# Preserving Generalization of Language Models in Few-shot Continual Relation Extraction

**Quyen Tran[1]*, Thanh Nguyen[2]*, Anh Nguyen[2]*, Nam Le Hai[2], Trung Le[3], Linh Ngo Van [2]†, Thien Huu Nguyen[4]**

[1]VinAI Research, [2]Hanoi University of Science and Technology,
[3]Monash University, [4]University of Oregon

## Abstract

Few-shot Continual Relations Extraction (FCRE) is an emerging and dynamic area of study where models can sequentially integrate knowledge from new relations with limited labeled data while circumventing catastrophic forgetting and preserving prior knowledge from pre-trained backbones. In this work, we introduce a novel method that leverages often-discarded language model heads. By employing these components via a mutual information maximization strategy, our approach helps maintain prior knowledge from the pre-trained backbone and strategically aligns the primary classification head, thereby enhancing model performance. Furthermore, we explore the potential of Large Language Models (LLMs), renowned for their wealth of knowledge, in addressing FCRE challenges. Our comprehensive experimental results underscore the efficacy of the proposed method and offer valuable insights for future work.

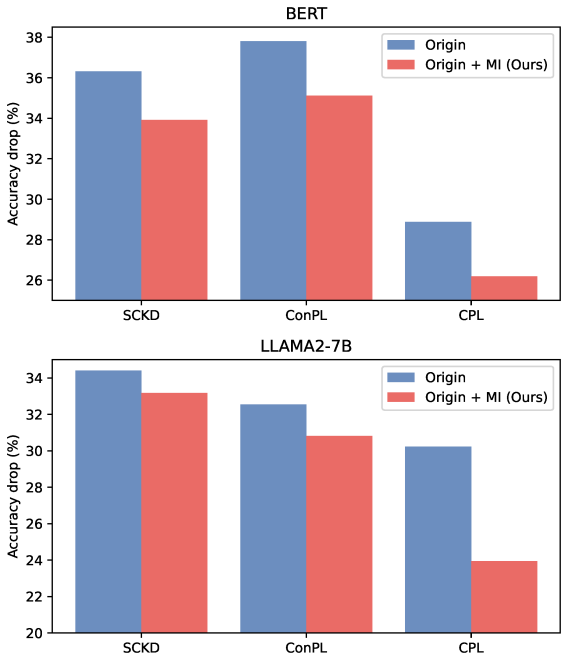


Figure 1: Accuracy drop (%) after learning eight tasks of methods on TACRED *5-way-5-shot*. Lower is better.

## 1 Introduction

Continual Relations Extraction (CRE) (Nguyen et al., 2023; Le et al., 2024c; Nguyen et al.) is a learning scenario that requires a model to identify emerging relationships between entities or objects in texts (Baldini Soares et al., 2019; Lai et al., 2022; Man et al., 2022) while maintaining the accuracy of existing classifications and avoiding the problem of *Catastrophic forgetting* (Thrun and Mitchell, 1995; French and Chater, 2002; Le et al., 2024b; Hai et al., 2024; Phan et al., 2022). In many real-world situations, models must learn from a few new samples due to the limited availability of labeled training data for relations. As a result, Few-short Continual Relation Extraction (FCRE) methods have been proposed (Qin and Joty, 2022; Chen et al., 2023) to enable models to solve new tasks where each new relation has only a minimal number of corresponding samples. However, due to the lack of data, FCRE models are often biased towards the current task compared to related scenarios, which can lead to forgetting previous knowledge and losing highly general priori from the pre-trained backbone. Thus, the challenge of FCRE is not only catastrophic forgetting but also severe overfitting.

Recent works (Wang et al., 2023; Qin and Joty, 2022; Chen et al., 2023) tackles these issues by employing memory-based approaches inspired by traditional Continual Learning methods (Rolnick et al., 2019; Buzzega et al., 2020; Lopez-Paz and Ranzato, 2017; Van et al., 2022; Le et al., 2024a), along with various strategies to enhance the model's ability to distinguish relation representations. Nevertheless, these methods solely fine-tune pre-trained BERT-based backbones for few-shot tasks, which leads to eroding prior knowledge from

*Equally contributed.
†Corresponding author: linhnv@soict.hust.edu.vn



*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13771–13784
November 12-16, 2024 

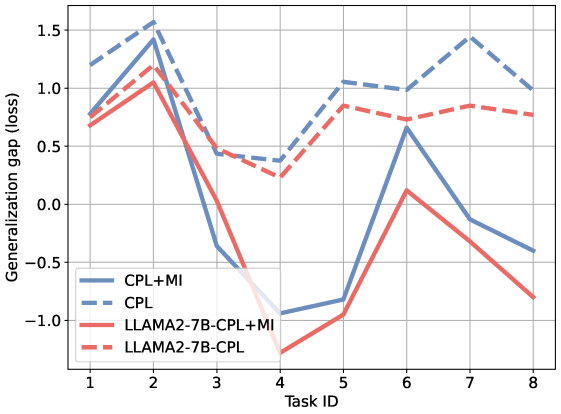


Figure 2: Generalization gap regarding loss of models after training each task (TACRED *5-way-5-shot*, seed=100).

the pre-trained model and hindering the final performance. Additionally, these methods often neglect the pre-trained LM head in favor of training a new classifier from scratch, even though this component contains rich and general knowledge that remains untapped. Therefore, we propose our *Mutual Information Maximization (MIM) strategy* that leverages pre-trained LM heads during training FCRE models for the first time. Our proposed strategy not only helps preserve the knowledge on the backbone but also assists in aligning the main classifier to improve representation learning. Extensive experimental results on benchmark datasets demonstrate the effectiveness of our novel approach in preserving the pre-trained LM's generalization capability and reducing forgetting, leading to remarkable results.

Furthermore, pre-trained Large Language Models (LLMs) (Touvron et al., 2023; Jiang et al., 2023) with billions of parameters are known for their excellence in autoregressive text generation tasks. They have also been extensively studied in text classification and information extraction (Zhao et al., 2021; Wei et al., 2023). However, these models often underperform compared to discriminative encoder models like BERT due to their generation-focused mechanism. To address this, recent work (Li et al., 2023) proposed replacing ineffective LLM heads with classification heads in the restricted space of the classification problem. This approach has shown promise, but the potential of LLMs in CL, specifically in FCRE, remains underexplored. Therefore, we conduct extensive experiments to answer: How the performance would LLMs yeild for FCRE? How will limited data in this scenario impact the generalization of LLMs? We also assess the effectiveness of our MIM strategy when using LLM heads, which were eliminated due to their unsuitability. The results offer valuable insights for the community.

To sum up, our main contributions are twofold:

- First, we introduce a novel approach to enhance FCRE models by strategically leveraging the LM heads. Through maximizing mutual information between these components and the primary classifiers, we can better preserve prior knowledge from pre-trained backbones, as well as strengthen representation learning. The experimental results demonstrate our effectiveness.

- We also investigate the application of pre-trained LLMs to FCRE tasks, including evaluating the effectiveness of the proposed method when using LLM heads, which were discarded in classification-based problems due to their unsuitability. Our comprehensive experimental results offer valuable insights.

## 2 Related work

**Continual Learning (CL)** is a learning scenario that requires models to continually acquire new knowledge from a sequence of tasks while preventing the loss of previously learned information. The main challenge in CL is *catastrophic forgetting* (French, 1993). To address this problem, memory-based approaches prove to be effective methods for both machine learning (Rebuffi et al., 2017; Shin et al., 2017) and NLP problems (Wang et al., 2019; Han et al., 2020). In particular, models need to save a few representative samples from the current task in a memory buffer and replay these samples when learning new tasks to review old knowledge.

**Fewshot Continual Relation Extraction** is a challenging scenario, which was introduced by (Qin and Joty, 2022) for Relation Extraction problems. This challenge arises due to the limited availability of data for new tasks, coupled with the high cost and time involved in obtaining high-quality data. Recent work like Wang et al. (2023); Chen et al. (2023); Ma et al. (2024) propose memory-based solutions, which suggest imposing objective functions on the embedding space and classification head. Specifically, Wang et al. (2023) employs serial objective functions based on contrastive and distillation, Qin and Joty (2022) leverage extra training data from unlabeled text, and Chen et al.

(2023) proposes a consistent prototype learning strategy to help the model distinguish between different relation representations, thus enhancing representation learning efficiency.

However, in these methods, eliminating the pre-trained LM head and training a new classifier still leads to overfitting and forgetting due to limited data, as it emphasizes discriminative features only. To address this problem, we propose a novel approach that leverages LM heads, which are often overlooked in pre-trained models for downstream tasks. Our method not only helps preserve prior knowledge from the backbone but also supports the training of the main classifier, thereby further reducing both catastrophic forgetting and overfitting.

## 3 Background

### 3.1 Problem Formulation

In the setting of FCRE, a model needs to continually acquire new knowledge from a series of tasks. For each task $t$, also denoted as $\mathcal{T}^t$, the model is trained on the training set $D^t = \{(x_i^t, y_i^t)\}_{i=1}^{N \times K}$. Here, $N$ and $K$ represent the number of classes in the new relation set $R^t$ and the number of samples corresponding to each relation, respectively. Each sample $(x_i^t, y_i^t)$ consists of a sentence $x_i$ with a pair of entities $(e_h, e_t)$ and a relation label $y_i \in R^t$. This type of task is also known as *"N-way-K-shot"*. Once task $\mathcal{T}^t$ is completed, $D^t$ is no longer available for future learning. Finally, the model will be evaluated on all task data so far in order to identify relations in $\tilde{R}^t = \bigcup_{i=1}^{t} R^i$.

### 3.2 Existing Concept of FCRE Models

Current FCRE methods (Wang et al., 2023; Chen et al., 2023; Ma et al., 2024) have considered tackling two main issues: catastrophic forgetting and overfitting. This has been achieved by exploiting the power of pre-trained BERTs and various motivated techniques which can divided into 3 main groups, including (i) using objective functions (i.e., $\mathcal{L}_0$) to enhance representation learning ability, (ii) implementing a prompt design, and (iii) employing a memory management strategy to store and retrieve knowledge of old tasks. In this paper, we propose a novel strategy that can flexibly integrate with and improve these methods (Figure 3).

Moreover, to explore the potential of pre-trained LLMs when dealing with the FCRE problems, we need to apply the current SOTA methods for LLMs, which were originally designed for "encoder-only" models. On the other hand, the examined LLMs (LLAMA2, Mistral) are "decoder-only", operating in the auto-regressive mechanism (Xie, 2017; Yang et al., 2019). Due to the differences between these models, we have to modify the original designs mentioned above (see Sec. 4.2).

## 4 Proposed Method

In this section, we first present our efficient strategy in Section 4.1 that can flexibly adapt to the existing FCRE methods and enhance model performance. After that, in Section 4.2, we explain in detail the motivation and research questions when investigating LLMs in FCRE.

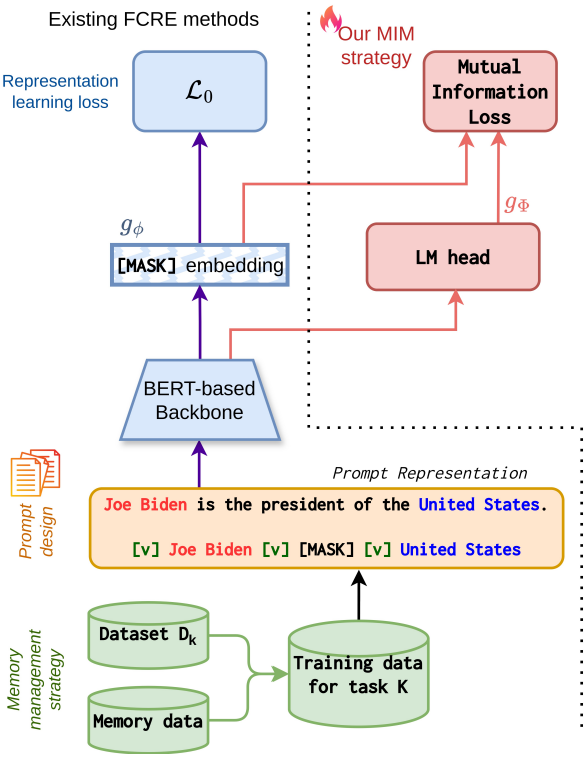


Figure 3: Our Framework

### 4.1 Mutual Information Maximization (MIM)

According to recent work (Li et al., 2023; Xu et al., 2023), using pre-trained LMs (BERTs) with their classification heads often leads to poor results. This is because the models must return responses in the vocabulary's high-dimensional space (i.e., $\|V\|$). Therefore, in downstream tasks like Relation Extraction, LM heads of pre-trained LMs are often discarded. Instead, existing work (Wang et al., 2023; Ma et al., 2024) opt for training a classification head across tasks as a better solution. However, in FCRE, training a new classifier from scratch often encourages models to emphasize only discrimina-

**FewRel** *(10-way 5-shot)*

| Method | $\mathcal{T}^1$ | $\mathcal{T}^2$ | $\mathcal{T}^3$ | $\mathcal{T}^4$ | $\mathcal{T}^5$ | $\mathcal{T}^6$ | $\mathcal{T}^7$ | $\mathcal{T}^8$ |
|---|---|---|---|---|---|---|---|---|
| SCKD | 94.75 | 82.83 | 76.21 | 72.19 | 70.61 | 67.15 | 64.86 | 62.98 |
| SCKD+MI | **94.75** | **83.88** | **76.71** | **72.34** | **70.78** | **67.36** | **65.08** | **63.95** ↑ 0.97 |
| ConPL** | **95.18** | 79.63 | 74.54 | 71.27 | 68.35 | 63.86 | 64.74 | 62.46 |
| ConPL+MI | 95.02 | **81.42** | **77.23** | **74.21** | **69.64** | **67.74** | **66.44** | **64.50** ↑ 2.04 |
| CPL | **94.87** | 85.14 | 78.80 | 75.10 | 72.57 | 69.57 | 66.85 | 64.50 |
| CPL+MI | 94.69 | **85.58** | **80.12** | **75.71** | **73.90** | **70.72** | **68.42** | **66.27** ↑ 1.77 |

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

| Method | $\mathcal{T}^1$ | $\mathcal{T}^2$ | $\mathcal{T}^3$ | $\mathcal{T}^4$ | $\mathcal{T}^5$ | $\mathcal{T}^6$ | $\mathcal{T}^7$ | $\mathcal{T}^8$ |
|---|---|---|---|---|---|---|---|---|
| SCKD | **88.42** | 79.35 | 70.61 | **66.78** | 60.47 | 58.05 | 54.41 | 52.11 |
| SCKD+MI | 87.55 | **79.39** | **70.70** | 66.68 | **61.94** | **59.81** | **55.10** | **53.63** ↑ 1.52 |
| ConPL** | **88.77** | 69.64 | 57.50 | 52.15 | 58.19 | 55.01 | 52.88 | 50.97 |
| ConPL+MI | 88.10 | **83.03** | **73.19** | **65.21** | **59.77** | **60.99** | **58.88** | **52.98** ↑ 2.01 |
| CPL | **86.27** | 81.55 | 73.52 | 68.96 | 63.96 | 62.66 | 59.96 | 57.39 |
| CPL+MI | 85.67 | **82.54** | **75.12** | **70.65** | **66.79** | **65.17** | **61.25** | **59.48** ↑ 2.09 |

Table 1: Accuracy (%) of different BERT-based methods after training for each task on TACRED and FewRel in *5-shot* settings. We highlight the rows corresponding to our method. The best result in each group is in **bold**. **Results of ConPL are reproduced (see Section 5.1)

tive features derived from sparse data streams and memory buffers. This biased behavior can make the model seriously overfit and rapidly lose prior knowledge from the pre-trained backbone and, thus, hinder the final performance.

Therefore, we propose an MIM strategy that exploits the overlooked LM head to solve the drawbacks of existing FCRE methods. Intuitively, leveraging knowledge from pre-trained LM heads will support the primary classifier, aiding the model in capturing information more holistically and better preserving old knowledge of the pre-trained backbone. In particular, inspired by (Guo et al., 2022), we aim at maximizing Mutual Information *(MI)* between latent representations on the LM head branch and on our main classifier branch as follows:

$$MI = I[g_\phi(\mathbf{x}), g_\Phi^{LM}(\mathbf{x})] \tag{1}$$

where $g_\phi$ corresponds to the class-discriminative feature representation at the classification head, $g_\Phi^{LM}$ denotes the representation at the LM head. According to (van den Oord et al., 2018):

$$MI \geq \log B + \text{InfoNCE}(\{x_i\}_{i=1}^B; h) \tag{2}$$

where we have defined

$$\begin{aligned}&\text{InfoNCE}(\{x_i\}_{i=1}^B; h) = \\ &\frac{1}{B}\sum_{i=1}^{B} \log \frac{h(g_\phi(\mathbf{x_i}), g_\Phi^{LM}(\mathbf{x_i}))}{\sum_{j=1}^{B} h(g_\phi(\mathbf{x_i}), g_\Phi^{LM}(\mathbf{x_j}))}, \\ &h(g_\phi(\mathbf{x_i}), g_\Phi^{LM}(\mathbf{x_j})) = \exp \frac{g_\phi(\mathbf{x_i})^T W g_\Phi^{LM}(\mathbf{x_j})}{\tau}\end{aligned} \tag{3}$$

where $\tau$ is the temperature, $B$ is mini-batch size and $W$ is a trainable parameter. Then, the MI loss function in our implementation is:

$$\mathcal{L}_{MI} = -\sum_{(x_i,y_i)\in D_{train}^k} \text{InfoNCE}(\{x_i\}_{i=1}^B; h) \tag{4}$$

Therefore, the objective function of the model can be summarized as:

$$\mathcal{L} = \mathcal{L}_0 + \mathcal{L}_{MI} \tag{5}$$

where $\mathcal{L}_0$ is the loss function of the original method. In this work, to demonstrate the effectiveness of our method, we integrate it into three existing methods CPL (Ma et al., 2024), ConPL (Chen et al., 2023) and SCKD (Wang et al., 2023) (see Appendix A.2).

**Discussion:** Although using pre-trained LM heads directly in downstream tasks is challenging, this does not hinder us from tapping into their wealth of knowledge to enhance our model performance in FCRE.

- First, maintaining the LM heads while fine-tuning them with a carefully controlled learning rate encourages the pre-trained backbones to retain prior knowledge and inherent behaviors. Thus, this strategy can mitigate the risk of overfitting, especially when models are trained on limited data for each task, enhancing their overall robustness and reliability.

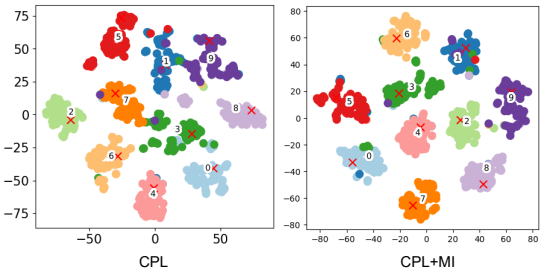


Figure 4: *t*-SNE visualization for representation of 10 relations from Task 1 on the main classification branch after the last task (FewRel *10 way - 5shot*).

- Second, applying MIM on different representation layers of the data will be a powerful aid for $\mathcal{L}_0$ in learning representations. Specifically, the mutual information of samples with the same label will be enhanced, while the information corresponding to features of different labels will be restricted. As a result, feature vectors of the same class will become more condensed, and representations of different classes will be more separated.

### 4.2 Exploiting LLMs for FCRE

**Motivations and Research questions** Pre-trained LLMs (Touvron et al., 2023; Jiang et al., 2023) are known for containing rich knowledge with billions of parameters, which have achieved impressive results in auto-regressive text generation tasks. These models have also been extensively examined in classification-based problems (Zhao et al., 2021; Wei et al., 2023). However, these models often do not outperform discriminative encoder models such as BERT because their original generation-focused mechanism, which generates answers over a large vocabulary, may not capture task-specific patterns as efficiently as label-supervised BERT models. To address this drawback, recent work (Li et al., 2023) proposed directly extracting latent representations from the final LLaMA decoder layer and mapping them into the label space through feed-forward layers. Specifically, the LLM heads, which have been found ineffective, are removed and replaced by a classification head trained from scratch using CrossEntropy loss. This approach has shown promising results. However, exploration in the area of Continual Learning, specifically Few-shot Continual Relation Extraction (FCRE), has not yet been thoroughly investigated. Therefore, in this work, we conduct extensive experiments to answer the following research questions ***(RQs)***:

- ***RQ1:*** *How the performance would LLMs yield in FCRE tasks?* Will it yield significantly better results compared to conventional BERT-based models? How will the limited data in the FCRE scenario impact the generalization of this model class? It would be interesting to examine the behavior of an LLM, which contains rich prior knowledge in the context of the FCRE problem, where each task only has very little data, and the model will usually be forgotten and severely overfit.

- ***RQ2:*** Our study also aims to assess *the effectiveness of employing our MIM strategy for LLMs*, particularly in addressing the challenges of forgetting prevention and overfitting reduction. Does using LLM heads according to our strategy eliminate the prejudice about the unsuitability of LLMs in classification-based problems, specifically FCRE?

**How to adapt BERT-based FCRE methods to LLMs?** Because current FCRE methods are used for BERT-based backbones, which are *"encoder-only"* language models. It is essential to modify their original design to adapt to *"decoder-only"* LLMs like LLAMA2-7B (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023), which operate in the auto-regressive mechanism (Xie, 2017; Yang et al., 2019; et al, 2023). See illustration in Figure 6, Appendix. In particular:

- (i) The prompted inputs will be in the form of: *"[Original sentence]. The relation between [Entity 1] and [Entity 2] is [Answer]"*;

- (ii) The embedding used for the main classifier (i.e., $g_\phi(\cdot)$) is now the embedding of the word *"is"* in the corresponding input, instead of "[MASK] embedding" in Figure 3.

## 5 Experimental Results

In this part, we first present the experiment setup in Section 5.1, followed by the results that demonstrate the effectiveness of our proposed method

(Section 5.2) when using BERT-based backbones. We then discuss the investigation results of using pre-trained LLMs for FCRE tasks in Section 5.3.

### 5.1 Experiment Setup

In our experiments, we use three current state-of-the-art methods as baselines, including: SCKD (Wang et al., 2023), ConPL (Chen et al., 2023), and CPL (Ma et al., 2024). Besides, the models are evaluated using pre-trained models consisting of BERT (Devlin et al., 2018), LLAMA2-7B (Touvron et al., 2023), and Mistral-7B (Jiang et al., 2023), on two benchmark datasets: FewRel (Han et al., 2018) and TACRED (Zhang et al., 2017). We note that we have reproduced the results of ConPL (Chen et al., 2023) under the same setting as SCKD and CPL. The reason is that the evaluation strategy in this paper is impractical for continual learning scenarios. Please refer to Appendix A for more details.

### 5.2 Evaluation

**a. Using LM heads significantly improves the model's accuracy.** Table 1 reports the results of baselines and our proposed method ***(+MI)***, which exploits pre-trained LM heads beside the primary classifiers. In general, our method consistently helps improve the performance of existing methods in all cases. On both datasets, our strategy improved the final accuracy by around 2% when integrated with CPL and ConPL and around 1% when combined with SCKD. Moreover, considering accuracy after learning immediate tasks, ConPL+MI, when using our proposed strategy, can exceed the original version by about 15% on TACRED.

**b. Exploiting the LM head effectively helps reduce forgetting and overfitting.** Figure 1 and Table 2 show the accuracy drop after completing 8 tasks in various cases. The results indicate that our method significantly helps reduce forgetting for the baselines by approximately 1 to 3%. Moreover, Figure 2 shows generalization gaps (i.e., $\delta$ = test loss − train loss) after training each task of different models. The results show that our MIM strategy helps the models minimize these gaps significantly, thereby increasing their generalization.

**c. The LM head supports representation learning.** Figure 4 presents representations in the latent space of CPL model before and after exploiting our MIM strategy (CPL+MI) on data of Task 1, after learning 8 tasks. It can be seen that the test features belonging to different categories of CPL+MI are better separated and therefore achieve better results. In addition, we provide a t-SNE visualization about features of the first task in the latent space on the LM head after learning the final tasks (Figure 5), confirming the benefits when taking advantage of this component to enhance the performance of models.

| | FewRel | | TACRED | |
|---|---|---|---|---|
| | Original | + MI | Original | + MI |
| SCKD | 36.31 | **33.92** | 31.77 | **30.80** |
| ConPL | 37.80 | **35.12** | 32.72 | **30.52** |
| CPL | 28.88 | **26.19** | 30.37 | **28.42** |

Table 2: Accuracy drop (%) after learning eight tasks of methods on the FewRel and TACRED in *5-shot* settings.

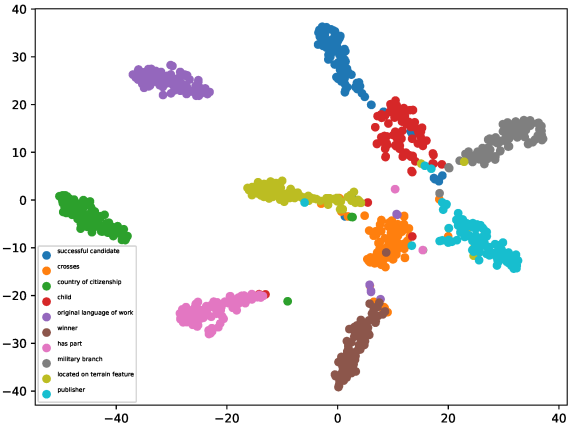


Figure 5: *t*-SNE visualization of the representation of 10 relations from the first task of CPL+MI on the LM head after the last task (FewRel *10-way 5-shot*).

**d. Ablation study**

For further analyzing the effectiveness of our proposed method, we make an ablation study and present the experimental results in Table 3. Regarding ConPL, it becomes evident that our MIM (e.i., +MI) plays a pivotal role compared to the loss components proposed in the original paper. Specifically, the elimination of each loss component among $L_{cc}$, $L_{dc}$ and $L_{fc}$ leads to only a marginal decline in performance. However, removing MI results in a notable decrease in accuracy across tasks, except for tasks 1 and 5. In the case of the SCKD, we note a substantial impact when excluding the distillation element (i.e., $L_{dst}$). This underscores the pivotal role of this component in mitigating forgetting while our proposed MI mechanism continues to enhance the performance of the overall model.

Moreover, we also explore a scenario in which the LM head is frozen to retain the knowledge from the pretraining phase fully. We notice inconsis-

| Method | $\mathcal{T}^1$ | $\mathcal{T}^2$ | $\mathcal{T}^3$ | $\mathcal{T}^4$ | $\mathcal{T}^5$ | $\mathcal{T}^6$ | $\mathcal{T}^7$ | $\mathcal{T}^8$ |
|---|---|---|---|---|---|---|---|---|
| ConPL + MI | 88.10 | **83.03** | 73.19 | 65.21 | 59.77 | **60.99** | **58.88** | **52.98** |
| w.o MI (ConPL) | **88.77** | 69.64 | 57.50 | 52.15 | 58.19 | 55.01 | 52.88 | 50.97 |
| w.o $L_{cc}$† | 88.15 | 83.01 | 73.11 | 65.16 | 58.70 | 60.06 | 58.61 | 52.69 |
| w.o $L_{dc}$† | 88.10 | 82.67 | 73.10 | 65.06 | 58.70 | 60.36 | 58.71 | 52.84 |
| w.o $L_{fc}$† | 88.06 | 81.15 | 72.04 | 63.15 | 56.26 | 59.30 | 57.69 | 50.10 |
| freeze LM head | 88.22 | 80.27 | **77.15** | **67.72** | **59.62** | 57.75 | 54.73 | 52.10 |
| SCKD + MI | 87.55 | **79.39** | 70.70 | **66.78** | **61.94** | **59.81** | 55.10 | **53.63** |
| w.o MI (SCKD) | **88.42** | 79.35 | 70.61 | 66.68 | 60.47 | 58.05 | 54.41 | 52.11 |
| w.o $L_{dst}$† | 87.61 | 77.15 | 67.21 | 62.21 | 57.11 | 54.98 | 50.53 | 50.38 |
| w.o $aug$† | 87.66 | 78.06 | 69.29 | 66.16 | 61.06 | 59.71 | 55.05 | 53.38 |
| w.o $L_{dst}$ and $aug$† | 87.37 | 76.81 | 65.88 | 62.03 | 56.81 | 52.87 | 49.41 | 46.09 |
| freeze LM head | 87.61 | 78.41 | **70.62** | 65.98 | 61.33 | 58.90 | **55.19** | 51.99 |
| CPL + MI | 85.67 | **82.54** | **75.12** | **70.65** | **66.79** | **65.17** | **61.25** | **59.48** |
| freeze LM head | **86.17** | 80.52 | 73.84 | 69.03 | 64.33 | 62.36 | 60.19 | 57.99 |

Table 3: Ablation study on TACRED in the *5-way-5- shot* setting. †The components in the ablation study of the existing methods are described in Appendix A.2.

| | FewRel | | TACRED | |
|---|---|---|---|---|
| | Original | + MI | Original | + MI |
| SCKD | 62.98 | **63.45** | 52.11 | **53.63** |
| Llama2-7B-SCKD | 65.14 | **66.58** | 54.26 | **55.17** |
| ConPL | 62.46 | **64.50** | 50.97 | **52.98** |
| Llama2-7B-ConPL | 63.97 | **65.18** | 54.72 | **56.07** |
| CPL | 64.50 | **66.27** | 57.39 | **59.48** |
| Llama2-7B-CPL | 69.87 | **72.08** | 58.03 | **62.04** |
| Mistral-7B-CPL | 71.89 | **75.02** | 64.11 | **65.48** |

Table 4: Final Accuracy (%) of methods after training the final task in the 5-shot settings.

tent changes during the task learning process, with certain tasks demonstrating performance improvements while others exhibit declines. We hypothesize that in specific cases, the LM's pretraining-derived general knowledge can facilitate recognizing specific relations. Consequently, fine-tuning the model on domain-restricted data might compromise this capability. Conversely, for other relations, the general knowledge of the pretraining stage may not hold significant value.

### 5.3 Using LLM for FCRE

**RQ1: How the performance would LLMs yield in FCRE tasks?** Table 4 depicts the increase in final accuracy after learning 8 FCRE tasks when the BERT-based backbone is replaced by the LLM backbone. Specifically, improvements can be as much as 3.75% in the case of LLAMA-2-7B, and 8.75% for Mistral-7B across both datasets. In addition, Table 6 shows the full results of FCRE models on both datasets. Mostly, during the training of eight tasks, the LLMs tend to provide higher accuracy than the BERT-based models. For some immediate tasks, LLAMA2-7B can achieve up to 16% higher accuracy than BERT-based models in TACRED, although their accuracy can be slightly lower in other cases. Besides, the differences in performance after training the first task and the last task (Accuracy drop - column $\Delta \downarrow$) in LLMs are smaller than in BERT-based models, from 2 to 5% in the case of LLAMA2-7B and as much as 8% for Mistral-CPL. These experimental results confirm the general superiority of LLM in solving FCRE compared to the class of conventional BERT-based models.

On the other hand, pre-trained LLMs are known to be knowledge-rich models with high generalization capabilities. However, for the first task, LLMs achieve accuracies of around 96% on FewRel and around 86% on TACRED, having no clear advantage over BERT-based models. Besides, the results in Table 6 clearly demonstrate the degradation of prior knowledge when applying pre-trained LLM in FCRE. In particular, the model's accuracy can drop by 30 - 32% for LLAMA2-7B and by 20 - 25% for Mistral-7B, after training 8 tasks.

Thanks to thorough training on large datasets, LLMs with billions of parameters contain a wealth of knowledge and have great potential in downstream tasks. However, in some cases, with the current operating mechanism of an autoregressive decoder, employing such a model with billions of parameters, as opposed to one with hundreds of

**FewRel** *(10-way–5-shot)*

| Method | $\mathcal{T}^1$ | $\mathcal{T}^2$ | $\mathcal{T}^3$ | $\mathcal{T}^4$ | $\mathcal{T}^5$ | $\mathcal{T}^6$ | $\mathcal{T}^7$ | $\mathcal{T}^8$ | $\Delta \downarrow$ |
|---|---|---|---|---|---|---|---|---|---|
| SCKD | 94.75 | 82.83 | 76.21 | 72.19 | 70.61 | 67.15 | 64.86 | 62.98 | 31.77 |
| SCKD + MI | 94.75 | 83.88 | **76.71** | 72.34 | 70.78 | 67.36 | 65.08 | 63.45 | 31.30 |
| Llama2-7B-SCKD | **95.63** | 82.76 | 76.04 | 74.91 | 70.10 | 66.52 | 64.89 | 65.14 | 30.49 |
| Llama2-7B-SCKD + MI | 95.22 | **85.01** | 76.63 | **76.50** | **72.19** | **67.47** | **67.03** | **66.58** | **28.64** |
| ConPL** | **95.18** | 79.63 | 74.54 | 71.27 | 68.35 | 63.86 | 64.74 | 62.46 | 32.72 |
| ConPL + MI | 95.02 | 81.42 | 77.23 | 74.21 | 69.64 | 67.74 | **66.44** | 64.50 | 30.52 |
| Llama2-7B-ConPL | 94.72 | 82.43 | 75.07 | 73.95 | 72.67 | 65.80 | 63.41 | 63.79 | 30.93 |
| Llama2-7B-ConPL + MI | 94.50 | **83.75** | **77.61** | **74.78** | **72.83** | **68.01** | 63.98 | **65.18** | **29.32** |
| CPL | 94.87 | 85.14 | 78.80 | 75.10 | 72.57 | 69.57 | 66.85 | 64.50 | 30.37 |
| CPL + MI | 94.69 | 85.58 | 80.12 | 75.71 | 73.90 | 70.72 | 68.42 | 66.27 | 28.42 |
| Llama2-7B-CPL | 95.73 | 85.87 | 80.57 | 78.60 | 77.30 | 73.95 | 71.35 | 69.87 | 25.86 |
| Llama2-7B-CPL + MI | 95.63 | 87.14 | 83.25 | 80.59 | **79.20** | 76.41 | 74.62 | 72.08 | 23.55 |
| Mistral-7B-CPL | **96.57** | 86.80 | 83.31 | 79.45 | 77.17 | 74.24 | 73.59 | 71.89 | 24.68 |
| Mistral-7B-CPL + MI | 96.55 | **90.77** | **84.81** | **83.08** | 78.92 | **77.27** | **77.05** | **75.02** | **21.53** |

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

| Method | $\mathcal{T}^1$ | $\mathcal{T}^2$ | $\mathcal{T}^3$ | $\mathcal{T}^4$ | $\mathcal{T}^5$ | $\mathcal{T}^6$ | $\mathcal{T}^7$ | $\mathcal{T}^8$ | $\Delta \downarrow$ |
|---|---|---|---|---|---|---|---|---|---|
| SCKD | 88.42 | 79.35 | 70.61 | 66.78 | 60.47 | 58.05 | 54.41 | 52.11 | 36.31 |
| SCKD + MI | 87.55 | 78.39 | 69.70 | **66.88** | 61.94 | 59.81 | 55.10 | 53.63 | 33.92 |
| Llama2-7B-SCKD | **88.67** | 84.48 | 72.53 | 63.10 | 62.01 | 59.38 | 57.18 | 54.26 | 34.41 |
| Llama2-7B-SCKD + MI | 88.35 | **84.90** | **74.32** | 63.48 | **63.37** | **60.20** | **59.64** | **55.17** | **33.18** |
| ConPL** | **88.77** | 69.64 | 57.50 | 52.15 | 58.19 | 55.01 | 52.88 | 50.97 | 37.80 |
| ConPL + MI | 88.10 | 83.03 | 73.19 | 65.21 | 58.77 | **60.99** | **58.88** | 52.98 | 35.12 |
| Llama2-7B-ConPL | 87.26 | 81.72 | 73.04 | 65.67 | 60.96 | 58.47 | 56.49 | 54.72 | 32.54 |
| Llama2-7B-ConPL + MI | 86.88 | **83.11** | **73.83** | **67.58** | **61.87** | 60.31 | 56.83 | **56.07** | **30.81** |
| CPL | 86.27 | 81.55 | 73.52 | 68.96 | 63.96 | 62.66 | 59.96 | 57.39 | 28.88 |
| CPL + MI | 85.67 | **82.54** | 75.12 | 70.65 | 66.79 | 65.17 | 61.25 | 59.48 | 26.19 |
| Llama2-7B-CPL | **86.76** | 75.94 | 70.65 | 68.64 | 67.44 | 65.12 | 60.27 | 58.03 | 30.23 |
| Llama2-7B-CPL + MI | 85.55 | 77.91 | 76.49 | 74.99 | 69.15 | 68.19 | 64.19 | 62.04 | 23.51 |
| Mistral-7B-CPL | 86.67 | 80.98 | 77.16 | 73.24 | 70.05 | 67.70 | 67.04 | 64.11 | 22.56 |
| Mistral-7B-CPL + MI | 86.32 | 81.00 | **77.71** | **75.48** | **71.92** | **71.02** | **67.69** | **65.48** | **20.84** |

Table 5: Accuracy (%) of methods using different LMs after training for each task. We highlight the rows corresponding to our proposed method. The best result in each group is in **bold**. **Results of ConPL are reproduced. Columns $\Delta \downarrow$ present Accuracy drop after learning 8 tasks.

millions (BERT), proves exceedingly expensive for only marginal improvements in accuracy. Even on TACRED, the final accuracy of LLAMA2-7B-CPL is lower than that of CPL+MI, indicating that our method with the BERT-based model can effectively replace the LLM in this case. These findings necessitate the development of more effective methodologies to ensure the effectiveness of LLMs within this challenging setting

**RQ2: The effectiveness of exploiting our MIM strategy for LLMs in FCRE tasks** Figure 1 and Table 6 clearly show that our strategy significantly mitigates accuracy drop in LLMs, which could reach up to 6% on TACRED and 4% on FewRel, and better than on BERT-based models. Besides, Figure 2 consistently illustrates the effectiveness of our method in reducing overfitting. It can be said that with our proposed strategy, LLM heads are no longer an obstacle when applying pre-trained LLMs to classification tasks. On the contrary, using LLMs demonstrates the clearest and most significant improvement in mitigating catastrophic forgetting and reducing overfitting.

## 6 Conclusion

In this work, we introduce a novel method that utilizes pre-trained language model heads to maintain the generalization of LMs in FCRE problems. By making use of this often ignored component through a mutual information strategy, our approach also significantly improves the comprehensiveness of the representation on the main classifier. Additionally, we present comprehensive experimental results that demonstrate the impact of using LLMs for FCRE and provide valuable insights to the community.

## Limitations

- First, our proposed method and current investigations in this paper apply only to high-level RE tasks, where all entities are assumed to be given. Therefore, to achieve more practical results, it is motivating to consider end-to-end RE problems, covering entity recognition to relation extraction between entities in the future.

- Another potential limitation could arise from the fact that pre-trained LMs used in our work might inherit biases from their pre-training data. These biases can manifest in various forms, such as gender, racial, or cultural biases, and could be exacerbated in scenarios with limited labeled data, as in FCRE tasks. Our method endeavors to transfer the knowledge within the LMs to the classification head by leveraging Mutual Information (MI), which could inadvertently perpetuate biased representations. Such biased representations may have adverse consequences, potentially resulting in misidentifying relations associated with biased information. This raises an open question for the research community to investigate further, exploring the impact of bias on FCRE tasks when utilizing LLMs.

## Acknowledgements

This research has been supported by the Army Research Office (ARO) grant W911NF-21-1-0112, the NSF grant CNS-1747798 to the IUCRC Center for Big Learning, and the NSF grant # 2239570. This research is also supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract 2022-22072200003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government.

## 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*, Florence, Italy.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. 2020. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xiudi Chen, Hui Wu, and Xiaodong Shi. 2023. Consistent prototype learning for few-shot continual relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7409–7422. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Hugo Touvron et al. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Robert M. French. 1993. Catastrophic interference in connectionist networks: Can it be predicted, can it be prevented? In *Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993]*, pages 1176–1177. Morgan Kaufmann.

Robert M. French and Nick Chater. 2002. Using noise to compute error surfaces in connectionist networks: A novel means of reducing catastrophic forgetting. *Neural Comput.*, 14(7):1755–1769.

Yiduo Guo, Bing Liu, and Dongyan Zhao. 2022. Online continual learning through mutual information maximization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8109–8126. PMLR.

Nam Le Hai, Trang Nguyen, Linh Ngo Van, Thien Huu Nguyen, and Khoat Than. 2024. Continual variational dropout: a view of auxiliary local variables in continual learning. *Machine Learning*, 113(1):281–323.

Xu Han, Yi Dai, Tianyu Gao, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2020. Continual relation learning via episodic memory activation and reconsolidation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6429–6440. Association for Computational Linguistics.

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

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Viet Lai, Hieu Man, Linh Ngo, Franck Dernoncourt, and Thien Nguyen. 2022. Multilingual subevent relation extraction: A novel dataset and structure induction method. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5559–5570.

Minh Le, An Nguyen, Huy Nguyen, Trang Nguyen, Trang Pham, Linh Van Ngo, and Nhat Ho. 2024a. Mixture of experts meets prompt-based continual learning. *Advances in Neural Information Processing Systems*.

Thanh-Thien Le, Viet Dao, Linh Nguyen, Thi-Nhung Nguyen, Linh Ngo, and Thien Nguyen. 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)*, pages 3632–3644.

Thanh-Thien Le, Manh Nguyen, Tung Thanh Nguyen, Linh Ngo Van, and Thien Huu Nguyen. 2024c. Continual relation extraction via sequential multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18444–18452.

Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu Lee Wang, Qing Li, and Xiaoqin Zhong. 2023. Label supervised llama finetuning. *CoRR*, abs/2310.01208.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6467–6476.

Shengkun Ma, Jiale Han, Yi Liang, and Bo Cheng. 2024. Making pre-trained language models better continual few-shot relation extractors. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10970–10983. ELRA and ICCL.

Hieu Man, Nghia Trung Ngo, Linh Ngo Van, and Thien Huu Nguyen. 2022. Selecting optimal context sentences for event-event relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11058–11066.

Huy Nguyen, Chien Nguyen, Linh Ngo, Anh Luu, and Thien Nguyen. 2023. A spectral viewpoint on continual relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9621–9629.

Manh Nguyen, Quang Duc Nguyen, Nam Le Hai, Linh Van Ngo, Sang Dinh, and Thien Huu Nguyen. Continual information extraction via sequential multi-task learning. *Available at SSRN 4895531*.

Hoang Phan, Anh Phan Tuan, Son Nguyen, Ngo Van Linh, and Khoat Than. 2022. Reducing catastrophic forgetting in neural networks via gaussian mixture approximation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 106–117. Springer.

Chengwei Qin and Shafiq R. Joty. 2022. Continual few-shot relation learning via embedding space regularization and data augmentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2776–2789. Association for Computational Linguistics.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2017. icarl: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5533–5542. IEEE Computer Society.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Gregory Wayne. 2019. Experience replay for continual learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 348–358.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2990–2999.

Sebastian Thrun and Tom M. Mitchell. 1995. Lifelong robot learning. *Robotics and Autonomous Systems*, 15(1):25–46. The Biology and Technology of Intelligent Autonomous Agents.

Hugo Touvron et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Linh Ngo Van, Nam Le Hai, Hoang Pham, and Khoat Than. 2022. Auxiliary local variables for improving regularization/prior approach in continual learning. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 16–28. Springer.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019. Sentence embedding alignment for lifelong relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 796–806. Association for Computational Linguistics.

Xinyi Wang, Zitao Wang, and Wei Hu. 2023. Serial contrastive knowledge distillation for continual few-shot relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12693–12706. Association for Computational Linguistics.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with chatgpt. *CoRR*, abs/2302.10205.

Ziang Xie. 2017. Neural text generation: A practical guide. *CoRR*, abs/1711.09534.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2023. Large language models for generative information extraction: A survey. *arXiv preprint arXiv:2312.17617*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

## A Implementation details

For each reported result, we conduct 6 independent runs with different random seeds and report the mean. Our code is available at https://github.com/thanhnx12/CRE-via-MMI

*Note:* As discussed in (Li et al., 2023), LLaMA-2-7B model gives better results compared with LLaMA-2-13B. Therefore, we opt to use LLaMA-2-7B to examine in our experiments.

### A.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.

### A.2 Baselines

In this work, we showcase our approach through thorough experiments using three recent SOTA methods in FCRE as the baselines, including:

- **SCKD** (Wang et al., 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.

  In this paper, to conduct the ablation study in Table 3, we denote $L_{dst}$ as the representative of all the losses serving the distillation and contrastive learning mentioned above and $aug$ as the augmentation technique on the memory buffer.

- **ConPL** (Chen et al., 2023) proposes a method that consists of three fundamental modules: a prototype-based classification module, a memory-enhanced module, and a novel consistent learning module that enforces distribution consistency to prevent forgetting. Additionally, ConPL leverages prompt learning to improve representation learning and incorporate focal loss to alleviate confusion among closely related classes.

  This paper conducts the ablation study in Table 3where the role of each component of ConPL's objective function is analyzed. In particular, $L_{cc}$ helps constrain the consistency between samples and corresponding prototypes of old tasks, $L_{dc}$ forces the consistency regarding the distribution of samples and prototypes, and $L_{fc}$ is a focal loss that alleviates the difficulty of choosing negative classes during inference.

- **CPL** (Ma et al., 2024) CPL proposes a Contrastive Prompt Learning framework, which designs prompts to generalize across categories and uses margin-based contrastive 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.

### A.3 Evaluation Protocol

**Metric** We use final average accuracy to evaluate methods in our experiments. The average accuracy at 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 $T_j$.

**Prediction mechanism** As mentioned in 5.1, our methods follow the evaluation strategy in the setting of SCKD and CPL. Specifically, during the testing phase, the learned model is required to evaluate all classes/ relations it has been trained on so far.

Note that in the original code repository of ConPL (e.g., Lines 18-53 in this file), this method follows a different evaluation process. In particular, after training on task $\mathcal{T}^k$, the model has been

**FewRel** *(10-way–5-shot)*

| Method | $\mathcal{T}^1$ | $\mathcal{T}^2$ | $\mathcal{T}^3$ | $\mathcal{T}^4$ | $\mathcal{T}^5$ | $\mathcal{T}^6$ | $\mathcal{T}^7$ | $\mathcal{T}^8$ | $\Delta\downarrow$ |
|---|---|---|---|---|---|---|---|---|---|
| SCKD | 94.75 | 82.83 | 76.21 | 72.19 | 70.61 | 67.15 | 64.86 | 62.98 | 31.77 |
| SCKD + MI | $94.75_{\pm 0.37}$ | $83.88_{\pm 0.67}$ | $76.71_{\pm 2.48}$ | $72.34_{\pm 1.43}$ | $70.78_{\pm 0.82}$ | $67.36_{\pm 0.73}$ | $65.08_{\pm 2.43}$ | $63.45_{\pm 2.44}$ | 31.30 |
| Llama2-7B-SCKD | $95.63_{\pm 0.56}$ | $82.76_{\pm 2.26}$ | $76.04_{\pm 4.22}$ | $74.91_{\pm .77}$ | $70.10_{\pm 3.63}$ | $66.52_{\pm 2.9}$ | 64.89 2.85 | $65.14_{\pm 1.52}$ | 30.49 |
| Llama2-7B-SCKD + MI | $95.22_{\pm 0.53}$ | $85.01_{\pm 2.4}$ | $76.63_{\pm 1.19}$ | $76.50_{\pm 1.28}$ | $72.19_{\pm 1.4}$ | $67.47_{\pm 1.87}$ | $67.03_{\pm 2.97}$ | $66.58_{\pm 2.11}$ | **28.64** |
| ConPL** | $95.18_{\pm 0.73}$ | $79.63_{\pm 1.27}$ | $74.54_{\pm 1.13}$ | $71.27_{\pm 0.85}$ | $68.35_{\pm 0.86}$ | $63.86_{\pm 2.03}$ | $64.74_{\pm 1.39}$ | $62.46_{\pm 1.54}$ | 32.72 |
| ConPL + MI | $95.02_{\pm 0.4}$ | $81.42_{\pm 1.93}$ | $77.23_{\pm 1.01}$ | $74.21_{\pm 1.5}$ | $69.64_{\pm 1.19}$ | $67.74_{\pm 1.52}$ | $66.44_{\pm 1.91}$ | $64.50_{\pm 1.15}$ | 30.52 |
| Llama2-7B-ConPL | $94.72_{\pm 1.15}$ | $82.43_{\pm 1.69}$ | $75.07_{\pm 1.62}$ | $73.95_{\pm 2.75}$ | $72.67_{\pm 1.51}$ | $65.80_{\pm 1.46}$ | $63.41_{\pm 2.15}$ | $63.79_{\pm 2.76}$ | 30.93 |
| Llama2-7B-ConPL + MI | $94.50_{\pm 0.57}$ | $83.75_{\pm 1.05}$ | $77.61_{\pm 1.27}$ | $74.78_{\pm 3.19}$ | $72.83_{\pm 2.74}$ | $68.01_{\pm 2.23}$ | $63.98_{\pm}3.1$ | $65.18_{\pm 1.99}$ | **29.32** |
| CPL | 94.87 | 85.14 | 78.80 | 75.10 | 72.57 | 69.57 | 66.85 | 64.50 | 30.37 |
| CPL + MI | $94.69_{\pm 0.7}$ | $85.58_{\pm 1.88}$ | $80.12_{\pm 2.45}$ | $75.71_{\pm 2.28}$ | $73.90_{\pm 1.8}$ | $70.72_{\pm 0.91}$ | $68.42_{\pm 1.77}$ | $66.27_{\pm 1.58}$ | 28.42 |
| Llama2-7B-CPL | $95.73_{\pm 0.92}$ | $85.87_{\pm 1.46}$ | $80.57_{\pm 1.74}$ | $78.60_{\pm 3.31}$ | $77.30_{\pm 2.41}$ | $73.95_{\pm 1.54}$ | $71.35_{\pm 3.75}$ | $69.87_{\pm 2.32}$ | 25.86 |
| Llama2-7B-CPL + MI | $95.63_{\pm 1.08}$ | $87.14_{\pm 1.94}$ | $83.25_{\pm 2.14}$ | $80.59_{\pm 2.37}$ | $79.20_{\pm 1.36}$ | $76.41_{\pm 2.13}$ | $74.62_{\pm}1.73$ | $72.08_{\pm 3.18}$ | 23.55 |
| Mistral-7B-CPL | $96.57_{\pm 0.40}$ | $86.80_{\pm 2.53}$ | $83.31_{\pm 1.94}$ | $79.45_{\pm 2.53}$ | $77.17_{\pm 2.2}$ | $74.24_{\pm 1.96}$ | $73.59_{\pm 2.00}$ | $71.89_{\pm 1.97}$ | 24.68 |
| Mistral-7B-CPL + MI | $96.55_{\pm 0.43}$ | $90.77_{\pm 2.11}$ | $84.81_{\pm 1.09}$ | $83.08_{\pm 1.5}$ | $78.92_{\pm 1.35}$ | $77.27_{\pm 2.06}$ | $77.05_{\pm 2.3}$ | $75.02_{\pm 1.67}$ | **21.53** |

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

| Method | $\mathcal{T}^1$ | $\mathcal{T}^2$ | $\mathcal{T}^3$ | $\mathcal{T}^4$ | $\mathcal{T}^5$ | $\mathcal{T}^6$ | $\mathcal{T}^7$ | $\mathcal{T}^8$ | $\Delta\downarrow$ |
|---|---|---|---|---|---|---|---|---|---|
| SCKD | 88.42 | 79.35 | 70.61 | 66.78 | 60.47 | 58.05 | 54.41 | 52.11 | 36.31 |
| SCKD + MI | $87.55_{\pm 0.48}$ | $78.39_{\pm 2.18}$ | $69.70_{\pm 1.75}$ | $66.88_{\pm 1.56}$ | $61.94_{\pm 2.87}$ | $59.81_{\pm 1.56}$ | $55.10_{\pm 3.63}$ | $53.63_{\pm 2.31}$ | 33.92 |
| Llama2-7B-SCKD | $88.67_{\pm 0.56}$ | $84.48_{\pm 2.26}$ | $72.53_{\pm 4.22}$ | $63.10_{\pm 4.77}$ | $62.01_{\pm 3.63}$ | $59.38_{\pm 2.90}$ | $57.18_{\pm 2.85}$ | $54.26_{\pm 1.52}$ | 34.41 |
| Llama2-7B-SCKD + MI | $88.35_{\pm 1.11}$ | $84.90_{\pm 2.59}$ | $74.32_{\pm 3.73}$ | $63.48_{\pm 2.03}$ | $63.37_{\pm 2.44}$ | $60.20_{\pm 3.54}$ | $59.64_{\pm 3.19}$ | $55.17_{\pm 2.68}$ | **33.18** |
| ConPL** | $88.77_{\pm 0.84}$ | $69.64_{\pm 1.93}$ | $57.50_{\pm 2.48}$ | $52.15_{\pm 1.59}$ | $58.19_{\pm 2.31}$ | $55.01_{\pm 3.12}$ | $52.88_{\pm 3.66}$ | $50.97_{\pm 3.41}$ | 37.80 |
| ConPL + MI | $88.10_{\pm 0.68}$ | $83.03_{\pm 3.38}$ | $73.19_{\pm 1.57}$ | $65.21_{\pm 3.04}$ | $58.77_{\pm 3.45}$ | $60.99_{\pm 1.61}$ | $58.88_{\pm 2.52}$ | $52.98_{\pm 1.68}$ | 35.12 |
| Llama2-7B-ConPL | $87.26_{\pm 1.22}$ | $81.72_{\pm 2.54}$ | $73.04_{\pm 2.92}$ | $65.67_{\pm 2.07}$ | $60.96_{\pm 4.39}$ | $58.47_{\pm 3.32}$ | $56.49_{\pm 3.2}$ | $54.72_{\pm 2.24}$ | 32.54 |
| Llama2-7B-ConPL + MI | $86.88_{\pm 1.03}$ | $83.11_{\pm 3.46}$ | $73.83_{\pm 2.88}$ | $67.58_{\pm 2.04}$ | $61.87_{\pm 4.16}$ | $60.31_{\pm 4.41}$ | $56.83_{\pm 2.57}$ | $56.07_{\pm 3.45}$ | **30.81** |
| CPL | 86.27 | 81.55 | 73.52 | 68.96 | 63.96 | 62.66 | 59.96 | 57.39 | 28.88 |
| CPL + MI | $85.67_{\pm 0.8}$ | $82.54_{\pm 2.98}$ | $75.12_{\pm 3.67}$ | $70.65_{\pm 2.75}$ | $66.79_{\pm 2.18}$ | $65.17_{\pm 2.48}$ | $61.25_{\pm 1.52}$ | $59.48_{\pm 3.53}$ | 26.19 |
| Llama2-7B-CPL | $86.76_{\pm 1.58}$ | $75.94_{\pm 4.76}$ | $70.65_{\pm 2.57}$ | $68.64_{\pm 3.03}$ | $67.44_{\pm 2.95}$ | $65.12_{\pm 3.85}$ | $60.27_{\pm 3.79}$ | $58.03_{\pm 1.98}$ | 30.23 |
| Llama2-7B-CPL + MI | $85.55_{\pm 0.74}$ | $77.91_{\pm 2.8}$ | $76.49_{\pm 2.79}$ | $74.99_{\pm 2.69}$ | $69.15_{\pm 3.65}$ | $68.19_{\pm 2.29}$ | $64.19_{\pm 3.01}$ | $62.04_{\pm 1.1}$ | 23.51 |
| Mistral-7B-CPL | $86.67_{\pm 0.81}$ | $80.98_{\pm 5.42}$ | $77.16_{\pm 4.96}$ | $73.24_{\pm 3.63}$ | $70.05_{\pm 2.5}$ | $67.70_{\pm 3.95}$ | $67.04_{\pm 3.12}$ | $64.11_{\pm 3.68}$ | 22.56 |
| Mistral-7B-CPL + MI | $86.32_{\pm 1.25}$ | $81.00_{\pm 3.2}$ | $77.71_{\pm 2.31}$ | $75.48_{\pm 2.59}$ | $71.92_{\pm 3.09}$ | $71.02_{\pm 2.84}$ | $67.69_{\pm 3.58}$ | $65.48_{\pm 1.97}$ | **20.84** |

Table 6: Accuracy (%) of methods using different LMs after training for each task. We highlight the rows corresponding to our proposed method. The best result in each group is in **bold**. **Results of ConPL are reproduced. Columns $\Delta\downarrow$ present Accuracy drop after learning 8 tasks.

trained on a set of $\tilde{R}^t$ relations. However, for each relation $r$, ConPL defines a set of negative candidate classes $M_r$, so that predictions are made on the set $(\tilde{R}^t \cap M_r)$. This means that the model does not make predictions with all the classes it has learned so far but rather with a predefined subset specific to each relation. While enhancing the performance reported for ConPL, this targeted prediction approach does not align with the practical requirements of CL. In this challenging scenario, each model has to dynamically adapt and make predictions across the expanding set of relations without relying on some fixed set of classes. Therefore, despite its efficacy in controlled evaluations, the ConPL method is impractical for real-world continual learning applications.

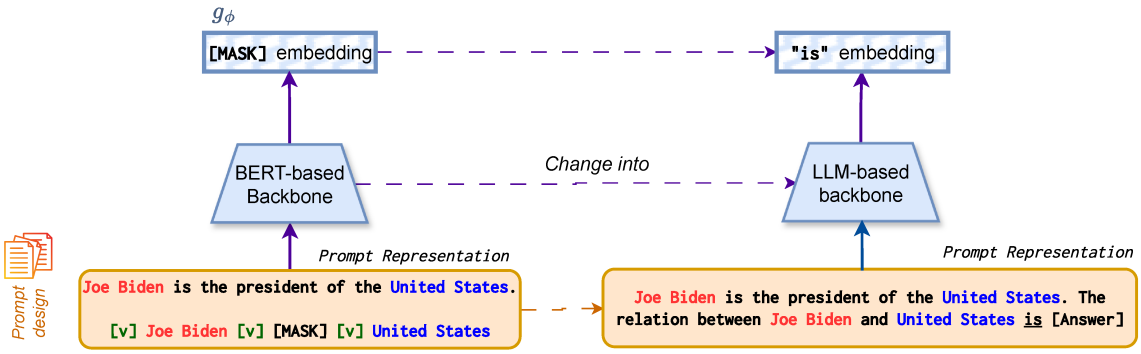


Figure 6: Adapting LLMs for FCRE problems