Few-Shot, No Problem: Descriptive Continual Relation Extraction

Nguyen Xuan Thanh [1*], **Anh Duc Le**[2*], **Quyen Tran** [3*], **Thanh-Thien Le** [3*],
Linh Ngo Van [2†], **Thien Huu Nguyen** [4]

[1]Oraichain Labs,
[2]Hanoi University of Science and Technology,
[3]VinAI Research,
[4]University of Oregon
thanh.nx@orai.io
anh.ld204628@sis.hust.edu.vn
{v.quyentt15,v.thienlt3}@vinai.io,
linhnv@soict.hust.edu.vn, thien@cs.uoregon.edu

Abstract

Few-shot Continual Relation Extraction is a crucial challenge for enabling AI systems to identify and adapt to evolving relationships in dynamic real-world domains. Traditional memory-based approaches often overfit to limited samples, failing to reinforce old knowledge, with the scarcity of data in few-shot scenarios further exacerbating these issues by hindering effective data augmentation in the latent space. In this paper, we propose a novel retrieval-based solution, starting with a large language model to generate descriptions for each relation. From these descriptions, we introduce a bi-encoder retrieval training paradigm to enrich both sample and class representation learning. Leveraging these enhanced representations, we design a retrieval-based prediction method where each sample "retrieves" the best fitting relation via a reciprocal rank fusion score that integrates both relation description vectors and class prototypes. Extensive experiments on multiple datasets demonstrate that our method significantly advances the state-of-the-art by maintaining robust performance across sequential tasks, effectively addressing catastrophic forgetting.

1 Introduction

Relation Extraction (RE) refers to classifying semantic relationships between entities within text into predefined types. Conventional RE tasks assume all relations are present at once, ignoring the fact that new relations continually emerge in the real world. Few-shot Continual Relation Extraction (FCRE) is a subfield of continual learning (Hai et al. 2024; Van et al. 2022; Phan et al. 2022; Tran et al. 2024a,b; Le et al. 2024a) where a model must continually assimilate new emerging relations while avoiding the forgetting of old ones, a task made even more challenging by the limited training data available. The importance of FCRE stems from its relevance to dynamic real-world applications, garnering increasing interest in the field (Chen, Wu, and Shi 2023a; Le et al. 2024c, 2025).

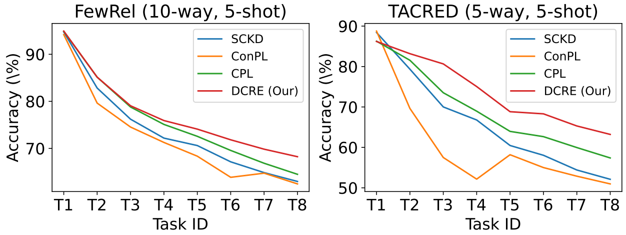


Figure 1: Existing FCRE methods face catastrophic forgetting due to the limited and poor quality of old training samples stored in the memory buffer.

State-of-the-art approaches to FCRE often rely on memory-based methods for continual learning (Lopez-Paz and Ranzato 2017; Nguyen et al. 2023; Le et al. 2024b; Dao et al. 2024). However, these methods frequently suffer from overfitting to the limited samples stored in memory buffers. This overfitting hampers the reinforcement of previously learned knowledge, leading to catastrophic forgetting—a marked decline in performance on learnt relations when new ones are introduced (Figure 1). The few-shot scenario of FCRE exacerbates these issues, as the scarcity of data not only impedes learning on new tasks, but also hinders helpful data augmentation, which are crucial in many methods (Shin et al. 2017).

In order to improve on these methods, we must not completely disregard them or dwell on their weaknesses, but rather contemplate their biggest strength. *Why do so many methods use the memory buffer in the first place?* The primary objective of these replay buffers is to rehearse and reinforce past knowledge, providing the model with something to "look back" at during training. However, these past samples may not always be representative of the entire class and can still lead to sub-optimal performance. Based on this observation, we propose a straightforward: besides relying

[*]These authors contributed equally.
[†]Corresponding Author

on potentially unrepresentative past samples, we leverage our knowledge of the past relations themselves. This insight leads to our approach of generating detailed descriptions for each relation. These descriptions inherently represent the class more accurately than the underlying information from a set of samples, serving as stable pivots for the model to align with past knowledge while learning new information. By using these descriptions, we create a more robust and effective method for Few-Shot Continual Relation Extraction, ensuring better retention of knowledge across tasks.

Overall, our paper makes the following contributions:

a. We introduce an innovative approach to Few-Shot Continual Relation Extraction that leverages Large Language Models (LLMs) to generate comprehensive descriptions for each relation. These descriptions serve as stable class representations in the latent space during training. Unlike the variability and limitations of a limited set of samples from the memory buffer, these descriptions define the inherent meaning of the relations, offer a more reliable anchor, significantly reducing the risk of catastrophic forgetting. Importantly, LLMs are employed exclusively for generating descriptions and do not participate in the training or inference processes, ensuring that our method incurs minimal computational overhead.

b. We design a bi-encoder retrieval learning framework for both sample and class representation learning. In addition to sample representation contrastive learning, we integrate a description-pivot learning process, ensuring alignment of samples which maximize their respective class samples proximity, while non-matching samples are distanced.

c. Building on the enhanced representations, we introduce the *Descriptive Retrieval Inference* (DRI) strategy. In this approach, each sample "retrieves" the most fitting relation using a reciprocal rank fusion score that integrates both class descriptions and class prototypes, effectively finalizing the retrieval-based paradigm that underpins our method.

2 Background

2.1 Problem Formulation

In Few-Shot Continual Relation Extraction (FCRE), a model must continuously assimilate new knowledge from a sequential series of tasks. For each t-th task, the model undergoes training on the dataset $D^t = \{(x_i^t, y_i^t)\}_{i=1}^{N \times K}$. Here, N represents the number of relations in the task R^t, and K denotes the limited number of samples per relation, reflecting the few-shot learning scenario. Each sample (x, y) includes a sentence x containing a pair of entities (e_h, e_t) and a relation label $y \in R$. This type of task setup is referred to as *"N-way-K-shot"* (Chen, Wu, and Shi 2023a). Upon completion of task t, the dataset D^t should not be extensively included in subsequent learning, as continual learning aims to avoid retraining on all prior data. Ultimately, the model's performance is assessed on a test set which encompasses all encountered relations $\tilde{R}^T = \bigcup_{t=1}^{T} R^t$.

For clarity, each task in FCRE can be viewed as a conventional relation extraction problem, with the key challenge being the scarcity of samples available for learning. The primary goal of FCRE is to develop a model that can consistently acquire new knowledge from limited data while retaining competence in previously learned tasks. In the following subsections, we will explore the key aspects of FCRE models as addressed by state-of-the-art studies.

2.2 Encoding Latent Representation

A key initial consideration in Relation Extraction is how to *formalize the latent representation* of the input, as the output of a Transformer (Vaswani et al. 2017) is a matrix. In this work, we adopt a method recently introduced by Ma et al. (2024). Given an input sentence x, which includes a head entity e_h and a tail entity e_t, we reformulate it into a Cloze-style phrase $T(x)$ by incorporating a [MASK] token, which represents the relation between the entities. Specifically, the template is structured as follows:

$$\begin{aligned} T(x) = x\,[v_{0:n_0-1}]\,e_h\,[v_{n_0:n_1-1}]\,[\texttt{MASK}] \\ [v_{n_1:n_2-1}]\,e_t\,[v_{n_2:n_3-1}]\,. \end{aligned} \tag{1}$$

Each $[v_i]$ denotes a learnable continuous token, and n_j determines the number of tokens in each phrase. In our specific implementation, we use BERT's [UNUSED] tokens as $[v]$. The soft prompt phrase length is set to 3 tokens, meaning n_0, n_1, n_2 and n_3 correspond to the values of 3, 6, 9, and 12, respectively. We then forward the templated sentence $T(x)$ through BERT to encode it into a sequence of continuous vectors, from which we obtain the hidden representation $\boldsymbol{z}$ of the input, corresponding to the position of the [MASK] token:

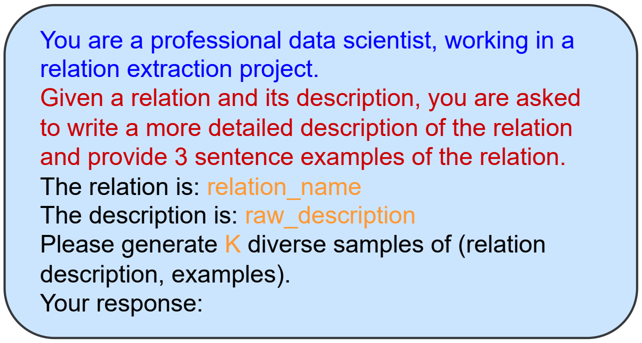
$$\boldsymbol{z} = [\mathcal{M} \circ T](x)[\text{position}(\texttt{[MASK]})], \tag{2}$$

where $\mathcal{M}$ denotes the backbone pre-trained language model. This latent representation is then passed through an MLP for prediction, enabling the model to learn which relation that best fills the [MASK] token.

2.3 Learning Latent Representation

In conventional Relation Extraction scenarios, a basic framework typically employs a backbone PLM followed by an MLP classifier to directly map the input space to the label space using Cross Entropy Loss. However, this approach proves inadequate in data-scarce settings (Snell, Swersky, and Zemel 2017). Consequently, training paradigms which directly target the latent space, such as contrastive learning, emerge as more suitable approaches. To enhance the semantic richness of the information extracted from the training samples, two popular losses are often utilized: *Supervised Contrastive Loss* and *Hard Soft Margin Triplet Loss*.

Supervised Contrastive Loss. To enhance the model's discriminative capability, we employ the Supervised Contrastive Loss (SCL) (Khosla et al. 2020). This loss function is designed to bring positive pairs of samples, which share the same class label, closer together in the latent space. Simultaneously, it pushes negative pairs, belonging to different classes, further apart. Let $\boldsymbol{z}_x$ represent the hidden vector



You are a professional data scientist, working in a relation extraction project.
Given a relation and its description, you are asked to write a more detailed description of the relation and provide 3 sentence examples of the relation.
The relation is: relation_name
The description is: raw_description
Please generate K diverse samples of (relation description, examples).
Your response:

Figure 2: Prompt to generate relation descriptions with LLMs.

output of sample x, the positive pairs $(\boldsymbol{z}_x, \boldsymbol{z}_p)$ are those who share a class, while the negative pairs $(\boldsymbol{z}_x, \boldsymbol{z}_n)$ correspond to different labels. The SCL is computed as follows:

$$\mathcal{L}_{\text{SC}}(x) = -\sum_{p \in P(x)} \log \frac{f(\boldsymbol{z}_x, \boldsymbol{z}_p)}{\sum_{u \in \mathcal{D} \backslash \{x\}} f(\boldsymbol{z}_x, \boldsymbol{z}_u)} \tag{3}$$

where $f(\mathbf{x}, \mathbf{y}) := \exp\left(\frac{\gamma(\mathbf{x}, \mathbf{y})}{\tau}\right)$, $\gamma(\cdot, \cdot)$ denotes the cosine similarity function, and τ is the temperature scaling hyperparameter. $P(x)$ and $\mathcal{D}$ denote the sets of positive samples with respect to sample x and the training set, respectively.

Hard Soft Margin Triplet Loss. To achieve a balance between flexibility and discrimination, the Hard Soft Margin Triplet Loss (HSMT) integrates both hard and soft margin triplet loss concepts (Hermans, Beyer, and Leibe 2017). This loss function is designed to maximize the separation between the most challenging positive and negative samples, while preserving a soft margin for improved flexibility. Formally, the loss is defined as:

$$\mathcal{L}_{\text{ST}}(x) =$$
$$-\log\left(1 + \max_{p \in P(\boldsymbol{x})} e^{\xi(\boldsymbol{z}_x, \boldsymbol{z}_p)} - \min_{n \in N(x)} e^{\xi(\boldsymbol{z}_x, \boldsymbol{z}_n)}\right), \tag{4}$$

where $\xi(\cdot, \cdot)$ denotes the Euclidean distance function. The objective of this loss is to ensure that the hardest positive sample is as distant as possible from the hardest negative sample, thereby enforcing a flexible yet effective margin.

During training, these two losses is aggregated and referred to as the *Sample-based learning loss*:

$$\mathcal{L}_{\text{Samp}} = \beta_{\text{SC}} \cdot \mathcal{L}_{\text{SC}} + \beta_{\text{ST}} \cdot \mathcal{L}_{\text{ST}} \tag{5}$$

3 Proposed Method

3.1 Label Descriptions

A core component of our method is achieving robust class latent representations, making class encoding crucial. To this end, having detailed definitions for each label, alongisde the hidden information extracted from the samples, is essential for our approach. In fact, the datasets used for benchmarking already provide each relation with a concise description, which we refer to as the *Raw description*. While leveraging these descriptions has shown promise in previous work (Luo et al. 2024), this approach remains limited due to its reliance on a one-to-one mapping between input embeddings and a single label description representation per task. This singular approach fails to offer rich, diverse, and robust information about the labels, leading to potential noise, instability, and suboptimal model performance.

To address these limitations, we employ Gemini 1.5 (Team et al. 2023; Reid et al. 2024) to generate K diverse, detailed, and illustrative descriptions for each relation. In particular, for each label, the respective raw description will be fed into the *LLM prompt*, serving as an expert-in-the-loop to guide the model. Our prompt template is depicted in Figure 2.

3.2 Description-pivot Learning

The single most valuable quality of class descriptions in our problem is that they are literal definitions of a relation, which makes them more accurate representations of that class than the underlying information from a set of samples. Thanks to this strength, they serve as stable knowledge anchors for the model to rehearse from, enabling effective reinforcement of old knowledge while assimilating new information. Unlike the variability of individual samples, a description remains consistent, providing a more reliable reference point for the model to rehearse from, effectively mitigating catastrophic forgetting.

To fully leverage this inherent advantage, we integrate these descriptions into the training process, framing the task as one of retrieving definition, which embodies real-world meaning, rather than a straightforward categorical classification. By doing so, we capitalize on the unchanging nature of descriptions, making them the focal point of our model's learning. Specifically, we incorporate two description-centric losses to enhance this retrieval-oriented approach:

$$\mathcal{L}_{\text{Des}} = \beta_{\text{HM}} \cdot \mathcal{L}_{\text{HM}} + \beta_{\text{MI}} \cdot \mathcal{L}_{\text{MI}}. \tag{6}$$

Here, $\mathcal{L}_{\text{HM}}$ and $\mathcal{L}_{\text{MI}}$ denote the Hard Margin Loss and the Mutual Information Loss, respectively. These losses are elaborated upon in the following paragraphs.

Hard Margin Loss. The Hard Margin Loss leverages label descriptions to refine the model's ability to distinguish between hard positive and hard negative pairs. Given the output hidden vectors $\{\boldsymbol{d}_x^k\}_{k=1,\ldots,K}$ from BERT corresponding to the label description of sample x, and $\boldsymbol{z}_p$ and $\boldsymbol{z}_n$ representing the hidden vectors of positive and negative samples respectively, the loss function is formulated to maximize the alignment between $\boldsymbol{d}_x^k$ and its corresponding positive sample, while enforcing a strict margin against negative samples.

Specifically, the loss is formulated as follows:

$$\mathcal{L}_{\text{HM}}(x) = \sum_{k=1}^{K} \mathcal{L}_{\text{HM}}^{k}(x), \tag{7}$$

$$\begin{aligned}\mathcal{L}_{\text{HM}}^{k}(x) = & \sum_{p \in P_{\text{H}}(x)} (1 - \gamma(\boldsymbol{d}_x^k, \boldsymbol{z}_p))^2 \\ & + \sum_{n \in N_{\text{H}}(x)} max(0, m - 1 + \gamma(\boldsymbol{d}_x^k, \boldsymbol{z}_n))^2, \end{aligned} \tag{8}$$

where m is a margin hyperparameter; $\gamma(\cdot, \cdot)$ denotes the cosine similarity function; $P_{\text{H}}(x)$ and $N_{\text{H}}(x)$ represent the sets of hard positive and hard negative samples, respectively. They are determined by comparing the similarity between $\boldsymbol{d}_x^k$ and both positive and negative pairs, specifically focusing on the most challenging pairs where the similarity to negative samples is close to or greater than that of positive samples, defined as follows:

$$\begin{aligned}P_{\text{H}}(x) = \{p \in P(x) | 1 - \gamma(\boldsymbol{d}_x^k, \boldsymbol{z}_p) \\ > min_{n \in N(x)}(1 - \gamma(\boldsymbol{d}_x^k, \boldsymbol{z}_n)), \forall k \in [K]\},\end{aligned} \tag{9}$$

$$\begin{aligned}N_{\text{H}}(x) = \{n \in N(x) | 1 - \gamma(\boldsymbol{d}_x^k, \boldsymbol{z}_n) \\ < max_{p \in P(x)}(1 - \gamma(\boldsymbol{d}_x^k, \boldsymbol{z}_p)), \forall k \in [K]\}.\end{aligned} \tag{10}$$

By utilizing the label description vectors $\{\boldsymbol{d}_x^k\}$, optimizing $\mathcal{L}_{\text{HM}}(x)$ effectively sharpens the model's decision boundary, reducing the risk of confusion between similar classes and improving overall performance in few-shot learning scenarios. The loss penalizes the model more heavily for misclassifications involving these hard samples, ensuring that the model pays particular attention to the most difficult cases, thereby enhancing its discriminative power.

Mutual Information Loss. The Mutual Information (MI) Loss is designed to maximize the mutual information between the input sample's hidden representation $\boldsymbol{z_x}$ of $\boldsymbol{x}$ and its corresponding retrieved descriptions, promoting a more informative alignment between them. Let $\boldsymbol{d}_n$ be a hidden vector of other label descriptions than $\boldsymbol{x}$. According to van den Oord, Li, and Vinyals (2018), the Mutual Information $MI(x)$ between the input embedding $\boldsymbol{z}_x$ and its corresponding label description follows the following inequation:

$$MI \geq \log B + \text{InfoNCE}(\{x_i\}_{i=1}^{B}; h), \tag{11}$$

where we have defined:

$$\text{InfoNCE}(\{x_i\}_{i=1}^{B}; h) = \frac{1}{B} \sum_{i=1}^{B} \log \frac{\sum_{k=1}^{K} h(\boldsymbol{z_i}, \boldsymbol{d}_i^k)}{\sum_{j=1}^{B} \sum_{k=1}^{K} h(\boldsymbol{z_j}, \boldsymbol{d}_j^k)}, \tag{12}$$

where $h(\boldsymbol{z}_j, \boldsymbol{d}_j^k) = \exp\left(\frac{\boldsymbol{z}_j^T W \boldsymbol{d}_j^k}{\tau}\right)$. Here, τ is the temperature, B is mini-batch size and W is a trainable parameter.

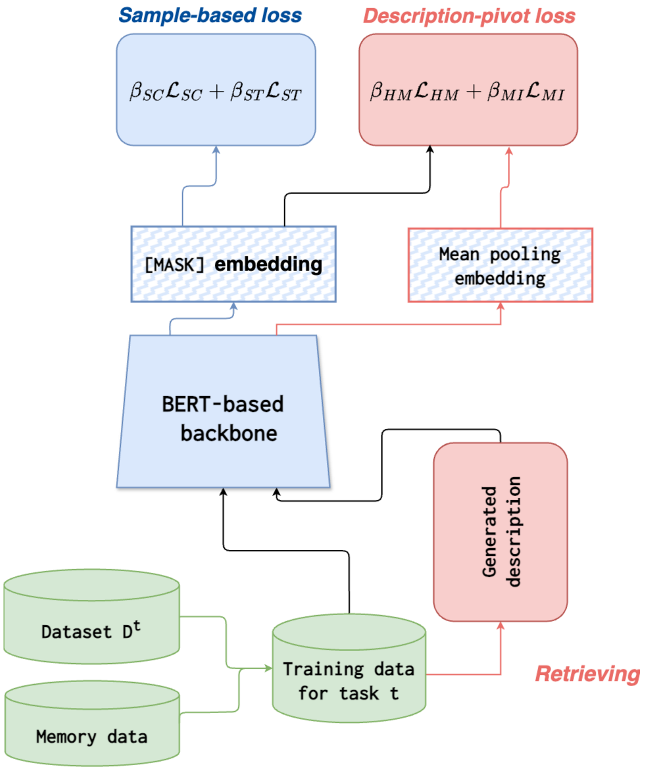


Figure 3: Our Framework.

Finally, the MI loss function in our implementation is:

$$\mathcal{L}_{\text{MI}}(x) = - \log \frac{\sum_{k=1}^{K} h(\boldsymbol{z}_x, \boldsymbol{d}_x^k)}{\sum_{k=1}^{K} h(\boldsymbol{z}_x, \boldsymbol{d}_x^k) + \sum_{n \in N(x)} \sum_{k=1}^{K} h(\boldsymbol{z}_x, \boldsymbol{d}_n^k)} \tag{13}$$

This loss ensures that the representation of the input sample is strongly associated with its corresponding label, while reducing its association with incorrect labels, thereby enhancing the discriminative power of the model.

Joint Training Objective Function. Our model is trained using a combination of the *Sample-based learning loss* mentioned in Section 2.3 and our description-pivot loss $\mathcal{L}_{\text{Des}}$, weighted by their respective coefficients:

$$\mathcal{L}(x) = \mathcal{L}_{\text{Samp}} + \mathcal{L}_{\text{Des}} \tag{14}$$

$$\begin{aligned} &= \beta_{\text{SC}} \cdot \mathcal{L}_{\text{SC}}(x) + \beta_{\text{ST}} \cdot \mathcal{L}_{\text{ST}}(x) \\ &\quad + \beta_{\text{HM}} \cdot \mathcal{L}_{\text{HM}}(x) + \beta_{\text{MI}} \cdot \mathcal{L}_{\text{MI}}(x), \end{aligned} \tag{15}$$

where β_{SC}, β_{ST}, β_{HM}, and β_{MI} are hyperparameters. This joint objective enables the model to leverage the strengths of each individual loss, facilitating robust and effective learning in Few-Shot Continual Relation Extraction tasks.

Training Procedure. Algorithm 1 outlines the end-to-end training process at each task $\mathcal{T}^j$, with Φ_{j-1} denoting the model after training on the previous $j-1$ tasks. In line with

FewRel *(10-way–5-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	$\mathcal{T}^6$	$\mathcal{T}^7$	$\mathcal{T}^8$	$\Delta\downarrow$
RP-CRE	$93.97_{\pm0.64}$	$76.05_{\pm2.36}$	$71.36_{\pm2.83}$	$69.32_{\pm3.98}$	$64.95_{\pm3.09}$	$61.99_{\pm2.09}$	$60.59_{\pm1.87}$	$59.57_{\pm1.13}$	34.40
CRL	$94.68_{\pm0.33}$	$80.73_{\pm2.91}$	$73.82_{\pm2.77}$	$70.26_{\pm3.18}$	$66.62_{\pm2.74}$	$63.28_{\pm2.49}$	$60.96_{\pm2.63}$	$59.27_{\pm1.32}$	35.41
CRECL	$93.93_{\pm0.22}$	$82.55_{\pm6.95}$	$74.13_{\pm3.59}$	$69.33_{\pm3.87}$	$66.51_{\pm4.05}$	$64.60_{\pm1.92}$	$62.97_{\pm1.46}$	$59.99_{\pm0.65}$	33.94
ERDA	$92.43_{\pm0.32}$	$64.52_{\pm2.11}$	$50.31_{\pm3.32}$	$44.92_{\pm3.77}$	$39.75_{\pm3.34}$	$36.36_{\pm3.12}$	$34.34_{\pm1.83}$	$31.96_{\pm1.91}$	60.47
SCKD	$94.77_{\pm0.35}$	$82.83_{\pm2.61}$	$76.21_{\pm1.61}$	$72.19_{\pm1.33}$	$70.61_{\pm2.24}$	$67.15_{\pm1.96}$	$64.86_{\pm1.35}$	$62.98_{\pm0.88}$	31.79
ConPL**	$\mathbf{95.18}_{\pm\mathbf{0.73}}$	$79.63_{\pm1.27}$	$74.54_{\pm1.13}$	$71.27_{\pm0.85}$	$68.35_{\pm0.86}$	$63.86_{\pm2.03}$	$64.74_{\pm1.39}$	$62.46_{\pm1.54}$	32.72
CPL	94.87	85.14	78.80	75.10	72.57	69.57	66.85	64.50	30.37
CPL + MI	$94.69_{\pm0.7}$	$\mathbf{85.58}_{\pm\mathbf{1.88}}$	$\mathbf{80.12}_{\pm\mathbf{2.45}}$	$\underline{75.71_{\pm2.28}}$	$\underline{73.90_{\pm1.8}}$	$\underline{70.72_{\pm0.91}}$	$\underline{68.42_{\pm1.77}}$	$\underline{66.27_{\pm1.58}}$	28.42
DCRE	$\underline{94.93_{\pm0.39}}$	$\underline{85.14_{\pm2.27}}$	$\underline{79.06_{\pm1.68}}$	$\mathbf{75.92}_{\pm\mathbf{2.03}}$	$\mathbf{74.10}_{\pm\mathbf{2.53}}$	$\mathbf{71.83}_{\pm\mathbf{2.17}}$	$\mathbf{69.84}_{\pm\mathbf{1.48}}$	$\mathbf{68.24}_{\pm\mathbf{0.79}}$	**26.69**

TACRED *(5-way-5-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	$\mathcal{T}^6$	$\mathcal{T}^7$	$\mathcal{T}^8$	$\Delta\downarrow$
RP-CRE	$87.32_{\pm1.76}$	$74.90_{\pm6.13}$	$67.88_{\pm4.31}$	$60.02_{\pm5.37}$	$53.26_{\pm4.67}$	$50.72_{\pm7.62}$	$46.21_{\pm5.29}$	$44.48_{\pm3.74}$	42.84
CRL	$88.32_{\pm1.26}$	$76.30_{\pm7.48}$	$69.76_{\pm5.89}$	$61.93_{\pm2.55}$	$54.68_{\pm3.12}$	$50.92_{\pm4.45}$	$47.00_{\pm3.78}$	$44.27_{\pm2.51}$	44.05
CRECL	$87.09_{\pm2.50}$	$78.09_{\pm5.74}$	$61.93_{\pm4.89}$	$55.60_{\pm5.78}$	$53.42_{\pm2.99}$	$51.91_{\pm2.95}$	$47.55_{\pm3.38}$	$45.53_{\pm1.96}$	41.56
ERDA	$81.88_{\pm1.97}$	$53.68_{\pm6.31}$	$40.36_{\pm3.35}$	$36.17_{\pm3.65}$	$30.14_{\pm3.96}$	$22.61_{\pm3.13}$	$22.29_{\pm1.32}$	$19.42_{\pm2.31}$	62.46
SCKD	$\underline{88.42_{\pm0.83}}$	$79.35_{\pm4.13}$	$70.61_{\pm3.16}$	$66.78_{\pm4.29}$	$60.47_{\pm3.05}$	$58.05_{\pm3.84}$	$54.41_{\pm3.47}$	$52.11_{\pm3.15}$	36.31
ConPL**	$\mathbf{88.77}_{\pm\mathbf{0.84}}$	$69.64_{\pm1.93}$	$57.50_{\pm2.48}$	$52.15_{\pm1.59}$	$58.19_{\pm2.31}$	$55.01_{\pm3.12}$	$52.88_{\pm3.66}$	$50.97_{\pm3.41}$	37.80
CPL	86.27	$\underline{81.55}$	73.52	$\underline{68.96}$	63.96	62.66	59.96	57.39	28.88
CPL + MI	$85.67_{\pm0.8}$	$\underline{82.54_{\pm2.98}}$	$\underline{75.12_{\pm3.67}}$	$\underline{70.65_{\pm2.75}}$	$\underline{66.79_{\pm2.18}}$	$\underline{65.17_{\pm2.48}}$	$\underline{61.25_{\pm1.52}}$	$\underline{59.48_{\pm3.53}}$	26.19
DCRE	$86.20_{\pm1.35}$	$\mathbf{83.18}_{\pm\mathbf{8.04}}$	$\mathbf{80.65}_{\pm\mathbf{3.06}}$	$\mathbf{75.05}_{\pm\mathbf{3.07}}$	$\mathbf{68.83}_{\pm\mathbf{5.05}}$	$\mathbf{68.30}_{\pm\mathbf{4.28}}$	$\mathbf{65.30}_{\pm\mathbf{2.74}}$	$\mathbf{63.21}_{\pm\mathbf{2.39}}$	**22.99**

Table 1: Accuracy (%) of methods using BERT-based backbone after training for each task. The best results are in **bold**. **Results of ConPL are reproduced

memory-based continual learning methods, we maintain a memory buffer $\tilde{M}_{j-1}$ that stores a few representative samples from all previous tasks $\mathcal{T}^1, \ldots, \mathcal{T}^{j-1}$, along with a relation description set $\tilde{E}_{j-1}$ that holds the descriptions of all previously encountered relations.

1. **Initialization** (Line 1–2): The model for the current task, Φ_j, is initialized with the parameters of Φ_{j-1}. We update the relation description set $\tilde{E}_j$ by incorporating new relation descriptions from E_j.
2. **Training on the Current Task** (Line 3): We train Φ_j on D_j to learn the novel relations introduced in in $\mathcal{T}^j$.
3. **Memory Update** (Lines 4–8): We select L representative samples from D_j for each relation $r \in R_j$. These are the L samples whose latent representations are closest to the 1-means centroid of all class samples. These samples constitute the memory M_r, leading to an updated overall memory $\tilde{M}_j = \tilde{M}_{j-1} \cup M_j$ and an updated relation set $\tilde{R}_j = \tilde{R}_{j-1} \cup R_j$.
4. **Prototype Storing** (Line 9): A prototype set $\tilde{P}_j$ is generated based on the updated memory $\tilde{M}_j$. We generate a prototype set $\tilde{P}_j$ based on the updated memory $\tilde{M}_j$.
5. **Memory Training** (Line 10): We refine Φ_j by training on the augmented memory set $\tilde{M}_j^*$, ensuring that the model preserves knowledge of relations from previous tasks.

Algorithm 1: Training procedure at each task $\mathcal{T}^j$

```
Input: $\Phi_{j-1}, \tilde{R}_{j-1}, \tilde{M}_{j-1}, \tilde{K}_{j-1}, D_j, R_j, K_j$.
Output: $\Phi_j, \tilde{M}_j, \tilde{K}_j, \tilde{P}_j$.
 1: Initialize $\Phi_j$ from $\Phi_{j-1}$
 2: $\tilde{K}_j \leftarrow \tilde{K}_{j-1} \cup K_j$
 3: Update $\Phi_j$ by L on $D_j$ (train on current task)
 4: $\tilde{M}_j \leftarrow \tilde{M}_{j-1}$
 5: for each $r \in R_j$ do
 6:     pick $L$ samples in $D_j$ and add them into $\tilde{M}_j$
 7: end for
 8: $\tilde{R}_j \leftarrow \tilde{R}_{j-1} \cup R_j$
 9: Update $\tilde{P}_j$ with new data in $D_j$ (for inference)
10: Update $\Phi_j$ by $\mathcal{L}$ on $\tilde{M}_j$ and $D_j^*$ (train on memory)
```

3.3 Descriptive Retrieval Inference

Traditional methods such as Nearest Class Mean (NCM) (Ma et al. 2024) predict relations by selecting the class whose prototype has the smallest distance to the test sample x. While effective, this approach relies solely on distance metrics, which may not fully capture the nuanced relationships between a sample and the broader context provided by class descriptions.

Rather than merely seeking the closest prototype, we aim to retrieve the class description that best aligns with the input, thereby leveraging the inherent semantic meaning of the label. To achieve this, we introduce *Descriptive Retrieval In-*

FewRel *(10-way–5-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	$\mathcal{T}^6$	$\mathcal{T}^7$	$\mathcal{T}^8$	Δ ↓
CPL	$\mathbf{97.25}_{\pm\mathbf{0.30}}$	$\mathbf{89.29}_{\pm\mathbf{2.51}}$	$85.56_{\pm 1.21}$	$82.10_{\pm 2.02}$	$79.96_{\pm 2.72}$	$78.41_{\pm 3.22}$	$76.42_{\pm 2.25}$	$75.20_{\pm 2.33}$	22.05
DCRE	$96.92_{\pm 0.16}$	$88.95_{\pm 1.72}$	$\mathbf{87.12}_{\pm\mathbf{1.52}}$	$\mathbf{85.44}_{\pm\mathbf{1.91}}$	$\mathbf{84.89}_{\pm\mathbf{2.12}}$	$\mathbf{83.52}_{\pm\mathbf{1.46}}$	$\mathbf{81.64}_{\pm\mathbf{0.69}}$	$\mathbf{80.34}_{\pm\mathbf{0.55}}$	**16.58**

TACRED *(5-way-5-shot)*

Method	$\mathcal{T}^1$	$\mathcal{T}^2$	$\mathcal{T}^3$	$\mathcal{T}^4$	$\mathcal{T}^5$	$\mathcal{T}^6$	$\mathcal{T}^7$	$\mathcal{T}^8$	Δ ↓
CPL	$88.74_{\pm 0.44}$	$85.16_{\pm 5.38}$	$78.35_{\pm 4.46}$	$77.50_{\pm 4.04}$	$76.01_{\pm 5.04}$	$76.30_{\pm 4.41}$	$74.51_{\pm 5.06}$	$73.83_{\pm 4.91}$	14.91
DCRE	$\mathbf{89.06}_{\pm\mathbf{0.59}}$	$\mathbf{87.41}_{\pm\mathbf{5.54}}$	$\mathbf{84.91}_{\pm\mathbf{3.38}}$	$\mathbf{84.18}_{\pm\mathbf{2.44}}$	$\mathbf{82.74}_{\pm\mathbf{3.64}}$	$\mathbf{81.92}_{\pm\mathbf{2.33}}$	$\mathbf{79.34}_{\pm\mathbf{2.89}}$	$\mathbf{79.10}_{\pm\mathbf{2.37}}$	**9.96**

Table 2: Accuracy (%) of methods using LLM2Vec-based backbone after training for each task. The best results are in **bold**.

ference (DRI), a retrieval mechanism fusing two distinct reciprocal ranking scores. This approach not only considers the proximity of a sample to class prototypes but also incorporates cosine similarity measures between the sample's hidden representation $\boldsymbol{z}$ and relation descriptions generated by an LLM. This dual focus on both spatial and semantic alignment ensures that the final prediction is informed by a richer, more robust understanding of the relations.

Given a sample x with hidden representation $\boldsymbol{z}$ and a set of relation prototypes $\{\boldsymbol{p}_r\}_{r=1}^n$, the inference process begins by calculating the negative Euclidean distance between $\boldsymbol{z}$ and each prototype $\boldsymbol{p}_r$:

$$\mathbf{E}(x, r) = -\|\boldsymbol{z} - \boldsymbol{p}_r\|_2, \tag{16}$$

$$\boldsymbol{p}_r = \frac{1}{L}\sum_{i=1}^{L} \boldsymbol{z}_i, \tag{17}$$

where L is the memory size per relation. Simultaneously, we compute the cosine similarity between the hidden representation and each relation description prototype, $\gamma(\boldsymbol{z}, \boldsymbol{d}_r)$. These two scores are combined into DRI score of sample $\boldsymbol{x}$ w.r.t relation r for inference, ensuring that predictions align with both label prototypes and relation descriptions:

$$\text{DRI}(x, r) = \frac{\alpha}{\epsilon + \text{rank}(\mathbf{E}(x, r))} + \frac{1 - \alpha}{\epsilon + \text{rank}(\gamma(\boldsymbol{z}, \boldsymbol{d}_r))}, \tag{18}$$

where $d_r = \frac{1}{K}\sum_{i=1}^{K} d_r^i$, rank$(\cdot)$ represents the rank position of the score among all relations. The α hyperparameter balances the contributions of the Euclidean distance-based score and the cosine similarity score in the final ranking for inference, and ϵ is a hyperparameter that controls the influence of lower-ranked relations in the final prediction. By adjusting ϵ, we can fine-tune the model's sensitivity to less prominent relations. Finally, the predicted relation label y^* is predicted as the one corresponding to the highest DRI score:

$$y_x^* = \underset{r=1,\ldots,n}{\text{argmax}}\, \text{DRI}(x, r) \tag{19}$$

This fusion approach for inference complements the learning paradigm, ensuring consistency and reliability throughout the FCRE process. By effectively balancing the strengths of protoype-based proximity and description-based semantic similarity, it leads to more accurate and robust predictions across sequential tasks.

4 Experiments

4.1 Settings

We conduct experiments using two pre-trained language models, BERT (Devlin et al. 2019) and LLM2Vec (BehnamGhader et al. 2024), on two widely used benchmark datasets for Relation Extraction: FewRel (Han et al. 2018) and TACRED (Zhang et al. 2017). We benchmark our methods against state-of-the-art baselines: **SCKD** (Wang, Wang, and Hu 2023), **RP-CRE** (Cui et al. 2021), **CRL** (Zhao et al. 2022), **CRECL** (Hu et al. 2022), **ERDA** (Qin and Joty 2022), **ConPL** (Chen, Wu, and Shi 2023b), **CPL** (Ma et al. 2024), **CPL+MI** (Tran et al. 2024c).

4.2 Experiment results

Our proposed method yields state-of-the-art accuracy. Table 1 presents the results of our method and the baselines, all using the same pre-trained BERT-based backbone. Our method consistently outperforms all baselines across the board. The performance gap between our method and the strongest baseline, CPL, reaches up to 3.74% on FewRel and 5.82% on TACRED.

To further validate our model, we tested it on LLM2Vec, which provides stronger representation learning than BERT. As shown in Table 2, our model again surpasses CPL, with accuracy drops of only 16.58% on FewRel and 9.96% on TACRED.

These results highlight the effectiveness of our method in leveraging semantic information from descriptions, which helps mitigate forgetting and overfitting, ultimately leading to significant performance improvements.

Exploiting additional descriptions significantly enhances representation learning. Figure 4 presents t-SNE visualizations of the latent space of relations without (left) and with (right) the use of descriptions during training. The visualizations reveal that incorporating descriptions markedly improves the quality of the model's representation learning. For instance, the brown-orange and purple-green class pairs, which are closely clustered and prone to misclassification in the left image, are more distinctly separated in the right image. Additionally, Figure 5 illustrates that our strategy, which leverages refined descriptions, captures more semantic knowledge related to the labels than the approach using raw descriptions. This advantage bridges the gap imposed by the challenges of few-shot continual learning scenarios,

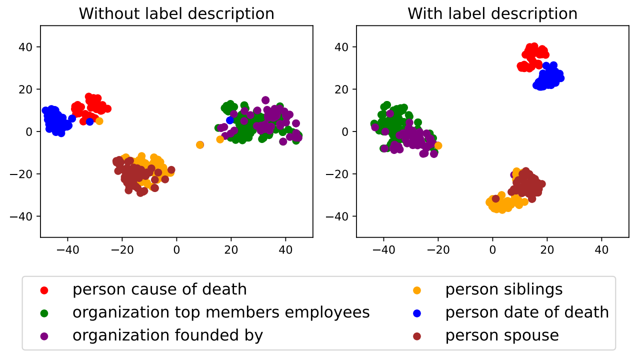


Figure 4: t-SNE visualization of the representations of 6 relations post-training, with and without descriptions, using our retrieval strategy.

leading to superior performance. Figure 6 shows the perfomance of our model on TACRED as the number of generated expert descriptions per training varies. The results indicate that the model performance generally improves from $K = 3$ and peaks at $K = 7$.

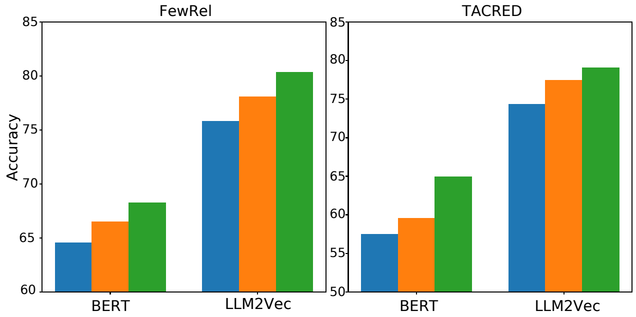


Figure 5: The impact of refined descriptions generated by LLMs. The green, orange, and blue bars show respectively the final accuracies of DCRE when using refined descriptions, original descriptions, and without using descriptions.

Our retrieval-based prediction strategy notably enhances model performance. Table 3 demonstrates that by leveraging the rich information from generated descriptions, our proposed strategy improves the model's performance by up to 1.31% on FewRel and 6.66% on TACRED compared to traditional NCM-based classification. The harmonious integration of NCM-based prototype proximity and description-based semantic similarity enables our strategy to deliver more accurate and robust predictions across sequential tasks.

4.3 Ablation study

Table 8 present evaluation results that closely examine the role of each component in the objective function during training. The findings underscore the critical importance of $\mathcal{L}_{\text{MI}}$ and $\mathcal{L}_{\text{HM}}$, both of which leverage instructive descriptions from LLMs, aided by *Raw descriptions*. Because when we ablate one of them, the final accuracy can be reduced by 6% on the BERT-based model, and 10% on the LLM2VEC-based model.

Method	FewRel		TACRED	
	BERT	LLM2Vec	BERT	LLM2Vec
NCM	66.93	79.26	58.26	75.00
DRI (Ours)	**68.24**	**80.34**	**63.21**	**79.10**

Table 3: DRI and NCM prediction.

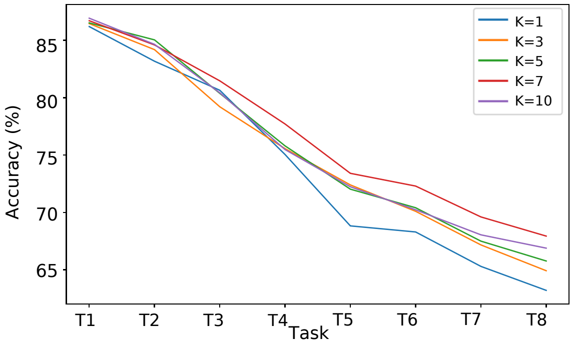


Figure 6: Model performance when varying K, on *TACRED 5-way 5-shot*.

Method	BERT		LLM2Vec	
	FewRel	TACRED	FewRel	TACRED
DCRE (Our)	**68.24**	**63.21**	**80.34**	**79.10**
w/o $\mathcal{L}_{\text{SC}}$	67.58	62.11	78.39	77.01
w/o $\mathcal{L}_{\text{MI}}$	65.10	57.23	70.61	74.17
w/o $\mathcal{L}_{\text{HM}}$	66.20	62.46	77.22	74.75
w/o $\mathcal{L}_{\text{ST}}$	67.54	59.56	77.48	73.77

Table 4: Ablation study.

5 Conclusion

In this work, we propose a novel retrieval-based approach to address the challenging problem of Few-shot Continual Relation Extraction. By leveraging large language models to generate rich relation descriptions, our bi-encoder training paradigm enhances both sample and class representations and also enables a robust retrieval-based prediction method that maintains performance across sequential tasks. Extensive experiments demonstrate the effectiveness of our approach in advancing the state-of-the-art and overcoming the limitations of traditional memory-based techniques, underscoring the potential of language models and retrieval techniques for dynamic real-world relationship identification.

References

BehnamGhader, P.; Adlakha, V.; Mosbach, M.; Bahdanau, D.; Chapados, N.; and Reddy, S. 2024. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. *arXiv preprint arXiv:2404.05961*.

Chen, X.; Wu, H.; and Shi, X. 2023a. Consistent Prototype Learning for Few-Shot Continual Relation Extraction. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7409–7422. Toronto, Canada: Association for Computational Linguistics.

Chen, X.; Wu, H.; and Shi, X. 2023b. Consistent Prototype Learning for Few-Shot Continual Relation Extraction. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, 7409–7422. Association for Computational Linguistics.

Cui, L.; Yang, D.; Yu, J.; Hu, C.; Cheng, J.; Yi, J.; and Xiao, Y. 2021. Refining Sample Embeddings with Relation Prototypes to Enhance Continual Relation Extraction. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 232–243. Online: Association for Computational Linguistics.

Dao, V.; Pham, V.-C.; Tran, Q.; Le, T.-T.; Ngo, L.; and Nguyen, T. 2024. Lifelong Event Detection via Optimal Transport. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 12610–12621.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Hai, N. L.; Nguyen, T.; Van, L. N.; Nguyen, T. H.; and Than, K. 2024. Continual variational dropout: a view of auxiliary local variables in continual learning. *Machine Learning*, 113(1): 281–323.

Han, X.; Zhu, H.; Yu, P.; Wang, Z.; Yao, Y.; Liu, Z.; and Sun, M. 2018. FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4803–4809. Brussels, Belgium: Association for Computational Linguistics.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In Defense of the Triplet Loss for Person Re-Identification. arXiv:1703.07737.

Hu, C.; Yang, D.; Jin, H.; Chen, Z.; and Xiao, Y. 2022. Improving Continual Relation Extraction through Prototypical Contrastive Learning. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., *Proceedings of the 29th International Conference on Computational Linguistics*, 1885–1895. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 18661–18673. Curran Associates, Inc.

Le, M.; Luu, T. N.; The, A. N.; Le, T.-T.; Nguyen, T.; Nguyen, T. T.; Van, L. N.; and Nguyen, T. H. 2025. Adaptive Prompting for Continual Relation Extraction: A Within-Task Variance Perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Le, M.; Nguyen, A.; Nguyen, H.; Nguyen, T.; Pham, T.; Ngo, L. V.; and Ho, N. 2024a. Mixture of Experts Meets Prompt-Based Continual Learning. In *Advances in Neural Information Processing Systems*.

Le, T.-T.; Dao, V.; Nguyen, L.; Nguyen, T.-N.; Ngo, L.; and Nguyen, T. 2024b. SharpSeq: Empowering Continual Event Detection through Sharpness-Aware Sequential-task Learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3632–3644.

Le, T.-T.; Nguyen, M.; Nguyen, T. T.; Van, L. N.; and Nguyen, T. H. 2024c. Continual relation extraction via sequential multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18444–18452.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.

Luo, D.; Gan, Y.; Hou, R.; Lin, R.; Liu, Q.; Cai, Y.; and Gao, W. 2024. Synergistic Anchored Contrastive Pre-training for Few-Shot Relation Extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17): 18742–18750.

Ma, S.; Han, J.; Liang, Y.; and Cheng, B. 2024. Making Pre-trained Language Models Better Continual Few-Shot Relation Extractors. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 10970–10983. Torino, Italia: ELRA and ICCL.

Nguyen, H.; Nguyen, C.; Ngo, L.; Luu, A.; and Nguyen, T. 2023. A spectral viewpoint on continual relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 9621–9629.

Phan, H.; Tuan, A. P.; Nguyen, S.; Linh, N. V.; and Than, K. 2022. Reducing catastrophic forgetting in neural networks via gaussian mixture approximation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 106–117. Springer.

Qin, C.; and Joty, S. 2022. Continual Few-shot Relation Learning via Embedding Space Regularization and Data Augmentation. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2776–2789. Dublin, Ireland: Association for Computational Linguistics.

Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillicrap, T.; Alayrac, J.-b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Shin, H.; Lee, J. K.; Kim, J.; and Kim, J. 2017. Continual Learning with Deep Generative Replay. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical Networks for Few-shot Learning. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Tran, Q.; Phan, H.; Le, M.; Truong, T.; Phung, D.; Ngo, L.; Nguyen, T.; Ho, N.; and Le, T. 2024a. Leveraging Hierarchical Taxonomies in Prompt-based Continual Learning. arXiv:2410.04327.

Tran, Q.; Phan, H.; Tran, L.; Than, K.; Tran, T.; Phung, D.; and Le, T. 2024b. KOPPA: Improving Prompt-based Continual Learning with Key-Query Orthogonal Projection and Prototype-based One-Versus-All. arXiv:2311.15414.

Tran, Q.; Thanh, N.; Anh, N.; Hai, N.; Le, T.; Ngo, L.; and Nguyen, T. 2024c. Preserving Generalization of Language models in Few-shot Continual Relation Extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 13771–13784.

Van, L. N.; Hai, N. L.; Pham, H.; and Than, K. 2022. Auxiliary local variables for improving regularization/prior approach in continual learning. In *Pacific-Asia conference on knowledge discovery and data mining*, 16–28. Springer.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR*, abs/1807.03748.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, X.; Wang, Z.; and Hu, W. 2023. Serial Contrastive Knowledge Distillation for Continual Few-shot Relation Extraction. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 12693–12706. Toronto, Canada: Association for Computational Linguistics.

Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; and Manning, C. D. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 35–45. Copenhagen, Denmark: Association for Computational Linguistics.

Zhao, K.; Xu, H.; Yang, J.; and Gao, K. 2022. Consistent Representation Learning for Continual Relation Extraction. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 3402–3411. Dublin, Ireland: Association for Computational Linguistics.

A Appendix

B Experimental Details

B.1 Datasets

Our experiments utilize the following two benchmarks:

- **FewRel** (Han et al. 2018) includes 100 relations with 70,000 samples. Following Qin and Joty (2022), we employ a setup with 80 relations, partitioned into 8 tasks, each comprising 10 relations *(10-way)*. Task $\mathcal{T}^1$ includes 100 samples per relation, whereas the remaining tasks are characterized as few-shot tasks conducted under *5-shot* settings.
- **TACRED** (Zhang et al. 2017) encompasses 42 relations with 106,264 samples extracted from Newswire and Web documents. Consistent with the approach outlined by Qin and Joty (2022), we exclude instances labeled as ”no_relation” and allocate the remaining 41 relations across 8 tasks. Task $\mathcal{T}^1$ comprises 6 relations, each with 100 samples, while each subsequent tasks involve 5 relations *(5-way)* in *5-shot* setups.

B.2 Baselines

In this section, we briefly describe some state-of-the-art methods in FCRE that appear as benchmarking baselines in our evaluations, including:

- **SCKD** (Wang, Wang, and Hu 2023) adopts a systematic strategy for knowledge distillation, which aims to preserve old knowledge from previous tasks. Besides, this method employs contrastive learning techniques with pseudo samples to enhance the distinguishability between representations of different relations.
- **CPL** (Ma et al. 2024) proposes a Contrastive Prompt Learning framework, which designs prompts to generalize across categories and uses margin-based contrastive

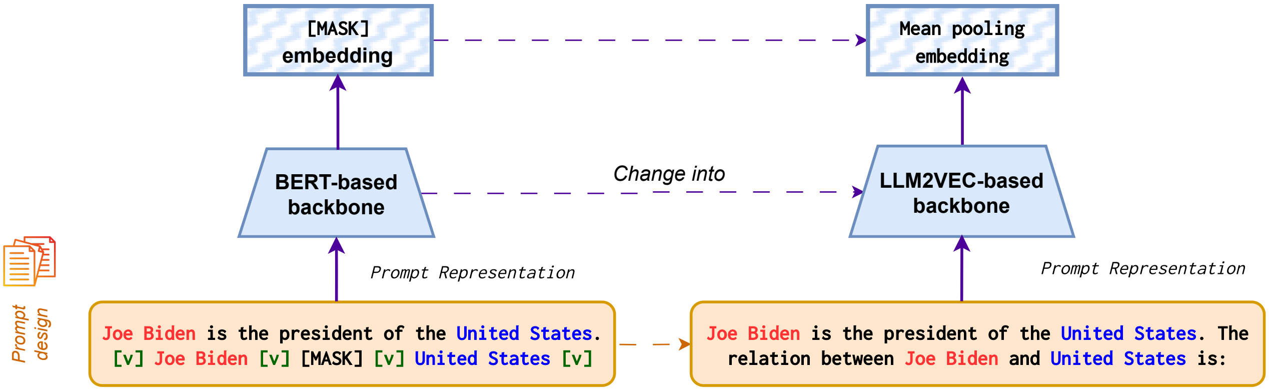

[MASK]
embedding
Mean pooling
embedding
BERT-based
backbone
Change into
LLM2VEC-based
backbone
Prompt Representation
Prompt Representation
Prompt
design
Joe Biden is the president of the United States.
[v] Joe Biden [v] [MASK] [v] United States [v]
Joe Biden is the president of the United States. The
relation between Joe Biden and United States is:

Figure 7: Adapting LLM2Vec for ours architecture

learning to handle hard samples, thus reducing catastrophic forgetting and overfitting. Besides, the authors employ a memory augmentation strategy to generate diverse samples with ChatGPT, further mitigating overfitting in low-resource scenarios of FCRE.

- **RP-CRE** (Cui et al. 2021): The approach tackles Continual Relation Extraction (CRE) by capitalizing on memorized samples to mitigate the forgetting of previous relations. It employs K-means clustering to identify prototypes that represent each relation based on stored samples. These prototypes are subsequently utilized to refine embeddings of subsequent samples, enabling the model to preserve knowledge of past relations while learning new ones. This methodology enhances memory utilization compared to previous CRE models, resulting in improved performance.
- **CRL** (Zhao et al. 2022): addresses the issue of catastrophic forgetting by adopting a consistent representation learning strategy. It emphasizes the preservation of stable relation embeddings via contrastive learning and knowledge distillation during the replay of memorized samples. The method involves supervised contrastive learning on a memory bank specific to each new task, followed by contrastive replay of memory samples and knowledge distillation to retain historical relation knowledge. Through this approach, effective alleviation of forgetting is achieved via consistent representation learning.
- **CRECL** (Hu et al. 2022): extends beyond conventional few-shot learning by imposing additional constraints on training data. It accomplishes this by integrating information regarding support instances to augment instance representations. Furthermore, it advocates for open-source task enrichment to facilitate cross-domain knowledge aggregation and introduces the TinyRel-CM dataset tailored specifically for few-shot relation classification with restricted training data. Experimental results illustrate its efficacy in enhancing performance under conditions of limited data availability.
- **ERDA** (Qin and Joty 2022): This study introduces Continual Few-Shot Relation Learning (CFRL) as a novel challenge, recognizing the constraints of current methodologies that demand substantial labeled data for new tasks. CFRL endeavors to acquire knowledge of novel relations with minimal data while averting catastrophic forgetting. Addressing this challenge, ERDA presents a methodology grounded in embedding space regularization and data augmentation. This strategy imposes constraints on relational embeddings and integrates supplementary relevant data through self-supervision. Extensive experimentation showcases ERDA's substantial performance enhancements over prior state-of-the-art approaches in CFRL scenarios.
- **ConPL** (Chen, Wu, and Shi 2023b) introduces a method comprising three core components: a prototype-based classification module, a memory-enhanced module, and a consistent learning module designed to maintain distribution consistency and mitigate forgetting. Furthermore, ConPL employs prompt learning to enhance representation learning and integrates focal loss to reduce confusion among closely related classes.
- **MI** (Tran et al. 2024c) introduces a novel framework leveraging often-discarded language model heads to preserve prior knowledge from pre-trained backbones. By employing a mutual information maximization strategy, this method aligns the primary classification head with retained backbone knowledge, enhancing model performance.

It is noteworthy that we reproduce the results of ConPL (Chen, Wu, and Shi 2023a) under the same settings as SCKD and CPL. This is because the evaluation strategy in the original paper is impractical for continual learning scenarios.

B.3 Architecture

- For BERT-based models: We use BERT-base-uncased checkpoint[1] on Hugging Face.
- For LLM2Vec-based models: We use the Meta-Llama-3-8B-Instruct-mntp-supervised [2] checkpoint on Hugging

[1] https://huggingface.co/bert-base-uncased

[2] https://huggingface.co/McGill-NLP/LLM2Vec-Meta-Llama-3-8B-Instruct-mntp-supervised

Face.

In addition, Figure 7 depicts the difference in architecture and input design when using the BERT-based backbone and the LLM2Vec-based backbone. Unlike BERT, which uses the ”MASK” token during pretraining, the LLM2Vec model does not. As suggested by BehnamGhader et al. (2024), the mean pooling strategy yields the best performance for LLM2Vec. Therefore, we replace the hybrid prompt with a hard prompt and adopt the mean pooling strategy to obtain the input representation.

B.4 Evaluation Protocol and Training Configurations

For each reported result, we conduct 6 independent runs with different random seeds and report the mean and the corresponding standard deviation.

Evaluation Metric

- **Final average accuracy:** We use final average accuracy to evaluate methods in our experiments. The average accuracy after training task T_j is calculated as follows:

$$ACC_j = \frac{1}{j}\sum_{i=1}^{j} ACC_{j,i}$$

  where $ACC_{j,i}$ is the accuracy on the test set of task T_i after training the model on task T_j.

- **Accuracy drop:** indicates the decrease in average accuracy after training all T tasks on each benchmark.

$$\Delta = ACC_T - ACC_1$$

Training Configuration Details of hyperparameter search:

- Learning rate: $\{\mathbf{1 \times 10^{-5}}, 2 \times 10^{-5}, 1 \times 10^{-4}\}$
- α: {**0.4**, 0.5}
- β_{SC}: {0.5, **1.0**, 1.5, 2.0, 2.5}
- β_{MI}: {0.5, **1.0**, 1.5, **2.0**, 2.5}
- β_{HM}: {**0.5**, **1.0**, 1.5, 2.0, 2.5}
- β_{ST}: {0.5, **1.0**, 1.5, 2.0, 2.5}

Additionally, Tables 5 and 6 provide the optimal values of hyperparameters for each model backbone.

B.5 Prompt Template

Table 9 illustrates a prompt used to generate label descriptions for a relation called *”place served by transport hub”* and its respective output, during training in our strategy.

C Additional experimental results

C.1 Effect of the number of generated descriptions for each relation

Table 7 reports the final accuracy of BERT-based models on two benchmarks when varying the number of generated descriptions K. The results show that using multiple generated descriptions is better than using just one, as many generated samples help provide more specific, diverse, and semantically rich information, thereby making representation learning more comprehensive and effective. However, if K is too large, our observations indicate that LLMs like Gemini can generate biased and low-quality samples, which negatively affect model performance. In particular, the best accuracy is achieved when $K = 3$ on FewRel and $K = 7$ on TACRED.

Moreover, we provide an ablation study demonstrating the importance of each component in the objective function when using the optimal value of K. The results depict that on both datasets, $\mathcal{L}_{MI}$, designed to maximize the mutual information between the input samples' hidden representations and their corresponding retrieved descriptions, plays the most vital role. Its absence can cause the model to lose 4-5% in final accuracy.

Hyperparameter	Value
Current-task training epochs	10
Rehearsal training epochs	10
Learning rate	1×10^{-5}
ϵ	60
α	0.4
Encoder output size	768
BERT input max length	256
Margin m for Hard Margin Loss	0.5
β_{SC}	1.0
β_{MI} (FewRel)	1.0
β_{MI} (TACRED)	2.0
β_{HM} (FewRel)	1.0
β_{HM} (TACRED)	0.5
β_{ST}	1.0
Soft prompt initialization	Random
Soft prompt phrase length	3
Soft prompt number of phrases	4

Table 5: Hyperparameters for the BERT-backbone setting

Hyperparameter	Value
Encoder output size	4096
Current-task training epochs	10
Rehearsal training epochs	10
Learning rate	1×10^{-5}
ϵ	60
α	0.4
Margin m for Hard Margin Loss	0.5
Lora alpha	16
Lora rank	8
Lora dropout	0.05
β_{SC}	1.0
β_{MI} (FewRel)	1.0
β_{MI} (TACRED)	2.0
β_{HM} (FewRel)	1.0
β_{HM} (TACRED)	0.5
β_{ST}	1.0

Table 6: Hyperparameters setting with LLM2vec backbone.

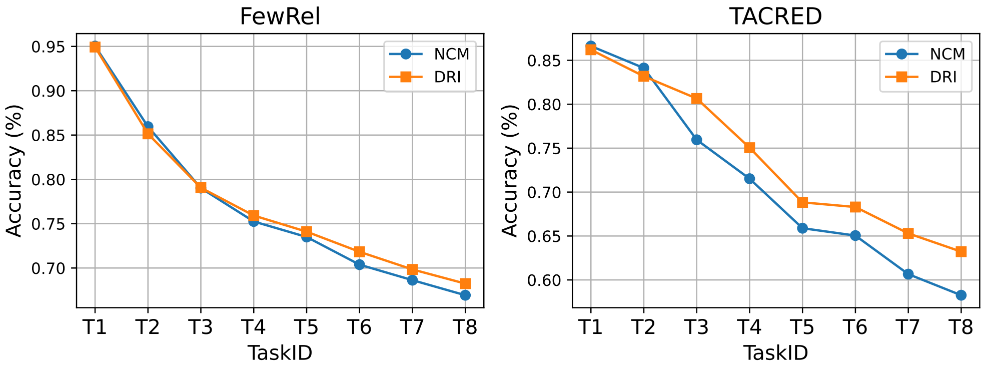


Figure 8: NCM and DRI prediction, BERT-based models

Num of generated descritions	FewRel	TACRED
K = 1	68.24	63.21
K = 3	**69.42**	64.92
K = 5	69.14	65.45
K = 7	68.92	**67.85**
K = 10	68.94	67.29

Table 7: Final accuracy (%) after 8 tasks, when varying the number of generated desciptions

Method	BERT	
	FewRel $(K = 3)$	TACRED $(K = 7)$
DCRE (Our)	**69.42**	**67.85**
w/o $\mathcal{L}_{\text{SC}}$	67.12	67.11
w/o $\mathcal{L}_{\text{MI}}$	65.96	62.41
w/o $\mathcal{L}_{\text{HM}}$	69.35	66.46
w/o $\mathcal{L}_{\text{ST}}$	69.03	66.12

Table 8: Ablation study with optimal values of K.

C.2 NCM and DRI prediction

Figures 8 and 9 compare model performance when using the conventional prediction method (NCM), and our proposed prediction strategy, DRI. The results show that, in most cases, our DRI outperforms NCM. In particular, on FewRel, DRI improves final accuracy by 1-2% compared to NCM. Notably, on the TACRED, this gap can be achieved by 4-5%.

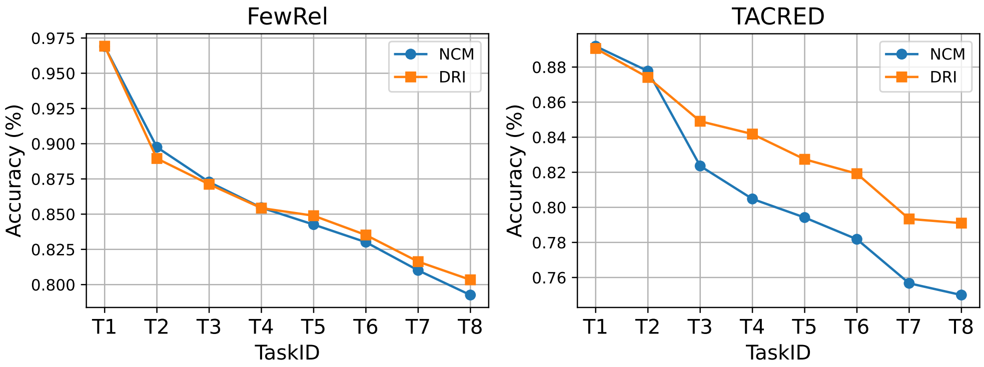


Figure 9: NCM and DRI prediction, LLM2Vec-based models

Prompt for generate label description

You are a professional data scientist, working in a relation extraction project.
Given a relation and its description, you are asked to write a more detailed description of the relation and provide 3 sentence examples of the relation.
The relation is: place served by transport hub
The description is: territorial entity or entities served by this transport hub (airport, train station, etc.)
Please generate **2** diverse samples of (relation description, examples).
Your response:

Output:

This relation identifies the geographic areas or communities that are directly connected to and supported by a specific transport hub, such as an airport, train station, or bus terminal, emphasizing the role of the hub in providing access to transportation services for these areas.
Examples:
- Heathrow Airport serves the Greater London area, providing access to travelers from London and surrounding towns.
- Union Station in Washington, D.C. serves the metropolitan area, facilitating travel for residents and visitors in Maryland and Virginia.
- Los Angeles International Airport (LAX) serves not only Los Angeles but also the wider Southern California region, including cities like Santa Monica and Long Beach.

This relation captures the locations or regions that benefit from the operational capabilities of a transport hub, highlighting its significance in regional connectivity and economic activity.
Examples:
- Chicago O'Hare International Airport serves the entire Chicago metropolitan area, enhancing accessibility for surrounding suburbs.
- Penn Station in New York City is a critical transport hub that serves travelers from New Jersey, New York, and Connecticut.
- Dallas/Fort Worth International Airport serves the Dallas and Fort Worth areas, acting as a key transit point for the entire North Texas region.

Table 9: Example generated label description