8	21.84	63.67	52.46	20.39	49.41
prompt2	54.82	56.94	68.43	82.0	27.29	65.5	71.58	30.29	57.11
prompt3	55.7	59.95	69.48	81.2	27.66	69.0	71.58	30.7	58.16
prompt4	52.19	58.8	72.15	87.0	31.56	72.17	73.77	35.4	60.38
prompt5	57.46	60.42	75.82	91.4	31.68	73.17	65.03	35.49	61.31
prompt6	59.65	62.5	75.43	90.8	29.76	73.83	68.85	35.74	62.07
					RelKB				
prompt1	42.11	45.14	67.76	77.6	21.97	70.17	70.49	38.51	54.22
prompt2	38.16	46.06	67.93	81.2	22.22	69.17	74.86	38.93	54.82
prompt3	43.42	47.22	64.59	78.4	22.34	69.0	72.13	30.37	53.43
prompt4	42.98	47.22	73.99	88.6	26.24	75.33	73.22	41.95	58.69
prompt5	45.61	51.39	78.04	88.4	26.73	75.83	72.13	48.07	60.78
prompt6	43.86	45.37	72.98	86.6	24.63	75.5	72.68	39.6	57.65
					SemRelLM				
prompt1	55.7	60.42	61.7	74.2	21.66	64.67	61.75	27.18	53.41
prompt2	56.14	59.26	66.43	79.2	27.35	64.67	55.74	30.29	54.89
prompt3	47.81	55.79	64.2	76.4	26.79	59.67	56.28	28.1	51.88
prompt4	53.95	61.11	71.76	87.4	30.57	69.17	77.6	36.16	60.97
prompt5	59.21	63.66	73.15	89.2	30.45	65.5	69.95	33.98	60.64
prompt6	58.33	61.11	74.71	90.4	31.81	65.67	69.4	36.74	61.02

Table 23: Accuracy (in %) for Uni-Aspect (prompts 1 to 4) and Multi-Aspect (prompts 5 and 6) Encoders on Analogy Tasks with bert-large

Model	BLESS	CogALexV	EVALution	K&H+N	ROOT09	Average
RelBERT-base	76.92	71.64	48.85	85.78	78.33	72.30
			RelKB			
UniPRE-base	86.54	73.38	55.75	91.04	84.33	78.21
MultiPRE-base	88.46	83.58	64.37	93.11	88.33	83.57
SemPRE-base	77.08	66.17	48.28	91.48	70.67	70.74
			SemRelLM			
UniPRE-base	84.29	78.36	60.92	91.11	88.33	80.6
MultiPRE-base	89.26	82.84	66.67	92.81	90.33	84.38
SemPRE-base	74.84	67.66	44.25	91.78	70.33	69.77
			RelKB+SemRelLM			
UniPRE-base	84.62	77.61	63.22	91.04	88.33	80.96
MultiPRE-base	89.1	83.58	66.67	93.56	90.33	84.65
SemPRE-base	75.64	67.66	43.68	91.93	74.67	70.72

Table 24: Micro F1-score (in %) on lexical relation classification. For our models, we use DeBERTa and train with RelBERT dataset.

BLESS	CogALexV	EVALution	K&H+N	ROOT09	Average
		RelKB+SemRelLM			
75.64	67.66	43.68	91.93	74.67	70.72
		RelKB			
77.08	66.17	48.28	91.48	70.67	70.74
		SemRelLM			
74.84	67.66	44.25	91.78	70.33	69.77

Table 25: Semantic Properties performance on lexical relation classification with deberta-base

BLESS	CogALexV	EVALution	K&H+N	ROOT09	Average
RelKB+SemRelLM					
71.15	74.38	43.68	87.78	67.67	68.93
RelKB					
77.4	73.88	44.83	88.52	72.0	71.33
SemRelLM					
71.31	74.13	45.4	87.33	69.0	69.43

Table 26: Semantic Properties performance on lexical relation classification with roberta-large

BLESS	CogALexV	EVALution	K&H+N	ROOT09	Average
RelKB+SemRelLM					
80.13	72.14	46.55	93.26	74.0	73.22
RelKB					
80.93	69.4	41.38	92.74	69.67	70.82
SemRelLM					
79.33	69.9	47.13	93.11	78.0	73.49

Table 27: Semantic Properties performance on lexical relation classification with deberta-large

BLESS	CogALexV	EVALution	K&H+N	ROOT09	Average
RelKB+SemRelLM					
68.75	73.13	44.25	82.0	65.0	66.63
RelKB					
58.65	72.14	33.91	81.41	57.33	60.69
SemRelLM					
63.46	72.14	39.66	83.78	63.33	64.47

Table 28: Semantic Properties performance on lexical relation classification with roberta-base

BLESS	CogALexV	EVALution	K&H+N	ROOT09	Average
RelKB+SemRelLM					
82.37	74.88	48.28	90.3	76.33	74.43
RelKB					
80.29	73.63	53.45	90.52	75.33	74.64
SemRelLM					
79.65	76.87	50.57	90.81	74.0	74.38

Table 29: Semantic Properties performance on lexical relation classification with bert-base

BLESS	CogALexV	EVALution	K&H+N	ROOT09	Average
RelKB+SemRelLM					
83.33	75.37	58.05	92.44	77.0	77.24
RelKB					
82.37	72.64	48.28	93.48	76.33	74.62
SemRelLM					
83.81	73.88	52.3	92.52	79.0	76.3

Table 30: Semantic Properties performance on lexical relation classification with bert-large

	BLESS	CogALexV	EVALution	K&H+N	ROOT09	Average
RelKB+SemRelLM						
prompt1	84.29	78.61	58.62	90.74	89.33	80.32
prompt2	83.33	78.11	55.75	90.96	88.0	79.23
prompt3	84.29	76.62	54.6	91.56	87.67	78.95
prompt4	84.62	77.61	63.22	91.04	88.33	80.96
prompt5	87.82	81.84	66.09	92.96	88.67	83.48
prompt6	89.1	83.58	66.67	93.56	90.33	84.65
RelKB						
prompt1	82.53	72.89	50.57	90.96	86.33	76.66
prompt2	86.06	73.13	52.87	88.81	87.33	77.64
prompt3	85.9	72.89	52.3	90.0	83.33	76.88
prompt4	86.54	73.38	55.75	91.04	84.33	78.21
prompt5	87.82	77.61	64.37	92.0	86.0	81.56
prompt6	88.46	83.58	64.37	93.11	88.33	83.57
SemRelLM						
prompt1	82.05	77.61	54.6	91.04	89.33	78.93
prompt2	82.21	78.61	57.47	90.81	90.0	79.82
prompt3	81.57	76.62	54.02	90.96	88.0	78.23
prompt4	84.29	78.36	60.92	91.11	88.33	80.6
prompt5	89.42	78.86	68.39	93.11	88.33	83.62
prompt6	89.26	82.84	66.67	92.81	90.33	84.38

Table 31: MultiPRE performance on lexical relation classification with deberta-base

	BLESS	CogALexV	EVALution	K&H+N	ROOT09	Average
RelKB+SemRelLM						
prompt1	85.9	76.37	63.22	91.26	85.67	80.48
prompt2	86.7	76.87	64.94	91.85	87.33	81.54
prompt3	88.3	75.12	62.07	91.93	87.67	81.02
prompt4	87.98	77.11	64.37	91.7	88.33	81.9
prompt5	85.74	75.87	64.94	92.22	83.67	80.49
prompt6	88.46	79.35	68.39	93.63	89.33	83.83
RelKB						
prompt1	80.93	72.89	56.32	89.85	84.67	76.93
prompt2	83.33	73.13	54.6	89.7	84.33	77.02
prompt3	84.78	72.89	50.57	89.04	83.0	76.06
prompt4	85.26	76.12	63.22	89.41	83.67	79.54
prompt5	86.22	75.37	65.52	91.04	85.0	80.63
prompt6	88.14	79.1	63.22	92.67	86.67	81.96
SemRelLM						
prompt1	84.13	76.87	62.64	91.33	84.0	79.79
prompt2	82.69	76.62	59.2	90.15	81.67	78.07
prompt3	82.53	76.37	60.92	90.44	83.0	78.65
prompt4	84.94	75.87	59.2	91.33	83.67	79.0
prompt5	84.62	75.37	64.37	90.59	83.67	79.72
prompt6	87.34	79.35	64.37	92.74	84.67	81.69

Table 32: MultiPRE performance on lexical relation classification with roberta-large

	BLESS	CogALexV	EVALution	K&H+N	ROOT09	Average
			RelKB+SemRelLM			
prompt1	89.58	79.85	68.97	93.78	87.67	83.97
prompt2	91.35	80.85	70.69	93.85	90.33	85.41
prompt3	88.62	80.85	66.67	93.78	90.0	83.98
prompt4	88.94	82.09	63.79	93.26	89.0	83.42
prompt5	91.67	86.32	67.24	93.56	87.33	85.22
prompt6	91.67	90.8	67.82	94.74	91.0	87.21
			RelKB			
prompt1	90.54	75.87	63.79	93.26	86.33	81.96
prompt2	91.35	75.37	58.05	93.26	89.33	81.47
prompt3	90.38	76.12	64.94	93.48	88.0	82.58
prompt4	88.14	78.86	67.24	93.48	83.67	82.28
prompt5	89.42	82.09	65.52	93.85	86.33	83.44
prompt6	90.54	86.82	64.94	94.37	83.0	83.93
			SemRelLM			
prompt1	88.94	80.35	63.22	93.7	89.67	83.18
prompt2	89.26	81.09	63.79	93.7	91.67	83.9
prompt3	88.3	80.6	65.52	92.81	90.33	83.51
prompt4	88.46	78.86	64.94	93.41	89.33	83.0
prompt5	91.51	83.33	66.09	94.52	87.33	84.56
prompt6	92.63	85.07	67.24	94.37	89.67	85.8

Table 33: MultiPRE performance on lexical relation classification with deberta-large

	BLESS	CogALexV	EVALution	K&H+N	ROOT09	Average
			RelKB+SemRelLM			
prompt1	85.42	77.11	62.07	90.15	83.33	79.62
prompt2	86.06	77.11	59.2	89.11	85.67	79.43
prompt3	85.42	78.86	61.49	89.78	86.33	80.38
prompt4	84.29	77.61	61.49	89.85	82.0	79.05
prompt5	86.86	80.6	64.37	90.74	87.0	81.91
prompt6	88.14	82.34	61.49	92.07	84.0	81.61
			RelKB			
prompt1	81.73	74.38	58.05	86.96	81.67	76.56
prompt2	83.81	74.13	56.9	88.07	84.33	77.45
prompt3	82.85	72.89	54.6	87.63	83.67	76.33
prompt4	82.85	74.88	54.6	87.26	81.0	76.12
prompt5	86.06	77.11	61.49	89.93	84.67	79.85
prompt6	87.66	81.84	62.64	91.11	86.0	81.85
			SemRelLM			
prompt1	85.58	76.37	62.64	89.41	82.67	79.33
prompt2	85.1	79.1	56.9	89.85	83.0	78.79
prompt3	85.42	79.6	58.62	89.85	84.0	79.5
prompt4	85.1	77.11	62.64	88.96	85.33	79.83
prompt5	88.14	80.35	60.92	90.67	85.0	81.02
prompt6	87.18	80.35	63.22	90.81	86.0	81.51

Table 34: MultiPRE performance on lexical relation classification with roberta-base

	BLESS	**CogALexV**	**EVALution**	**K&H+N**	**ROOT09**	**Average**
			RelKB+SemRelLM			
prompt1	85.58	75.37	58.05	89.85	82.0	78.17
prompt2	83.81	75.12	51.72	90.59	82.33	76.71
prompt3	84.13	75.87	56.32	90.67	81.67	77.73
prompt4	83.65	75.37	56.32	90.67	82.67	77.74
prompt5	86.7	78.36	58.05	90.59	85.67	79.87
prompt6	87.66	78.11	59.77	91.85	82.67	80.01
			RelKB			
prompt1	81.73	72.89	43.1	87.04	78.67	72.69
prompt2	83.17	72.89	47.13	89.48	83.33	75.2
prompt3	83.33	72.89	48.28	88.0	81.67	74.83
prompt4	82.21	73.38	52.3	88.37	79.0	75.05
prompt5	83.81	76.87	57.47	90.37	82.67	78.24
prompt6	85.9	77.61	59.77	91.41	83.33	79.6
			SemRelLM			
prompt1	84.46	74.38	54.6	89.26	80.67	76.67
prompt2	83.17	75.37	51.15	90.59	80.0	76.06
prompt3	78.53	73.38	51.15	89.11	79.0	74.23
prompt4	85.1	75.37	58.62	90.3	81.67	78.21
prompt5	85.58	76.87	58.62	90.81	85.33	79.44
prompt6	87.66	79.35	60.92	91.48	83.33	80.55

Table 35: MultiPRE performance on lexical relation classification with bert-base

	BLESS	**CogALexV**	**EVALution**	**K&H+N**	**ROOT09**	**Average**
			RelKB+SemRelLM			
prompt1	82.85	76.37	54.02	91.48	83.67	77.68
prompt2	85.26	76.62	54.6	91.85	84.0	78.47
prompt3	87.18	76.12	56.9	90.89	84.33	79.08
prompt4	86.86	77.36	63.79	92.67	82.33	80.6
prompt5	87.5	79.6	62.64	92.52	84.67	81.39
prompt6	87.66	78.36	66.09	92.59	83.67	81.67
			RelKB			
prompt1	83.81	73.63	55.17	89.93	81.67	76.84
prompt2	84.62	73.13	53.45	90.0	82.0	76.64
prompt3	84.29	72.89	55.75	88.81	81.67	76.68
prompt4	83.97	73.88	58.05	90.89	80.33	77.42
prompt5	86.54	77.11	62.64	91.11	83.67	80.21
prompt6	85.9	77.36	59.77	92.59	81.67	79.46
			SemRelLM			
prompt1	84.13	76.37	56.9	90.59	82.0	78.0
prompt2	83.33	77.36	54.02	91.7	84.0	78.08
prompt3	85.58	76.12	56.32	91.93	82.67	78.52
prompt4	85.58	76.37	58.05	92.74	81.33	78.81
prompt5	87.18	78.11	60.34	92.89	86.0	80.9
prompt6	88.62	79.35	60.92	92.67	84.0	81.11

Table 36: MultiPRE performance on lexical relation classification with bert-large