Bridging Language Gaps: Enhancing Few-Shot Language Adaptation

Anonymous ACL submission

Abstract

The disparity in language resources poses a challenge in multilingual NLP, with high-resource languages benefiting from extensive data, while low-resource languages lack sufficient data for effective training. Our **Co**ntrastive **L**anguage **A**lignment with **P**rompting (CoLAP) method addresses this gap by integrating contrastive learning with cross-lingual representations, facilitating task-specific knowledge transfer from high-resource to lower-resource languages. The primary advantage of our approach is its data efficiency, enabling rapid adaptation to new languages and reducing the need for large labeled datasets. We conduct experiments with multilingual encoder-only and decoder-only language models on natural language understanding tasks, including natural language inference and relation extraction, evaluating performance across both high- and low-resource languages. Our results demonstrate that CoLAP outperforms few-shot cross-lingual transfer baselines and in-context learning, even with limited available data. This effectively narrows the cross-lingual performance gap, contributing to the development of more efficient multilingual NLP techniques.[1]

1 Introduction

The adaptation of pretrained language models (PLMs) to specific downstream tasks is typically resource-intensive, requiring extensive labeled data and computational resources. This becomes particularly challenging in multilingual contexts, especially for languages not represented in the model's pretraining data. These low-resource languages are significantly underrepresented, primarily due to the scarcity of large-scale corpora necessary for self-supervised pretraining and the lack of labeled data required for task-specific fine-tuning (Lauscher et al., 2020; Wu and Dredze, 2020; Yang et al., 2022). Multilingual PLMs, such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and Mistral (Jiang et al., 2023), have been developed to address these challenges. These models are trained across a wide range of languages using self-supervised data within a unified embedding space, facilitating cross-lingual transfer (XLT) of information. They have shown remarkable results in multilingual tasks. Despite their broad versatility, multilingual PLMs often exhibit representation degradation, particularly in low-resource languages (Yang et al., 2022). This degradation stems from the skewed distribution of self-supervised training data, which disproportionately favors high-resource languages. As a result, the representational quality for low-resource languages, and particularly for those not included in the pretraining data, is substantially diminished, leading to a decline in downstream task performance (Winata et al., 2022).

To enhance model representations and downstream task performance, models are commonly fine-tuned using task-specific data. However, acquiring such labeled data at scale is a significant challenge, especially for languages already disadvantaged by insufficient pretraining data (Winata et al., 2022). While multilingual PLMs have advanced performance for these languages compared to monolingual models primarily trained on English data, XLT remains challenging as there are substantial performance disparities between different languages. This disparity becomes more evident when fine-tuning PLMs on task data using high-resource languages and subsequently evaluating the fine-tuned model on other languages (Guo et al., 2023). Few-shot learning techniques address this gap, enabling models to learn effectively from limited data. Recent works in this domain encompass a range of zero-shot cross-lingual transfer (ZS-XLT) and few-shot cross-lingual transfer (FS-XLT) methods, including full fine-tuning (Lauscher et al., 2020), in-context learning (Winata et al., 2022),

[1] Code: https://anonymous.4open.science/r/CoLAP-B2BE.

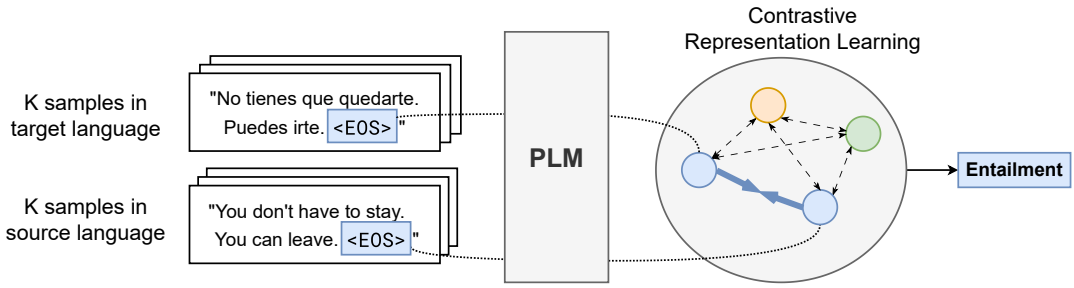

Figure 1: Illustration of our contrastive language alignment with prompting (CoLAP) approach for few-shot cross-lingual transfer applied to the natural language inference task.

prompting (Schick and Schütze, 2021), representation mixup (Yang et al., 2022; Xu et al., 2023), and contrastive learning (Chi et al., 2021). Among these, prompting techniques have shown efficacy in low-resource settings. Prompts express tasks as language modeling problems without introducing new model parameters, often articulated as natural language sentences (Qi et al., 2022; Nie et al., 2023). Despite these advances, full fine-tuning remains standard practice in the field (Zhao and Schütze, 2021; Zhou et al., 2023).

In our study, we propose contrastive language alignment with prompting (CoLAP), addressing the cross-lingual transfer gap by aligning representations of underrepresented languages with those from high-resource languages, especially English. Instead of aligning representations during language modeling, we focus on the application of pretrained models on specific downstream tasks. We hypothesize that the transfer of discriminative task-specific information from English to lower-resource languages can be achieved efficiently without relying on abundant self-supervised or task-specific labeled data. This builds upon prior research by Chi et al. (2021) and Yang et al. (2022), which demonstrated the effectiveness of such alignment in general domain representations and ZS-XLT. We posit that representations for downstream classification tasks are less complex than representations used for language modeling because they only capture information relevant to the downstream objective. As a result, these representations can be transferred more data-efficiently between languages. We evaluate a representation contrastive learning approach ($XRCL$) that aligns source and target language representations based on parallel translations. Additionally, we introduce a class contrastive learning objective ($XCCL$), which aligns representations of instances sharing the same class labels across languages, which does not require parallel translations. This facilitates the data annotation process and reduces costs.

Our experiments focus on natural language understanding tasks, including natural language inference and relation extraction, across 27 languages. We evaluate multilingual encoder-only and decoder-only models in the FS-XLT setting, where models are fine-tuned in a high-resource source language and subsequently adapted to the target language using few-shot learning. Our results demonstrate that CoLAP improves upon strong FS-XLT benchmarks but also outperforms in-context learning, even with limited available data. This methodology aims to narrow the cross-lingual performance gap, thereby extending the benefits of NLP models to a broader range of languages.

Contributions. **1)** We introduce CoLAP, a novel few-shot cross-lingual transfer method that leverages contrastive learning to efficiently transfer knowledge from high- to lower-resource languages. **2)** CoLAP enhances cross-lingual transfer performance for low-resource languages, including those not represented in PLM pretraining. **3)** We propose $XCCL$, a contrastive learning objective that enhances FS-XLT performance without relying on parallel translations. **4)** We propose a few-shot exemplar selection approach based on representation similarity improving data efficiency.

2 Related Work

Cross-lingual transfer involves transferring knowledge from a source language to one or more target languages (Ruder et al., 2019). It is grounded in pretraining language models on multilingual data, embedding languages within a shared universal space to facilitate cross-lingual information transfer. This approach is particularly effective for languages included in the pretraining dataset (Ruder et al., 2019; Ansell et al., 2021). Multilingual PLMs have extended this capability, showing promise in ZS-XLT even for languages not directly seen during pretraining (Devlin et al., 2019;

Conneau et al., 2020). The transfer's efficacy is influenced by the alignment quality of representations between source and target languages (Wu and Dredze, 2020; Cao et al., 2020), which is especially effective for topologically similar languages with large-scale pretraining corpora (Lauscher et al., 2020). Recent advancements in enhancing XLT performance include projecting target language representations into the English space (Yang et al., 2022; Xu et al., 2023), and employing data augmentation (Zheng et al., 2021). Notably, Chi et al. (2021) and Pan et al. (2021) improve cross-lingual representation alignment in their InfoXLM PLM, employing contrastive learning and large-scale text corpora during pretraining. In contrast, our work focuses on data-efficient cross-lingual transfer during task-specific fine-tuning, eliminating the need for an additional cross-lingual alignment step on self-supervised data.

Few-shot Learning and FS-XLT involve training models with a minimal number of labeled examples, challenging models to generalize from this limited data (Nie et al., 2023). Techniques explored for FS-XLT include fine-tuning (Lauscher et al., 2020), in-context learning (Winata et al., 2022), and prompting (Schick and Schütze, 2021). Prompting has shown notable performance improvements, mitigating hyperparameter sensitivity and performance variance issues prevalent in fine-tuning and in-context learning techniques (Zhao et al., 2021; Schmidt et al., 2022; Winata et al., 2022). Enhancements to prompting techniques have been made through the use of multilingual templates (Qi et al., 2022) and label words (Huang et al., 2022; Zhou et al., 2023), further improving their applicability in multilingual settings. Schmidt et al. (2023) note the difficulty in comparing studies due to varying few-shot settings. They suggest that averaging model checkpoints during the few-shot fine-tuning phase provides a robust baseline for FS-XLT. Our study contributes to this line of research by combining prompting techniques with contrastive learning to improve few-shot cross-lingual transfer performance.

3 Methodology

Our methodology is predicated on the hypothesis that the cross-lingual performance gap in downstream tasks for low-resource languages is largely attributable to their underdeveloped representation spaces within PLMs. Building on previous studies, we propose that aligning the representation spaces of low-resource languages with the more robust English representation space can significantly enhance task performance (Yang et al., 2022; Xu et al., 2023). This approach aims to transfer the discriminatory information embedded within the English representation space to these lower-resource languages. Prior research, including the works of Chi et al. (2021) and Yang et al. (2022), supports the effectiveness of this alignment during language modeling. However, these methods often rely on large-scale parallel corpora for extensive pretraining, making them infeasible for low-resource scenarios. To address this, we introduce `CoLAP`, a method specifically designed for few-shot cross-lingual transfer in low-resource contexts. `CoLAP` avoids introducing additional model parameters through prompting, and does not require large-scale labeled or self-supervised corpora. Our approach includes a task-agnostic representation contrastive learning objective, $XRCL$, which aligns multilingual vector representations of translated inputs. Unlike InfoXLM's contrastive learning objective, our method does not use a memory queue for negative pairs or mixup sampling, enhancing resource efficiency (Chi et al., 2021). In addition, we introduce a classification task-specific contrastive learning objective, $XCCL$, designed to transfer discriminative, class-specific features from high-resource to lower-resource languages, alleviating the dependency on parallel training datasets. Importantly, the additional computational complexity introduced by our contrastive learning objectives is dependent only on the number of samples per batch. Both `CoLAP` strategies are model-agnostic, making them applicable to any PLM.

3.1 Prompt-based Training

In prompt-based training approaches, tasks are reformulated as language modeling problems, where the model predicts tokens that serve as reference labels (Schick and Schütze, 2021). Specifically, a template $\mathcal{T}$ is applied to each input x, resulting in a prompted input $x_{\text{prompt}} = \mathcal{T}(x)$. The model then predicts task-specific label tokens based on the context provided by x_{prompt}. For models trained with a causal language modeling objective, we utilize the hidden state of the end-of-sequence token (`<EOS>`) to predict the labels. In contrast, for models trained with masked language modeling objectives, such as XLM-R, at least one `<mask>` token is included in the prompt , which is used to predict

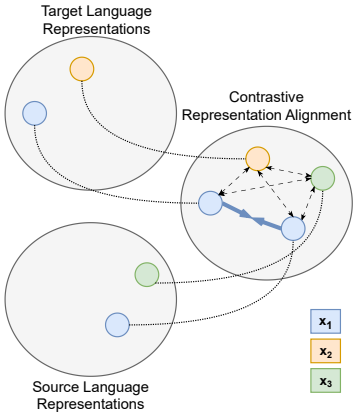

Figure 2: Illustration of cross-lingual contrastive representation alignment using the $XRCL$ objective.

the label tokens. To facilitate this prediction, labels are mapped to tokens in the model's vocabulary, represented as $\mathcal{R} \mapsto \mathcal{V}_{\mathcal{M}}$. Given an input x, the probability of predicting label y is denoted as:

$$p(y \mid x) = p(\texttt{<EOS>} = w_v \mid x_{\text{prompt}}) = \frac{\exp\left(w_v \cdot h_{\texttt{<EOS>}}\right)}{\sum_{v' \in \mathcal{V}} \exp\left(w_{v'} \cdot h_{\texttt{<EOS>}}\right)},$$

where $h_{\texttt{<EOS>}}$ is the hidden vector of the <EOS> token[2], and w_v denotes the pre-softmax vector corresponding to $v \in \mathcal{V}$ (Schick and Schütze, 2021). We employ English label words and language agnostic prompt templates, included in Table 3 (appendix).

3.2 Contrastive Learning Objectives

The **cross-lingual representation contrastive loss** term, abbreviated as $\mathcal{L}_{XRCL}$, aims to align the latent representations of instances in the target language with their counterparts in the source language. To achieve this, we utilize a model $\mathcal{M}$ to generate vector representations $r_i = \mathcal{M}(x_i)$ for each input sequence x_i sampled along with its class label y_i from the training set. We obtain these representations by prompting the PLM and extracting the hidden state of the <EOS> token. We treat the target language training dataset (D_T) and the source language training dataset (D_S) as parallel corpora, meaning that each instance $x_{i,T}$ in D_T has a direct translation $x_{i,S}$ in D_S.

For a given instance representation in the target language $r_{i,T}$, its direct translation in the source language $r_{i,S}$, forms a positive pair. Consequently, all other instances within the target language dataset are treated as negative pairs. This is formulated as follows:

$$r_i^+ = \{r_j \mid j = i, j \in D_S\}$$
$$r_i^- = \{r_j \mid j \neq i, j \in D_S\}$$

The $XRCL$ objective maximizes the similarity between these cross-lingual instance representations (positive pairs) while minimizing their similarity to other instances (negative pairs) (van den Oord et al., 2019). We compute the cross-lingual representation contrastive loss, $\mathcal{L}_{XRCL}$, for each instance representation in the target language $r_{i,T}$ as follows:

$$\mathcal{L}_{XRCL} = \sum_{i=1}^{N} -log \frac{exp\left(\phi(r_{i,T}, r_i^+)/\tau\right)}{exp\left(\phi(r_{i,T}, r_i^-)/\tau\right)},$$

where N is the number of instances in the target language's training set D_T. The parameter τ is a temperature scaling factor, and $\phi(r_i, r_i^-)$ calculates the cosine similarity between representation r_i and each representation in r_i^- using the formula $\sum_{j=1} r_i \cdot r_j / |r_i||r_j|$.

The **cross-lingual class contrastive loss**, denoted as $\mathcal{L}_{XCCL}$, addresses the limitations of related cross-lingual transfer methods, which often assume that contextual information is accurately preserved through direct translations between the source and target languages. In practice, translating instances requires careful adaptation of culture-specific terminology and evaluation metrics, which incurs significant cost in the data collection process (Ponti et al., 2020; Freitag et al., 2022; Winata et al., 2023). In contrast, our $XCCL$ objective does not rely on parallel translations between the source and target languages. Instead, it aligns instances across different languages based on shared class labels. Specifically, for each representation $r_{i,T}$ from the target language training dataset, we construct positive pairs with representations of instances in the source language training dataset that have the same class label. Instances from the source language with different class labels are considered negative pairs. The formulation is as follows:

$$r_i^+ = \{r_j \mid y_j = y_i, j \in D_S\}$$
$$r_i^- = \{r_j \mid y_i \neq y_j, j \in D_S\}$$

This approach aims to maximize the similarity between cross-lingual instance representations within the same class while minimizing similarity with different class representations. The key distinction of $\mathcal{L}_{XCCL}$ from $\mathcal{L}_{XRCL}$ lies in its assumption that class-specific information and can

[2]In masked language models, label tokens are predicted at the positions of the <mask> tokens rather than at the <EOS> token.

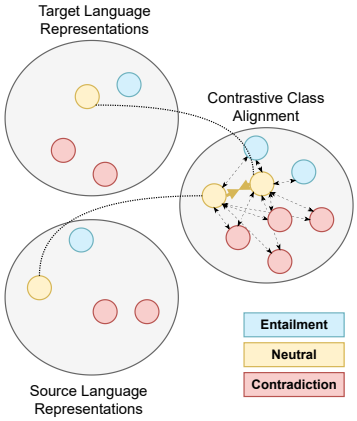

Figure 3: Illustration of cross-lingual class contrastive representation alignment using the $XCCL$ objective.

effectively be leveraged across languages. It posits that embeddings from instances within the same class should exhibit higher similarity compared to those from instances of different classes. The computation of the class-contrastive loss is analogous to $XRCL$, as shown below:

$$\mathcal{L}_{XCCL} = \sum_{i=1}^{N} -log \frac{exp\left(\phi(r_{i,T}, r_i^{+})/\tau\right)}{exp\left(\phi(r_{i,T}, r_i^{-})/\tau\right)}$$

4 Experiments

4.1 Models and Training

We investigate FS-XLT using models that are initially fine-tuned on downstream task data in a high-resource source language, such as English. During this phase, we employ the cross-entropy loss $\mathcal{L}_{CE} = -\log(z_y)$, where z_y represents the post-softmax probability of the vocabulary token corresponding to label y. Subsequently, the model is adapted to the target language under FS-XLT. During this few-shot adaptation phase, we integrate our contrastive loss terms. For the variant incorporating the cross-lingual representation contrastive objective $XRCL$, the total loss is calculated as $\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{XRCL}$. Similarly, for the class contrastive objective variant $XCCL$, the total loss is $\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{XCCL}$.

Training details. Unless specified otherwise, we adhere to training parameters that include a batch size of 64, a learning rate of 2e-5, and the AdamW optimizer (Loshchilov and Hutter, 2017). Models are fine-tuned for 5 epochs on the downstream task in English. For the few-shot adaptation to the target language, we train the models for 10 epochs using data from both the source and target languages. Unlike some related studies, we do not use a dedicated validation set during few-shot fine-tuning. This approach aligns with the few-shot premise, avoiding the need for additional labeled data that a validation set would require (Schmidt et al., 2023). Consequently, all models are trained for a fixed number of epochs on the few-shot data. We argue that this reflects real-world conditions more accurately, as performance is not skewed by the size or quality of a validation dataset.

Models. We evaluate a range of encoder-only and decoder-only PLMs that have been exposed to multilingual text during pretraining, including XLM-R Base (270 million parameters) (Conneau et al., 2020), Gemma 2 (2 billion parameters) (Gemma Team et al., 2024), and Mistral v0.3 (7 billion parameters) (Jiang et al., 2023). XLM-R is trained using full-fine-tuning, while we utilize 4-bit quantization and low-rank adapters (Hu et al., 2021; Dettmers et al., 2023) with $r = 16$ and $alpha = 32$ for the larger Gemma 2 and Mistral models.[3]

Baselines. We include relevant benchmarks proposed in related studies. To ensure fair model comparisons, we integrate relevant benchmarks in our XLT settings, with all models being reimplemented for consistency.

Regular fine-tuning (**FT**) directly predicts class labels from sequence representations, in contrast to prompting approaches that use tokens as reference labels (Lauscher et al., 2020; Schmidt et al., 2022). The model is fine-tuned using the standard cross-entropy loss $\mathcal{L}_{CE}$.

Checkpoint Averaging (**CA**), introduced by Schmidt et al. (2023), is a strong baseline for FS-XLT. It builds upon regular fine-tuning by averaging model weights during few-shot adaptation.

In the prompt-learning from cross-lingual templates (**PCT**) framework Qi et al. (2022) augment model inputs with multilingual prompt templates, optimizing the model using both the $\mathcal{L}_{CE}$ objective for non-augmented inputs and the $\mathcal{L}_{CE'}$ objective for augmented inputs. Additionally, they introduce a consistency loss term based on the Kullback-Leibler divergence between predicted token probabilities. In contrast to CoLAP, PCT is applied during both supervised fine-tuning and few-shot adaptation phases, which requires the target languages to be known and reduces the modularity of the fine-tuned language model.

In-context Learning (**ICL**) incorporates few-shot

[3]To ensure consistency and control for potential exposure to the downstream tasks included in our benchmarks, we do not evaluate instruction fine-tuned model variants.

examples directly into the input prompt, allowing the model to utilize these examples without updating its parameters. This approach significantly extends the prompt length during inference.[4]

4.2 Episode Sampling

In the few-shot task adaptation, episodes are randomly sampled from the training dataset of the target language. Each episode consists of K input sequences along with corresponding labels. The value of K is varied among the set $\{5, 10, 50, 100, 250\}$. Episodes comprise K uniformly sampled instances from the available class labels N_Y. In scenarios where $K \mod N_Y \neq 0$ the remaining instances are randomly selected.

For the computation of our cross-lingual contrastive loss terms, an episode is constructed to include K instances from the training dataset in the target language D_T and an additional K instances from the training dataset in the source language D_S. It is important to note that we assume D_T and D_S to be parallel corpora, meaning that all instances present in the training dataset of the target language D_T are also available in the training dataset of the source language D_S.

4.3 Datasets and Languages

XNLI contains 15 languages, each containing 7,500 examples. In natural language inference, the goal is to determine if a given hypothesis entails, contradicts, or is neutral relative to a premise. Initially annotated in English, the dataset was expanded to other languages using machine translation (Conneau et al., 2018). We assess model performance using the accuracy metric.

AmericasNLI (AmNLI) extends the XNLI dataset by adding 10 indigenous languages from the Americas, categorized as very low-resource languages due to their minimal presence in large-scale text corpora. The dataset is human-translated based on the Spanish XNLI. It includes 750 development and 750 test instances per language, which we utilize as our training and testing sets, respectively (Ebrahimi et al., 2022). Performance evaluation on AmericasNLI is based on accuracy.[5]

MultiTACRED is a multilingual relation extraction dataset covering 12 languages, created via machine translation of English annotations (Hennig et al., 2023). The primary task in sentence-level relation extraction involves categorizing a given set of entities into one of 41 distinct relation types. We utilize instances from the development set to construct few-shot episodes and assess model performance based on accuracy.

5 Results and Discussion

We summarize our results in Table 1. For each language in the datasets, we averaged accuracy metrics across five random seeds.

Our findings confirm our initial hypothesis that discriminative task-specific information can be effectively transferred from English to lower-resource languages using only a few labeled examples. Our CoLAP method consistently improves downstream performance over zero-shot baselines on natural language inference and relation extraction tasks across all evaluated few-shot settings. This improvement is observed for languages included during the pretraining of the PLMs (such as those in XNLI and MultiTACRED) as well as for unseen languages, like those in AmNLI. Furthermore, the performance gains provided by CoLAP are independent of the PLM architecture. Both encoder-only models (e.g., XLM-R) and decoder-only models (e.g., Gemma and Mistral) show increased FS-XLT performance when utilizing CoLAP. In the lowest-resource setting of $K = 5$, CoLAP demonstrates strong performance, outperforming all baseline approaches for the XLM-R and Gemma PLMs and performing on par with PCT for the Mistral PLM. These performance benefits remain consistent across different few-shot settings. Even in the most resource-rich scenario of $K = 250$ exemplars, CoLAP surpasses the performance of its benchmarks, except in the AmNLI dataset for Gemma and Mistral. Even in the high-resource setting of $K = 250$ exemplars, CoLAP exceeds the performance of its benchmarks. When compared to the strongest baseline, PCT (which also uses prompting), CoLAP achieves average performance gains of up to 1.84% for Gemma 2.

Our results demonstrate that prompting approaches, including CoLAP and PCT, outperform traditional fine-tuning methods in the FS-XLT setting. This indicates that regular fine-tuning requires significantly more exemplars to achieve performance comparable to the prompting models on average. While CoLAP is the most data-efficient ap-

[4] Due to context length and memory limitations, we benchmark ICL only for the $K = 5$ few-shot setting and discuss the results in Section 5.

[5] Otomí "oto" has less than 250 samples and therefore is not included in the $K = 250$ setting.

Model	K	XNLI			AmNLI			MultiTACRED		
		XLM-R	GEMMA 2B	MISTRAL 7B	XLM-R	GEMMA 2B	MISTRAL 7B	XLM-R	GEMMA 2B	MISTRAL 7B
FT	0	$73.21_{\pm1.0}$	$69.32_{\pm0.9}$	$61.98_{\pm1.3}$	$36.71_{\pm1.4}$	$40.99_{\pm0.2}$	$37.60_{\pm0.2}$	$24.57_{\pm1.3}$	$14.55_{\pm0.5}$	$10.96_{\pm0.5}$
PCT		$72.97_{\pm1.2}$	$71.81_{\pm0.7}$	$67.68_{\pm1.4}$	$37.98_{\pm0.9}$	$41.27_{\pm0.3}$	$37.37_{\pm0.2}$	$45.87_{\pm2.3}$	$34.16_{\pm0.8}$	$34.05_{\pm1.6}$
CoLAP		$73.15_{\pm1.2}$	$\mathbf{74.35}_{\pm0.7}$	$66.23_{\pm1.2}$	$35.79_{\pm0.7}$	$\mathbf{42.73}_{\pm0.3}$	$37.18_{\pm0.2}$	$\mathbf{51.58}_{\pm2.1}$	$32.48_{\pm0.9}$	$38.73_{\pm0.8}$
FT	5	$71.74_{\pm1.8}$	$70.38_{\pm0.8}$	$62.90_{\pm1.0}$	$39.57_{\pm1.4}$	$41.41_{\pm0.2}$	$37.58_{\pm0.2}$	$42.65_{\pm0.9}$	$17.53_{\pm0.6}$	$13.24_{\pm0.6}$
CA		$73.21_{\pm1.2}$	$70.38_{\pm0.8}$	$62.90_{\pm1.0}$	$37.94_{\pm0.9}$	$41.41_{\pm0.2}$	$37.58_{\pm0.2}$	$38.94_{\pm1.5}$	$17.53_{\pm0.6}$	$13.24_{\pm0.6}$
PCT		$72.52_{\pm1.3}$	$72.45_{\pm0.7}$	$67.94_{\pm1.2}$	$39.11_{\pm1.7}$	$41.99_{\pm0.3}$	$39.46_{\pm0.2}$	$69.01_{\pm1.2}$	$44.00_{\pm0.9}$	$43.35_{\pm1.6}$
CoLAP w/ $XRCL$		$73.56_{\pm1.2}$	$\mathbf{74.59}_{\pm0.7}$	$67.68_{\pm1.1}$	$40.01_{\pm1.4}$	$\mathbf{42.88}_{\pm0.3}$	$39.21_{\pm0.2}$	$\mathbf{69.26}_{\pm2.6}$	$43.47_{\pm1.0}$	$38.73_{\pm1.0}$
CoLAP w/ $XCCL$		$73.04_{\pm1.2}$	$74.54_{\pm0.7}$	$67.58_{\pm0.0}$	$39.71_{\pm1.0}$	$42.84_{\pm0.3}$	$39.20_{\pm0.2}$	$69.18_{\pm2.6}$	$42.91_{\pm1.0}$	$37.96_{\pm1.0}$
FT	10	$72.47_{\pm1.4}$	$70.53_{\pm0.8}$	$63.02_{\pm1.1}$	$42.17_{\pm1.4}$	$41.24_{\pm0.2}$	$37.88_{\pm0.2}$	$43.91_{\pm1.1}$	$18.35_{\pm0.6}$	$13.83_{\pm0.6}$
CA		$73.22_{\pm0.9}$	$70.53_{\pm0.8}$	$63.02_{\pm1.0}$	$40.28_{\pm0.5}$	$41.24_{\pm0.2}$	$37.88_{\pm0.2}$	$40.82_{\pm0.8}$	$18.35_{\pm0.6}$	$13.83_{\pm0.6}$
PCT		$72.92_{\pm0.6}$	$72.49_{\pm0.7}$	$68.11_{\pm1.2}$	$41.26_{\pm0.9}$	$42.41_{\pm0.3}$	$40.01_{\pm0.2}$	$70.19_{\pm3.2}$	$45.52_{\pm1.0}$	$46.63_{\pm1.6}$
CoLAP w/ $XRCL$		$73.69_{\pm0.9}$	$74.52_{\pm0.7}$	$67.47_{\pm1.1}$	$40.69_{\pm1.0}$	$42.36_{\pm0.3}$	$39.29_{\pm0.2}$	$70.16_{\pm3.8}$	$49.12_{\pm0.8}$	$41.98_{\pm0.9}$
CoLAP w/ $XCCL$		$73.15_{\pm0.9}$	$\mathbf{74.55}_{\pm0.7}$	$67.37_{\pm1.1}$	$41.09_{\pm1.2}$	$\mathbf{42.67}_{\pm0.3}$	$39.32_{\pm0.2}$	$\mathbf{70.32}_{\pm3.5}$	$47.43_{\pm0.8}$	$41.94_{\pm0.9}$
FT	50	$73.54_{\pm0.8}$	$73.77_{\pm0.6}$	$65.51_{\pm1.1}$	$\mathbf{44.76}_{\pm0.5}$	$41.68_{\pm0.2}$	$38.23_{\pm0.3}$	$46.88_{\pm1.5}$	$29.45_{\pm0.9}$	$20.68_{\pm0.8}$
CA		$73.55_{\pm0.8}$	$73.35_{\pm0.7}$	$65.25_{\pm1.1}$	$42.31_{\pm0.5}$	$41.66_{\pm0.2}$	$38.08_{\pm0.3}$	$46.22_{\pm0.5}$	$28.44_{\pm0.9}$	$19.88_{\pm0.8}$
PCT		$73.72_{\pm0.5}$	$74.53_{\pm0.6}$	$69.13_{\pm1.2}$	$43.38_{\pm0.6}$	$43.64_{\pm0.3}$	$41.94_{\pm0.2}$	$\mathbf{75.36}_{\pm1.3}$	$62.14_{\pm0.8}$	$63.33_{\pm1.3}$
CoLAP w/ $XRCL$		$73.88_{\pm0.7}$	$\mathbf{75.81}_{\pm0.6}$	$68.95_{\pm1.1}$	$41.98_{\pm0.4}$	$43.14_{\pm0.3}$	$39.84_{\pm0.2}$	$73.98_{\pm2.1}$	$59.76_{\pm0.9}$	$62.01_{\pm1.1}$
CoLAP w/ $XCCL$		$73.81_{\pm0.5}$	$75.73_{\pm0.6}$	$68.78_{\pm1.1}$	$41.98_{\pm0.4}$	$43.21_{\pm0.3}$	$40.02_{\pm0.2}$	$73.98_{\pm2.1}$	$57.90_{\pm0.9}$	$61.31_{\pm1.1}$
FT	100	$73.87_{\pm0.6}$	$74.38_{\pm0.6}$	$67.07_{\pm1.1}$	$\mathbf{45.94}_{\pm0.4}$	$42.76_{\pm0.2}$	$38.57_{\pm0.3}$	$48.87_{\pm1.7}$	$45.97_{\pm1.4}$	$42.11_{\pm1.9}$
CA		$73.95_{\pm0.5}$	$74.40_{\pm0.6}$	$66.91_{\pm1.1}$	$45.39_{\pm0.4}$	$42.70_{\pm0.2}$	$38.50_{\pm0.3}$	$48.41_{\pm1.6}$	$42.82_{\pm1.2}$	$41.06_{\pm1.8}$
PCT		$73.80_{\pm0.5}$	$75.25_{\pm0.6}$	$69.57_{\pm1.1}$	$45.23_{\pm0.4}$	$43.21_{\pm0.3}$	$42.69_{\pm0.2}$	$\mathbf{78.26}_{\pm8.7}$	$71.91_{\pm0.7}$	$74.67_{\pm1.1}$
CoLAP w/ $XRCL$		$74.09_{\pm0.3}$	$75.89_{\pm0.6}$	$69.62_{\pm1.0}$	$44.79_{\pm0.3}$	$43.64_{\pm0.3}$	$40.27_{\pm0.3}$	$77.29_{\pm3.5}$	$74.82_{\pm0.8}$	$73.97_{\pm1.1}$
CoLAP w/ $XCCL$		$74.03_{\pm0.4}$	$\mathbf{75.94}_{\pm0.6}$	$69.43_{\pm1.1}$	$45.18_{\pm0.2}$	$43.59_{\pm0.3}$	$40.25_{\pm0.3}$	$77.06_{\pm3.6}$	$71.29_{\pm0.9}$	$74.18_{\pm1.1}$
FT	250	$73.35_{\pm0.7}$	$74.51_{\pm0.6}$	$67.58_{\pm1.1}$	$48.53_{\pm1.2}$	$44.07_{\pm0.3}$	$41.43_{\pm0.3}$	$55.12_{\pm5.1}$	$82.47_{\pm1.2}$	$89.75_{\pm1.0}$
CA		$73.63_{\pm0.6}$	$74.58_{\pm0.6}$	$67.45_{\pm1.1}$	$48.20_{\pm1.1}$	$43.80_{\pm0.3}$	$40.54_{\pm0.3}$	$53.69_{\pm4.3}$	$84.81_{\pm1.2}$	$88.64_{\pm1.0}$
PCT		$73.03_{\pm0.6}$	$75.74_{\pm0.6}$	$70.25_{\pm1.1}$	$47.73_{\pm1.1}$	$45.28_{\pm0.3}$	$45.21_{\pm0.3}$	$82.19_{\pm4.5}$	$81.60_{\pm0.6}$	$87.98_{\pm0.5}$
CoLAP w/ $XRCL$		$74.00_{\pm0.3}$	$75.68_{\pm0.5}$	$70.37_{\pm1.1}$	$\mathbf{48.64}_{\pm1.2}$	$44.97_{\pm0.3}$	$42.45_{\pm0.2}$	$82.16_{\pm7.7}$	$87.48_{\pm0.7}$	$\mathbf{90.68}_{\pm1.0}$
CoLAP w/ $XCCL$		$73.60_{\pm0.6}$	$\mathbf{75.96}_{\pm0.6}$	$70.51_{\pm1.1}$	$48.44_{\pm0.7}$	$44.90_{\pm0.3}$	$42.56_{\pm0.2}$	$81.99_{\pm7.5}$	$86.25_{\pm0.7}$	$90.39_{\pm1.0}$

Table 1: Average accuracy across all non-English languages on the XNLI, AmNLI, and MultiTACRED datasets. Fine-tuning (FT), checkpoint averaging (CA), prompt-learning from cross-lingual templates (PCT), and CoLAP models are evaluated in few-shot adaptation to the target language following fine-tuning on English task data. The best results per dataset and few-shot setting are highlighted in **bold**.

proach, all evaluated few-shot models provide substantial performance improvements over zero-shot settings, even with as few as $K = 50$ exemplars. For the MultiTACRED dataset, with its large number of relation types, prompting-based approaches like CoLAP or PCT prove to be more suitable.

In Figure 4, we compare the performance of in-context learning with our CoLAP method, focusing on decoder-only PLMs, specifically Gemma 2 and Mistral, in the $K = 5$ few-shot setting. Both approaches start from the same prompting model fine-tuned on English task data. Our CoLAP method consistently outperforms ICL across all natural language understanding benchmarks, achieving average performance improvements of 6.41% for Gemma 2 and 6.93% for Mistral. The performance gap is particularly pronounced on the MultiTACRED dataset. Additionally, we include ablation study results in Table 4 (appendix), where we evaluate the individual components of CoLAP and explore combinations with related approaches.

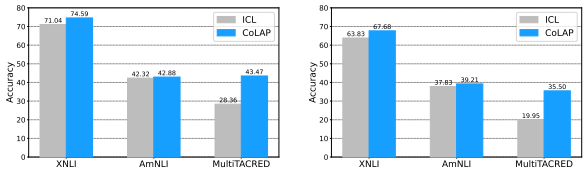

Figure 4: Performance comparison between in-context learning (ICL) and our CoLAP $XRCL$ method on XNLI, AmNLI, and MultiTACRED datasets with $K = 5$ exemplars for Gemma 2 (left) and Mistral (right).

5.1 Enhancing FS-XLT Without Parallel Translations

The results in Table 1 demonstrate that the CoLAP variants utilizing the $XCCL$ objective effectively transfer class-specific features from English to low-resource languages. Notably, the $XCCL$ variant enhances FS-XLT performance without relying on parallel translations of the few-shot examples, thereby simplifying the data annotation process. By eliminating the need for parallel data, $XCCL$ avoids the challenges associated with translating difficult or culturally specific examples and allows for the selection of more natural few-shot instances that share the same class labels. Despite not using parallel translations, the $XCCL$ variant exhibits only a minimal reduction of less than one percent in average few-shot performance ($K > 0$) compared to the $XRCL$ variant across all datasets and settings. While $XRCL$ outperforms $XCCL$ in very low-resource settings, as few as $K = 10$ exemplars are sufficient for $XCCL$ to match or exceed the performance of the $XRCL$ variant.

PLM	Model	K	XNLI			AmNLI			MultiTACRED		
			Random	High	Low	Random	High	Low	Random	High	Low
XLM-R	CoLAP w/ $XRCL$	5	73.56	74.41	74.22	40.01	38.86	38.81	69.26	72.22	69.62
	CoLAP w/ $XCCL$		73.04	74.38	74.12	39.71	39.32	39.33	69.18	71.72	69.36
XLM-R	CoLAP w/ $XRCL$	10	73.69	73.87	74.08	40.69	39.52	40.01	70.16	74.73	70.87
	CoLAP w/ $XCCL$		73.15	74.10	73.85	41.09	39.40	40.33	70.32	74.74	69.82

Table 2: Evaluation of exemplar selection using representation similarity scores, comparing "High" and "Low" similarity exemplars against random selection, with accuracy as the evaluation metric.

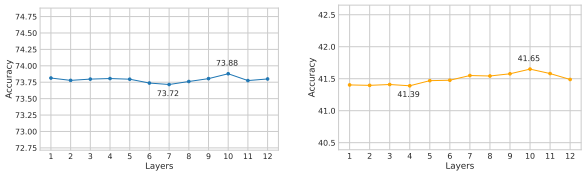

Figure 5: Comparative performance of utilizing different layer representations for contrastive learning in the CoLAP $XRCL$ model with K=50 exemplars. Results are shown for XNLI (left) and AmNLI (right) datasets.

5.2 Layer Selection for Contrastive Learning

To determine the most effective layer for contrastive representation alignment, we conduct experiments with the CoLAP $XRCL$ model, selecting $K = 50$ exemplars. The findings, illustrated in Figure 5, reveal that applying contrastive representation learning at the 10th layer of XLM-R yields the best performance for NLI tasks[6]. This observation suggests that contrastive learning objectives benefit from the general-domain information captured in the middle layers of the model. In contrast, initial layers process low-level syntactic features, whereas the final layers are more focused on extracting discriminative features relevant for the classification task at hand. This finding diverges from that of Chi et al. (2021), who identified the 8th layer as optimal for the contrastive learning objectives used during the pretraining of InfoXLM.

5.3 Few-Shot Exemplar Selection

Building upon the performance of our class contrastive learning objective, we investigate the potential of utilizing class representation similarity for selecting few-shot exemplars. We hypothesize that transferring information from English to other languages via contrastive learning is more sample efficient when selecting exemplars with high intra-class similarity and low inter-class similarity. Employing the XLM-R prompting model, fine-tuned on English task data, we extracted <EOS> representations from English training instances. We then created class prototypes by averaging these embeddings for each class. For each instance x_i, we compute the cosine similarity to its own class prototype P^+ and dissimilarity to other class prototypes P^-, using the formula $s_i = \phi(x_i, P^+) + N_Y - \phi(x_i, P^-)$, where N_Y is the total number of classes. The computational complexity of this procedure is determined by the pairwise similarity calculations, which depend on the number of instances and classes. We selected exemplars based on the highest or lowest similarity scores, focusing on sets of $K = 5$ and $K = 10$. Few-shot episodes consist of these selected exemplars in the source language and their translations in the target language. Table 2 shows that selecting exemplars based on representation similarity enhances efficiency for languages in XNLI and MultiTACRED, improving data efficiency by at least 50%. Notably, with just $K = 5$ exemplars selected based on our similarity criteria, CoLAP exceeds the performance of randomly selected sets of $K = 250$ exemplars for XNLI. However, for languages not included in the pretraining data, selecting exemplars based on similarity in English yields similar or worse results compared to random selection.

6 Conclusion

Our study showcases the effectiveness of contrastive learning in enhancing few-shot adaptation for PLMs. By aligning cross-lingual instance representations, we substantially improve performance in few-shot cross-lingual transfer on natural language inference and relation extraction tasks. In contrast to existing contrastive learning approaches, CoLAP does not require extensive pretraining on parallel multilingual corpora. Despite this, it achieves performance gains even with the smallest set of few-shot exemplars. We show that strong few-shot cross-lingual transfer can be accomplished without relying on parallel translations, which simplifies the data collection process and may reduce costs. Additionally, we demonstrate that selecting few-shot exemplars based on representation similarity enhances data efficiency for languages included in the PLM pretraining corpora.

[6]We note that these results may vary depending on the specific downstream tasks.

7 Limitations

In CoLAP, we introduce an effective strategy to narrow the cross-lingual transfer performance gap in few-shot cross-lingual transfer, tackling the challenge of imbalanced language representation in the pretraining data of current PLMs. Although our approach enhances data efficiency in improving downstream tasks, it does not directly resolve the underlying disparities that contribute to the cross-lingual transfer gap in PLMs.

A key consideration in CoLAP's implementation is the need for translations of the few-shot examples, predicated on the assumption that acquiring (machine) translations, especially in high-resource languages like English, is cost-effective and straightforward. Nonetheless, this requirement introduces an additional step in the process. While the class contrastive learning objective ($XCCL$) in CoLAP diminishes the dependency on parallel corpora, its counterpart, the cross-lingual representation contrastive learning objective ($XRCL$), requires translated instance pairs. Furthermore, the advantage of employing CoLAP with the $XCCL$ objective for not requiring parallel corpora during few-shot training, is limited to classification tasks.

References

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *The Eigth International Conference on Learning Representations, ICLR 2020*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Markus Freitag, David Vilar, David Grangier, Colin Cherry, and George Foster. 2022. A natural diet: Towards improving naturalness of machine translation output. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3340–3353, Dublin, Ireland. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda

Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Yiduo Guo, Yaobo Liang, Dongyan Zhao, Bing Liu, and Nan Duan. 2023. Analyzing and reducing the performance gap in cross-lingual transfer with fine-tuning slow and fast. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4002–4017, Toronto, Canada. Association for Computational Linguistics.

Leonhard Hennig, Philippe Thomas, and Sebastian Möller. 2023. MultiTACRED: A multilingual version of the TAC relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Toronto, Canada. Association for Computational Linguistics.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei, and Houfeng Wang. 2022. Zero-shot cross-lingual transfer of prompt-based tuning with a unified multilingual prompt. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11488–11497, Abu Dhabi, United Arab Emirates. 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, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *The Fifth International Conference on Learning Representations, ICLR 2017*.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2023. Cross-lingual retrieval augmented prompt for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8320–8340, Toronto, Canada. Association for Computational Linguistics.

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2021. Multilingual BERT post-pretraining alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 210–219, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Kunxun Qi, Hai Wan, Jianfeng Du, and Haolan Chen. 2022. Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1910–1923, Dublin, Ireland. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2023. Free lunch: Robust cross-lingual transfer via model checkpoint averaging. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5712–5730, Toronto, Canada. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 777–791, Online only. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. 2023. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3692–3702, Singapore. Association for Computational Linguistics.

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2022. Enhancing cross-lingual transfer by manifold mixup. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. In *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)*, pages 5751–5767, Online. Association for Computational Linguistics.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. In *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)*, pages 3403–3417, Online. Association for Computational Linguistics.

Meng Zhou, Xin Li, Yue Jiang, and Lidong Bing. 2023. Enhancing cross-lingual prompting with dual prompt augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11008–11020, Toronto, Canada. Association for Computational Linguistics.

A Training Details

Training is conducted on NVIDIA A6000, H100 and A100 GPUs. The process of fine-tuning XLM-R with prompt-based techniques on the English XNLI dataset takes approximately 10 hours on A6000 GPUs. For few-shot adaptation employing CoLAP across 14 languages from the XNLI dataset (English excluded) in the $K = 250$ setting, the duration is about 30 minutes.

In Table 3, we display the language agnostic prompting templates used in our CoLAP approach. For NLI tasks, we utilize the label words "Yes", "Maybe", and "No" are used, in line with Schick and Schütze (2021). For the PCT approach, we utilize the multilingual prompt templates introduced by Qi et al. (2022), such as "<premise> Question:<hypothesis>? Answer:<EOS>".

For relation extraction tasks, we replace the entity markers <E1> and <E2> with the respective entities mentioned in the context of the sentence. We adapt the multilingual PCT template to the relation extraction task "<sentence> Relation:<E1>, <E2>? Answer:<EOS>" and machine translate the English template using Google translate (Qi et al., 2022).

For MultiTACRED, we augment the input sentences for benchmarked models by including entity marker tokens <E1>, </E1>, <E2>, and </E2> indicating the position of the entities in the context. We reduce the number relation types from 41 to 31 by merging overlapping class labels for all models evaluated on MultiTACRED. The corresponding label mapping and template files are available at https://anonymous.4open.science/r/CoLAP-B2BE.

To accurately reflect the real-world few-shot performance, we consistently train all models for 10 epochs on the few-shot episodes. We observe that CoLAP models trained on episodes with K greater than 100 instances gain performance when trained up to 50 epochs.

B Ablation Studies

Table 4 displays the performance results of integrating various strategies from related work with our CoLAP method. We evaluate these combined model variants on both small ($K = 5$) and large ($K = 250$) few-shot settings. Our analysis confirms that the contrastive learning objectives are critical to CoLAP's performance, as removing them leads to a noticeable drop in performance. Additionally, combining both the $XRCL$ and $XCCL$ objectives yields better performance compared to using only one, though this requires parallel translations of the few-shot instances. The results also show that checkpoint averaging (CA) enhances robustness, particularly for larger PLMs, improving the stability of CoLAP models. While combining multilingual prompt templates through PCT with CoLAP shows potential, it underperforms compared to other approaches, suggesting that further research is needed to combine these methods.

C Detailed Results

The aggregated results presented in Table 1 are broken down by PLMs and individual languages in Tables 5, 6, 7, 8, 9, 10,11, 12, and 13 for detailed analysis. Given that English serves as the source language in our study, its accuracy scores are excluded from the average performance calculations.

Task	PLM	Template
XNLI, AmNLI	XLM-R	`<premise> <mask>, <hypothesis>`
	Gemma, Mistral	`<premise> <hypothesis>, <EOS>`
MultiTACRED	XLM-R	`<sentence> <E1> <mask> <E2>`
	Gemma, Mistral	`<sentence> <E1> <E2>, <EOS>`

Table 3: Language agnostic prompt templates for encoder-only and decoder-only language models on natural language inference and relation extraction tasks.

Model	K	XNLI			AmNLI			MultiTACRED		
		XLM-R	GEMMA 2B	MISTRAL 7B	XLM-R	GEMMA 2B	MISTRAL 7B	XLM-R	GEMMA 2B	MISTRAL 7B
CoLAP w/ $XRCL$	5	73.56	74.59	67.68	40.01	42.88	39.21	69.26	43.47	35.50
w/o $XRCL$		72.89	72.48	67.34	39.87	40.33	38.83	68.90	41.05	35.09
w/ $XCCL$		72.60	74.42	67.59	39.78	42.70	39.29	69.28	39.79	35.39
w/ CA		73.75	74.58	67.65	39.03	42.60	39.14	64.79	43.41	38.23
w/ PCT		72.95	74.22	67.32	38.16	42.05	38.91	69.87	43.18	30.67
CoLAP w/ $XRCL$	250	74.00	75.68	70.37	48.64	44.97	42.45	82.16	87.48	90.80
w/o $XRCL$		73.72	73.83	70.18	41.74	43.68	41.95	81.95	83.99	89.97
w/ $XCCL$		73.78	76.03	70.59	48.28	44.40	42.54	82.02	87.90	90.77
w/ CA		73.89	75.79	70.75	48.11	44.84	42.31	81.62	87.26	87.35
w/ PCT		73.45	75.38	70.20	47.68	44.57	42.13	84.97	86.38	87.15

Table 4: Model variants with (w/) or without (w/o) indicated architectural changes. Results that improve performance over the CoLAP variant with $XRCL$ are underlined.

Model	K	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg.
FT	0	84.07	71.38	77.25	76.05	75.42	78.66	77.23	69.22	75.69	65.10	72.33	72.62	65.66	74.75	73.59	73.21
PCT		85.31	72.96	73.01	73.00	73.00	73.02	73.02	72.93	73.00	72.86	72.96	72.96	72.89	72.98	72.99	72.97
CoLAP		84.37	72.00	77.23	75.77	75.93	78.74	77.23	69.18	75.67	65.17	71.62	72.46	65.41	73.81	73.95	73.16
FT	5	84.07	72.04	76.53	74.67	74.50	76.81	76.07	66.73	74.64	61.42	70.68	71.15	63.37	72.81	72.90	71.74
CA		84.07	72.33	77.57	75.66	75.69	78.2	77.16	69.57	75.83	63.85	71.66	72.35	66.81	74.06	74.26	73.21
PCT		85.31	72.38	77.03	74.61	75.21	78.36	77.14	67.88	75.76	59.31	72.26	71.38	65.03	74.49	74.46	72.52
CoLAP w/ XRCL		84.37	71.71	78.54	76.47	76.44	79.28	77.56	69.83	75.82	64.25	73.19	72.17	65.63	74.84	74.05	73.56
CoLAP w/ XCCL		84.37	71.84	77.74	76.17	76.00	78.33	76.87	68.94	75.23	63.99	72.01	72.10	65.85	73.30	74.23	73.04
FT	10	84.07	72.13	76.72	74.42	74.91	78.15	75.82	68.96	74.28	62.53	70.00	72.06	66.5	74.02	74.14	72.47
CA		84.07	72.27	77.19	75.20	75.45	78.75	76.68	70.04	75.27	64.0	70.97	72.71	67.13	74.75	74.70	73.22
PCT		85.31	72.58	77.08	75.78	75.60	78.55	77.09	68.79	75.41	61.57	71.05	70.83	66.57	74.86	75.05	72.91
CoLAP w/ XRCL		84.37	73.35	78.41	75.89	76.19	78.97	77.10	70.59	76.27	64.60	71.12	72.38	66.66	74.74	75.34	73.69
CoLAP w/ XCCL		84.37	72.82	77.40	75.90	75.79	78.51	76.66	69.80	74.98	64.37	70.64	71.58	66.71	74.32	74.63	73.15
FT	50	84.07	72.11	77.37	75.89	75.37	78.83	77.10	69.28	75.70	65.55	72.31	72.59	67.43	75.31	74.68	73.54
CA		84.07	71.86	77.36	75.77	75.54	78.89	76.96	69.28	75.49	65.70	72.27	72.92	67.83	75.14	74.65	73.55
PCT		85.31	72.62	78.00	76.20	75.73	78.79	77.95	69.68	76.10	63.71	73.11	72.22	66.66	75.56	75.82	73.72
CoLAP w/ XRCL		84.37	72.00	78.23	75.91	76.41	79.21	77.86	70.22	76.29	65.92	72.17	72.26	67.28	74.56	76.06	73.88
CoLAP w/ XCCL		84.37	72.54	77.88	76.01	75.95	79.31	77.38	70.12	76.09	65.54	72.44	72.77	67.11	74.69	75.46	73.81
FT	100	84.07	72.71	77.29	76.31	75.64	79.02	77.27	70.07	75.58	66.01	73.07	72.99	67.54	75.82	74.85	73.87
CA		84.07	72.50	77.37	76.40	75.74	79.15	77.32	69.97	75.80	65.98	73.20	73.01	67.88	76.11	74.86	73.95
PCT		85.31	72.46	77.60	75.92	76.04	78.66	78.14	69.61	76.13	64.31	73.48	72.16	66.72	76.29	75.71	73.80
CoLAP w/ XRCL		84.37	72.52	78.29	76.49	76.78	79.08	78.18	70.26	76.29	65.71	72.81	72.83	67.54	75.34	75.2	74.09
CoLAP w/ XCCL		84.37	72.56	78.09	76.81	76.35	79.11	77.73	70.23	76.08	65.76	72.81	73.04	67.19	75.69	74.93	74.03
FT	250	84.07	71.87	76.89	75.69	75.00	78.26	76.96	69.64	75.41	65.28	72.69	72.93	66.70	75.27	74.29	73.35
CA		84.07	72.16	76.99	76.00	75.19	78.46	77.07	70.08	75.76	65.58	72.94	73.29	67.15	75.56	74.59	73.63
PCT		85.31	72.00	76.61	75.16	74.94	77.97	76.83	69.50	75.11	63.87	73.33	71.45	65.35	75.18	75.12	73.03
CoLAP w/ XRCL		84.37	72.74	77.71	76.49	76.13	79.03	78.01	69.69	75.82	65.91	73.46	72.95	66.76	75.79	75.47	74.00
CoLAP w/ XCCL		84.37	72.16	77.38	75.93	75.50	78.55	77.19	69.57	75.14	65.94	73.02	72.99	66.44	75.37	75.15	73.60

Table 5: Average accuracy per language on the XNLI dataset for XLM-R.

Model	K	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg.
FT	0	87.76	57.74	76.61	77.29	73.71	81.34	82.06	61.22	71.80	62.38	57.23	71.62	53.55	73.17	70.70	69.32
PCT		88.76	68.18	75.39	78.84	75.11	80.56	79.32	69.80	74.59	62.50	68.56	67.96	54.89	72.53	77.09	71.81
CoLAP		88.32	74.51	80.06	78.32	79.14	84.15	81.86	68.26	77.92	64.15	70.52	70.64	59.10	76.19	76.11	74.35
FT	5	87.76	61.57	76.90	77.75	74.28	81.31	81.22	64.12	73.47	63.72	59.82	71.43	54.43	73.99	71.30	70.38
CA		87.76	61.57	76.90	77.75	74.28	81.31	81.22	64.12	73.47	63.72	59.82	71.43	54.43	73.99	71.30	70.38
PCT		88.76	69.76	75.95	79.52	75.35	80.84	78.85	70.43	75.73	62.98	69.26	69.01	55.68	73.39	77.55	72.45
CoLAP w/ XRCL		88.32	74.64	79.72	78.65	79.00	84.01	81.92	68.86	77.85	64.18	70.71	71.14	60.67	76.28	76.61	74.59
CoLAP w/ XCCL		88.32	74.77	79.86	78.65	79.13	84.12	82.05	68.55	77.77	64.28	70.49	70.94	60.34	76.53	76.09	74.54
FT	10	87.76	61.76	76.85	77.69	74.44	81.54	82.03	63.93	72.75	63.16	61.49	71.50	55.10	73.77	71.37	70.53
CA		87.76	61.76	76.85	77.69	74.44	81.54	82.03	63.93	72.75	63.16	61.49	71.50	55.10	73.77	71.37	70.53
PCT		88.76	69.25	76.07	79.34	75.59	80.95	79.41	70.50	75.28	63.07	69.63	69.17	55.28	73.63	77.74	72.49
CoLAP w/ XRCL		88.32	74.68	79.78	78.46	79.25	83.82	82.13	68.73	77.62	64.21	70.58	70.64	60.94	75.80	76.69	74.52
CoLAP w/ XCCL		88.32	74.73	80.02	78.66	79.37	83.96	81.96	68.46	77.94	63.93	70.85	70.60	60.62	76.20	76.44	74.55
FT	50	87.76	72.65	79.19	78.44	77.54	81.91	80.89	67.43	76.28	64.27	68.94	73.01	60.53	75.39	76.26	73.77
CA		87.76	71.00	79.16	78.38	77.28	81.77	81.03	67.01	75.97	64.16	67.91	72.93	59.35	75.13	75.88	73.35
PCT		88.76	70.92	78.66	80.41	77.75	81.45	81.11	72.41	76.85	64.86	72.80	71.73	59.65	75.90	78.92	74.53
CoLAP w/ XRCL		88.32	74.69	80.51	79.15	78.53	84.01	81.89	72.34	77.75	67.11	72.28	73.33	64.56	77.01	78.12	75.81
CoLAP w/ XCCL		88.32	75.57	79.70	79.39	79.52	83.73	82.67	71.15	77.75	65.72	72.63	73.54	64.42	77.14	77.25	75.73
FT	100	87.76	74.44	79.67	79.97	78.08	81.86	82.30	69.51	76.95	63.86	69.03	71.88	63.43	73.33	77.05	74.38
CA		87.76	74.21	79.84	79.78	78.19	81.86	82.38	69.75	76.95	64.13	69.07	72.20	62.93	73.51	76.83	74.40
PCT		88.76	71.78	79.48	80.60	78.38	82.31	81.25	73.43	78.01	66.25	74.02	73.06	58.98	76.52	79.50	75.25
CoLAP w/ XRCL		88.32	75.31	80.14	79.97	79.58	82.86	82.16	72.42	77.03	66.62	72.81	74.12	63.95	77.28	78.18	75.89
CoLAP w/ XCCL		88.32	76.05	80.46	80.06	79.14	83.15	82.49	72.61	77.48	66.50	72.40	74.09	63.70	77.43	77.59	75.94
FT	250	87.76	74.61	79.17	79.60	78.08	82.24	81.16	70.50	76.92	64.88	70.62	72.27	62.17	74.42	76.51	74.51
CA		87.76	74.84	79.22	79.61	78.14	82.28	81.07	70.62	77.10	65.09	70.42	72.15	62.60	74.38	76.57	74.58
PCT		88.76	72.33	79.89	80.83	79.64	82.69	81.97	74.05	78.62	65.68	74.10	73.95	59.97	77.00	79.68	75.74
CoLAP w/ XRCL		88.32	75.03	79.35	79.93	78.91	83.00	81.94	72.58	77.35	66.24	72.69	73.73	63.35	77.34	78.14	75.68
CoLAP w/ XCCL		88.32	76.00	80.12	80.14	80.00	83.39	81.60	72.42	77.46	66.40	72.75	73.70	63.99	77.51	77.94	75.96

Table 6: Average accuracy per language on the XNLI dataset for Gemma 2 2B.

Model	K	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg.
FT	0	89.22	60.38	68.00	76.37	63.29	77.33	75.51	56.51	68.10	40.96	52.77	56.95	49.38	62.46	59.66	61.98
PCT		90.32	64.71	80.66	79.20	58.50	82.63	81.32	58.86	79.42	42.34	60.90	61.82	51.90	66.95	78.26	67.68
CoLAP		90.12	64.39	77.49	76.99	61.84	78.46	77.94	59.38	76.13	40.78	62.02	57.86	52.34	65.37	76.29	66.23
FT	5	89.22	61.37	68.55	76.07	63.73	77.65	76.62	57.99	70.65	40.99	54.38	57.66	50.13	63.15	61.72	62.90
CA		89.22	61.37	68.55	76.07	63.73	77.65	76.62	57.99	70.65	40.99	54.38	57.66	50.13	63.15	61.72	62.90
PCT		90.32	65.66	80.61	79.29	58.03	82.75	81.44	58.85	79.50	42.72	61.48	61.98	53.34	67.10	78.37	67.94
CoLAP w/ XRCL		90.12	66.43	78.79	78.72	64.01	79.34	78.71	62.12	77.97	41.56	64.15	59.36	54.47	65.41	76.54	67.68
CoLAP w/ XCCL		90.12	66.51	78.75	78.75	63.85	79.25	78.60	61.83	77.96	41.41	63.98	58.94	54.33	65.54	76.47	67.58
FT	10	89.22	61.31	67.41	76.88	64.14	77.52	76.06	57.45	71.03	40.90	54.64	57.83	50.51	63.15	63.43	63.02
CA		89.22	61.31	67.41	76.88	64.14	77.52	76.06	57.45	71.03	40.90	54.64	57.83	50.51	63.15	63.43	63.02
PCT		90.32	64.85	81.29	79.28	57.61	82.83	81.37	59.49	79.96	42.85	61.02	62.92	53.41	67.71	78.91	68.11
CoLAP w/ XRCL		90.12	66.46	78.93	77.03	64.76	78.81	78.19	61.19	77.59	42.06	63.78	58.37	54.66	65.53	77.27	67.47
CoLAP w/ XCCL		90.12	66.22	78.93	76.86	64.71	78.85	77.95	60.87	77.54	41.98	63.80	58.15	54.48	65.49	77.34	67.37
FT	50	89.22	63.27	69.63	78.85	65.97	79.93	77.99	61.27	72.50	41.15	58.14	59.11	53.15	63.92	72.26	65.51
CA		89.22	63.23	68.93	78.74	65.82	79.69	77.95	61.02	72.09	41.19	57.96	59.00	52.61	63.90	71.35	65.25
PCT		90.32	65.62	82.07	79.99	60.56	83.01	81.32	61.20	80.92	45.13	61.94	63.45	55.03	67.70	79.89	69.13
CoLAP w/ XRCL		90.12	67.63	79.74	78.78	66.90	80.29	79.47	64.11	79.24	42.34	65.14	61.10	55.84	66.39	78.37	68.95
CoLAP w/ XCCL		90.12	67.29	79.60	78.58	66.63	80.04	79.38	63.82	79.07	42.35	65.14	60.85	55.60	66.35	78.22	68.78
FT	100	89.22	65.17	74.16	79.16	66.99	81.45	79.66	61.71	75.37	42.61	58.91	60.76	53.01	65.36	74.69	67.07
CA		89.22	64.63	73.33	79.22	66.83	81.34	79.56	61.80	75.40	42.36	58.88	60.79	53.09	65.30	74.23	66.91
PCT		90.32	66.10	82.10	79.67	61.87	83.42	82.13	61.80	80.65	46.63	61.82	64.17	55.23	67.99	80.45	69.57
CoLAP w/ XRCL		90.12	68.68	80.42	79.18	67.92	80.93	79.99	65.03	79.63	42.77	65.53	61.83	56.53	67.15	79.05	69.62
CoLAP w/ XCCL		90.12	68.18	80.06	79.01	67.82	80.79	79.88	64.65	79.34	42.67	65.48	61.76	56.17	67.26	78.97	69.43
FT	250	89.22	65.75	76.97	78.70	68.21	81.18	80.26	61.91	75.15	43.07	59.35	62.21	52.16	65.86	75.40	67.58
CA		89.22	65.47	76.55	78.82	67.79	81.21	80.34	61.63	75.12	42.88	58.80	62.20	52.25	65.63	75.56	67.45
PCT		90.32	66.92	82.97	79.88	64.15	83.71	81.45	61.80	81.33	48.23	63.20	64.71	55.39	68.81	80.91	70.25
CoLAP w/ XRCL		90.12	69.25	81.75	80.27	68.58	81.86	80.46	65.35	80.71	43.62	65.74	63.19	57.22	67.29	79.82	70.37
CoLAP w/ XCCL		90.12	69.40	81.88	80.10	68.83	81.98	80.29	65.60	80.87	43.87	65.90	63.54	57.10	67.75	80.08	70.51

Table 7: Average accuracy per language on the XNLI dataset for Mistral 7B.

Model	K	aym	bzd	cni	gn	hch	nah	oto	quy	shp	tar	Avg.
FT	0	36.80	37.60	36.27	38.13	37.20	37.94	35.96	38.27	35.73	35.20	36.91
PCT		38.42	38.16	38.02	37.89	36.96	40.81	38.93	36.82	36.29	37.49	37.98
CoLAP		35.47	35.47	34.93	36.53	34.80	37.67	35.83	36.53	35.07	35.60	35.79
FT	5	39.17	40.53	40.37	38.93	38.07	40.07	40.17	40.30	40.63	37.47	39.57
CA		37.60	38.77	39.27	37.27	36.57	39.60	38.34	36.77	38.83	36.43	37.95
PCT		37.64	40.87	38.58	39.64	38.87	39.48	40.02	37.62	39.27	39.14	39.11
CoLAP w/ XRCL		38.77	40.69	39.28	41.57	38.88	40.11	40.24	40.29	39.68	40.59	40.01
CoLAP w/ XCCL		38.77	40.51	39.17	40.69	37.92	41.60	40.43	39.33	38.96	39.73	39.71
FT	10	41.43	40.30	42.63	43.60	39.77	43.36	42.91	41.63	43.80	42.30	42.17
CA		39.80	39.53	40.77	40.70	38.03	42.85	41.14	39.37	41.00	39.57	40.28
PCT		41.32	41.05	40.52	42.09	39.24	42.70	41.69	41.16	41.32	41.50	41.26
CoLAP w/ XRCL		38.93	39.39	40.03	42.21	39.95	42.20	41.26	40.32	40.96	41.65	40.69
CoLAP w/ XCCL		39.81	40.43	40.53	41.55	40.19	42.14	41.79	41.04	41.15	42.29	41.09
FT	50	45.37	44.43	44.93	47.53	41.67	45.46	43.72	45.50	45.57	43.43	44.76
CA		41.63	42.60	44.13	44.37	39.07	43.63	41.71	40.83	43.8	41.30	42.31
PCT		43.04	42.86	45.39	45.58	40.19	45.67	43.75	41.68	44.08	41.52	43.38
CoLAP w/ XRCL		39.17	41.81	41.31	43.65	40.69	42.49	42.38	39.57	42.75	42.64	41.65
CoLAP w/ XCCL		39.92	42.11	41.79	44.53	41.15	43.22	41.55	40.03	42.77	42.75	41.98
FT	100	47.37	47.10	48.47	48.57	40.73	48.27	44.39	43.77	47.4	43.30	45.94
CA		46.40	46.03	45.60	48.20	42.83	45.66	44.05	45.67	46.63	42.80	45.39
PCT		44.99	47.79	46.03	47.20	41.26	47.43	45.62	43.87	44.70	43.44	45.23
CoLAP w/ XRCL		44.37	44.53	47.55	48.16	42.88	45.07	43.69	43.97	45.49	42.13	44.78
CoLAP w/ XCCL		44.83	45.68	47.79	48.13	42.93	46.50	43.32	44.29	45.36	43.01	45.18
FT	250	48.70	52.50	50.37	50.47	43.47	50.51	–	48.17	47.73	44.87	48.53
CA		49.20	50.90	50.53	49.77	42.87	50.51	–	47.13	48.63	44.27	48.20
PCT		47.20	50.93	48.26	50.96	41.25	51.47	–	47.46	46.48	45.57	47.73
CoLAP w/ XRCL		48.69	51.23	51.69	51.83	43.03	48.95	–	49.46	46.73	46.13	48.64
CoLAP w/ XCCL		48.56	51.76	50.88	50.11	42.72	50.49	–	48.69	46.69	46.08	48.44

Table 8: Average accuracy per language on the AmNLI dataset for XLM-R.

Model	K	aym	bzd	cni	gn	hch	nah	oto	quy	shp	tar	Avg.
FT	0	42.00	40.93	40.67	42.93	38.27	43.63	38.37	40.67	41.87	40.53	40.99
PCT		39.73	42.53	39.07	46.53	36.53	44.99	41.58	41.07	42.13	38.53	41.27
CoLAP		42.00	44.40	42.13	48.00	38.80	47.97	40.37	42.27	42.00	39.33	42.73
FT	5	42.24	40.53	40.16	45.09	38.32	44.04	38.66	41.41	42.64	40.96	41.41
CA		42.24	40.53	40.16	45.09	38.32	44.04	38.66	41.41	42.64	40.96	41.41
PCT		42.80	43.11	40.04	47.60	36.00	45.44	40.78	41.07	43.24	39.87	41.99
CoLAP w/ XRCL		41.79	43.71	42.21	46.45	39.97	47.67	40.78	43.47	42.99	39.76	42.88
CoLAP w/ XCCL		42.48	43.84	42.61	46.80	39.07	47.59	41.07	42.27	42.32	40.37	42.84
FT	10	41.95	40.69	40.93	43.28	38.24	44.31	38.98	40.88	42.37	40.75	41.24
CA		41.95	40.69	40.93	43.28	38.24	44.31	38.98	40.88	42.37	40.75	41.24
PCT		41.91	44.36	41.87	47.47	35.96	46.70	41.31	40.36	43.38	40.76	42.41
CoLAP w/ XRCL		41.84	43.39	41.60	46.24	39.47	47.51	39.97	42.19	42.16	39.28	42.36
CoLAP w/ XCCL		42.53	43.15	42.29	46.59	39.63	47.18	40.48	43.25	42.32	39.23	42.67
FT	50	42.40	42.21	41.39	44.85	38.64	43.90	38.88	41.33	43.41	39.79	41.68
CA		42.67	42.11	41.65	44.21	38.99	43.69	39.20	41.28	42.93	39.87	41.66
PCT		42.53	45.56	42.84	47.82	38.04	46.70	41.93	44.58	45.11	41.24	43.64
CoLAP w/ XRCL		43.20	45.04	43.76	47.01	39.39	47.40	38.77	43.04	43.65	40.19	43.14
CoLAP w/ XCCL		43.47	44.83	42.88	46.75	40.88	47.05	39.44	43.25	43.65	39.89	43.21
FT	100	43.87	44.72	43.39	45.73	39.04	43.98	39.63	41.63	43.47	42.13	42.76
CA		43.33	44.51	42.91	46.48	38.69	44.17	40.61	41.44	43.23	41.68	42.70
PCT		42.27	44.67	43.16	47.69	37.47	46.12	41.44	42.53	44.89	41.91	43.21
CoLAP w/ XRCL		43.68	46.45	43.15	46.77	40.27	47.45	40.32	43.07	43.92	41.33	43.64
CoLAP w/ XCCL		43.31	45.49	43.81	47.44	40.35	47.34	40.61	42.83	43.63	41.09	43.59
FT	250	43.92	47.55	44.03	47.09	40.05	45.83	–	43.12	44.00	41.07	44.07
CA		42.56	47.07	44.21	46.99	40.05	45.39	–	42.88	43.79	41.25	43.80
PCT		44.09	47.24	43.87	49.56	39.11	48.19	–	44.93	46.93	43.64	45.28
CoLAP w/ XRCL		45.20	48.11	44.45	48.03	40.40	49.40	–	42.85	43.71	42.61	44.97
CoLAP w/ XCCL		45.60	47.31	44.99	47.81	39.73	48.08	–	44.11	44.37	42.08	44.90

Table 9: Average accuracy per language on the AmNLI dataset for Gemma 2 2B.

Model	K	aym	bzd	cni	gn	hch	nah	oto	quy	shp	tar	Avg.
FT	0	37.47	37.60	38.40	37.07	36.93	42.68	36.10	37.07	37.87	34.80	37.60
PCT		36.93	38.00	36.13	38.40	35.47	39.57	36.90	37.33	38.67	36.27	37.37
CoLAP		37.33	38.00	36.27	37.47	36.00	39.84	37.30	38.40	37.47	33.73	37.18
FT	5	37.68	36.51	38.32	36.40	35.47	42.09	36.23	38.43	39.63	35.04	37.58
CA		37.68	36.51	38.32	36.40	35.47	42.09	36.23	38.43	39.63	35.04	37.58
PCT		39.80	38.53	40.07	39.93	38.20	41.60	38.24	39.87	41.87	36.53	39.46
CoLAP w/ XRCL		40.93	39.49	38.16	41.49	37.15	41.52	38.13	40.00	40.08	35.17	39.21
CoLAP w/ XCCL		40.51	38.93	37.84	41.44	37.63	41.19	38.48	40.08	40.03	35.87	39.20
FT	10	38.72	38.08	38.69	36.80	36.03	41.71	35.43	37.81	39.95	35.55	37.88
CA		38.72	38.08	38.69	36.80	36.03	41.71	35.43	37.81	39.95	35.55	37.88
PCT		39.73	39.67	39.47	42.53	37.40	44.04	39.04	39.33	41.87	37.00	40.01
CoLAP w/ XRCL		40.11	39.17	37.76	40.83	37.52	42.06	39.39	40.80	40.69	34.59	39.29
CoLAP w/ XCCL		40.27	39.36	38.21	40.59	36.91	41.98	39.04	40.67	40.69	35.52	39.32
FT	50	40.24	38.61	36.80	37.63	35.65	42.36	35.67	38.53	42.48	34.37	38.23
CA		39.87	38.67	36.27	37.68	34.99	41.90	35.67	38.53	42.32	34.88	38.08
PCT		41.67	42.60	41.93	43.73	39.07	43.97	41.78	41.53	41.93	41.20	41.94
CoLAP w/ XRCL		42.13	40.24	38.32	42.59	37.87	40.84	39.09	40.77	41.79	34.72	39.84
CoLAP w/ XCCL		42.29	40.43	38.43	42.67	37.65	41.65	39.06	40.96	42.43	34.61	40.02
FT	100	41.28	37.44	37.81	38.69	33.89	41.98	36.28	38.96	43.12	36.27	38.57
CA		41.12	38.19	37.60	38.43	33.87	41.79	35.94	38.88	43.41	35.79	38.50
PCT		43.29	42.98	40.93	44.93	38.89	46.34	42.11	42.27	43.07	42.09	42.69
CoLAP w/ XRCL		41.39	41.68	39.23	43.57	38.03	42.49	39.30	39.63	42.37	35.01	40.27
CoLAP w/ XCCL		41.73	41.60	39.09	43.17	38.13	42.47	39.22	39.71	42.51	34.91	40.25
FT	250	43.23	41.68	41.12	43.36	35.95	43.01	–	42.43	44.48	37.63	41.43
CA		42.69	40.64	40.00	41.60	35.41	42.49	–	40.96	44.27	36.77	40.54
PCT		48.44	45.20	43.91	47.73	39.91	48.60	–	45.16	43.78	44.18	45.21
CoLAP w/ XRCL		44.53	43.39	41.25	44.69	39.63	43.58	–	42.96	43.55	38.45	42.45
CoLAP w/ XCCL		44.64	43.55	41.97	45.39	39.57	43.39	–	42.59	43.44	38.51	42.56

Table 10: Average accuracy per language on the AmNLI dataset for Mistral 7B.

Model	K	en	ar	de	es	fi	fr	hi	hu	ja	pl	ru	tr	zh	Avg.
FT	0	85.18	23.20	26.32	30.96	25.61	34.29	22.8	24.80	12.61	23.60	29.58	17.20	23.90	24.57
PCT		87.38	40.40	49.39	54.39	50.81	53.88	38.00	46.00	41.44	50.00	52.50	33.60	40.00	45.87
CoLAP		87.94	55.60	48.99	57.74	50.41	60.41	47.60	46.40	49.55	49.20	60.83	40.00	52.20	51.58
FT	5	85.18	42.88	44.37	45.44	43.50	45.55	40.16	41.52	35.41	44.08	45.42	39.20	44.29	42.65
CA		85.18	37.76	41.13	45.10	40.41	45.06	36.32	34.88	31.44	38.24	43.08	32.72	41.07	38.93
PCT		87.38	69.80	71.56	73.01	69.61	73.27	67.10	71.60	64.64	66.80	71.56	65.50	63.66	69.01
CoLAP w/ XRCL		87.94	71.60	72.15	73.05	70.00	71.76	66.24	70.64	63.96	69.36	70.25	66.96	65.17	69.26
CoLAP w/ XCCL		87.94	72.08	72.23	72.72	70.41	71.59	66.00	70.40	64.14	69.04	69.42	66.88	65.27	69.18
FT	10	85.18	44.80	45.43	46.03	45.28	47.18	43.04	42.16	36.49	44.8	45.92	39.60	46.15	43.91
CA		85.18	39.76	43.08	44.69	41.30	45.71	39.68	36.96	33.87	40.80	44.50	35.84	43.61	40.82
PCT		87.38	68.20	73.28	73.54	69.82	73.78	67.50	72.40	67.79	70.30	72.29	67.50	65.85	70.19
CoLAP w/ XRCL		87.94	69.36	74.90	74.90	71.14	72.33	66.00	71.12	62.61	71.76	71.08	68.96	67.80	70.16
CoLAP w/ XCCL		87.94	69.84	74.74	74.90	71.14	72.82	65.84	71.12	63.42	72.32	71.17	68.96	67.61	70.32
FT	50	85.18	46.64	48.91	48.70	46.26	50.45	44.72	45.92	40.72	48.40	48.50	43.04	50.34	46.88
CA		85.18	46.0	47.13	50.38	45.85	50.2	44.64	43.76	40.18	46.96	48.42	40.96	50.15	46.22
PCT		87.38	74.10	77.73	80.23	76.02	78.16	72.90	75.20	70.61	75.80	77.60	72.80	73.17	75.36
CoLAP w/ XRCL		87.94	75.12	76.11	79.50	76.18	78.94	68.00	74.56	66.22	75.68	75.58	70.96	71.41	74.02
CoLAP w/ XCCL		87.94	75.12	75.71	79.33	76.26	78.86	68.48	74.64	66.22	75.44	75.08	71.28	71.32	73.98
FT	100	85.18	48.88	49.80	50.21	48.94	50.86	48.80	48.40	42.70	49.84	49.67	46.16	52.20	48.87
CA		85.18	48.72	50.28	49.71	47.64	50.94	46.88	47.68	42.70	49.52	50.00	45.12	51.71	48.41
PCT		87.38	76.60	81.68	83.16	78.66	80.31	76.60	78.50	72.30	78.70	81.56	75.90	75.12	78.26
CoLAP w/ XRCL		87.94	77.60	80.16	80.08	79.27	81.80	72.40	77.36	69.91	81.68	78.58	74.40	74.24	77.29
CoLAP w/ XCCL		87.94	77.76	79.68	79.33	78.94	81.31	72.80	77.28	70.09	80.88	78.42	74.16	74.05	77.06
FT	250	85.18	57.28	56.6	56.23	54.23	56.49	55.04	56.16	47.03	56.96	55.50	53.04	56.88	55.12
CA		85.18	55.68	55.14	54.90	53.25	55.27	53.6	54.64	45.32	55.12	54.08	51.04	56.20	53.69
PCT		87.38	81.60	84.51	85.46	81.71	83.78	80.20	83.00	78.15	83.40	82.40	81.00	81.10	82.19
CoLAP w/ XRCL		87.94	83.92	86.56	82.18	83.50	81.71	81.44	83.20	75.32	87.84	80.83	81.52	77.85	82.16
CoLAP w/ XCCL		87.94	84.32	84.45	82.18	84.07	81.63	81.28	83.60	76.40	86.96	79.83	81.84	77.37	81.99

Table 11: Average accuracy per language on the MultiTACRED dataset for XLM-R.

Model	K	en	ar	de	es	fi	fr	hi	hu	ja	pl	ru	tr	zh	Avg.
FT	0	85.03	10.91	21.68	21.85	13.72	15.97	9.86	18.93	4.91	16.64	14.88	15.08	10.20	14.55
PCT		87.10	23.34	35.84	46.26	29.13	46.80	34.48	36.49	22.01	34.32	42.13	29.27	29.80	34.16
CoLAP		87.05	35.74	37.20	50.41	36.44	44.84	24.94	31.35	17.21	27.76	35.63	21.49	26.70	32.48
FT	5	85.03	11.71	26.04	24.12	18.89	19.91	13.27	20.49	6.58	21.72	16.77	18.33	12.49	17.53
CA		85.03	11.71	26.04	24.12	18.89	19.91	13.27	20.49	6.58	21.72	16.77	18.33	12.49	17.53
PCT		87.10	33.04	48.64	56.37	42.20	57.74	42.98	44.42	30.04	44.40	48.93	39.21	39.99	44.00
CoLAP w/ XRCL		87.05	49.00	44.40	66.53	46.71	57.06	33.44	41.30	28.75	40.72	45.18	32.80	35.75	43.47
CoLAP w/ XCCL		87.05	50.20	46.00	60.74	45.90	56.70	32.15	41.37	26.10	41.84	43.98	33.76	36.23	42.91
FT	10	85.03	13.31	26.92	25.74	19.49	20.04	13.39	20.89	7.01	22.24	18.54	19.49	13.11	18.35
CA		85.03	13.31	26.92	25.74	19.49	20.04	13.39	20.89	7.01	22.24	18.54	19.49	13.11	18.35
PCT		87.10	34.00	53.68	59.43	39.58	59.15	43.86	48.35	29.44	49.12	48.02	38.26	43.33	45.52
CoLAP w/ XRCL		87.05	47.92	55.40	62.50	51.72	64.08	38.93	49.68	34.27	48.00	50.54	40.98	45.45	49.12
CoLAP w/ XCCL		87.05	46.76	53.44	61.76	50.56	62.71	37.53	47.07	31.06	47.40	48.61	40.01	42.24	47.43
FT	50	85.03	21.45	38.52	43.63	33.03	34.49	20.45	29.91	19.50	31.56	28.72	29.67	22.42	29.45
CA		85.03	19.77	38.40	42.80	32.30	32.91	19.56	28.99	17.48	31.44	27.31	29.35	20.95	28.44
PCT		87.10	50.28	70.24	72.87	62.68	73.25	57.97	63.75	48.18	64.96	63.82	59.42	58.25	62.14
CoLAP w/ XRCL		87.05	59.15	66.16	72.02	59.47	72.71	49.68	57.42	46.40	62.20	62.60	53.12	56.16	59.76
CoLAP w/ XCCL		87.05	56.90	64.28	70.07	60.07	71.66	47.71	57.02	42.37	60.28	58.89	52.88	52.61	57.90
FT	100	85.03	39.10	53.44	53.98	52.34	44.11	36.28	46.14	37.78	48.52	53.21	45.05	41.70	45.97
CA		85.03	35.21	51.20	54.14	49.57	41.79	32.19	43.25	32.30	46.28	49.54	41.83	36.56	42.82
PCT		87.10	60.46	77.60	80.21	74.39	80.80	68.00	73.85	60.11	74.24	73.49	70.57	69.17	71.91
CoLAP w/ XRCL		87.05	73.10	81.44	84.06	76.61	81.99	66.64	75.74	62.68	77.40	78.68	70.68	68.78	74.82
CoLAP w/ XCCL		87.05	69.76	79.40	80.22	70.54	79.96	62.30	72.13	59.77	75.32	75.39	65.67	65.00	71.29
FT	250	85.03	81.58	87.96	88.50	86.63	72.77	75.20	85.39	74.84	87.48	86.28	84.27	78.70	82.47
CA		85.03	81.92	92.88	94.29	90.41	76.60	75.22	88.10	74.06	90.00	88.89	86.81	78.48	84.81
PCT		87.10	70.27	86.88	85.23	84.86	83.99	80.45	82.62	71.66	85.84	84.80	80.29	82.28	81.60
CoLAP w/ XRCL		87.05	85.88	92.72	94.53	90.66	90.13	82.90	85.61	78.39	92.52	90.28	83.91	82.33	87.48
CoLAP w/ XCCL		87.05	85.83	92.44	93.76	89.55	88.97	81.36	84.82	76.92	90.98	87.42	81.60	81.31	86.25

Table 12: Average accuracy per language on the MultiTACRED dataset for Gemma 2 2B.

Model	K	en	ar	de	es	fi	fr	hi	hu	ja	pl	ru	tr	zh	Avg.
FT	0	85.08	6.26	18.00	13.60	9.30	16.20	5.29	12.51	5.99	12.32	15.39	9.70	6.78	10.95
PCT		87.68	5.13	48.40	53.42	30.26	50.51	8.74	39.30	23.47	42.96	42.82	25.98	37.60	34.05
CoLAP		85.68	19.00	53.44	50.77	39.56	55.27	18.85	52.13	29.41	43.52	33.22	32.24	37.34	38.73
FT	5	85.08	8.82	21.48	20.93	12.36	18.97	7.06	13.88	6.67	14.76	16.40	10.51	7.02	13.24
CA		85.08	8.82	21.48	20.93	12.36	18.97	7.06	13.88	6.67	14.76	16.40	10.51	7.02	13.24
PCT		87.68	12.19	58.56	59.87	42.05	57.86	15.72	47.96	40.33	53.44	49.41	35.84	46.92	43.35
CoLAP w/ XRCL		85.68	19.00	53.44	50.77	39.56	55.27	18.85	52.13	29.41	43.52	33.22	32.24	37.34	38.73
CoLAP w/ XCCL		85.68	20.53	49.44	48.27	37.32	53.14	18.69	49.48	30.58	45.92	31.70	30.47	37.96	37.79
FT	10	85.08	9.02	22.28	21.55	13.68	19.23	6.62	14.64	7.16	16.16	17.25	10.71	7.63	13.83
CA		85.08	9.02	22.28	21.55	13.68	19.23	6.62	14.64	7.16	16.16	17.25	10.71	7.63	13.83
PCT		87.68	15.00	60.32	63.94	44.94	63.71	19.09	50.04	41.64	56.32	56.00	40.10	48.47	46.63
CoLAP w/ XRCL		85.68	22.77	54.88	56.12	40.86	58.75	21.34	51.81	33.78	50.00	38.36	35.17	39.95	41.98
CoLAP w/ XCCL		85.68	21.49	55.92	55.85	41.97	56.82	22.24	52.31	36.06	48.70	38.93	33.87	39.07	41.94
FT	50	85.08	13.91	28.96	27.58	22.30	32.30	10.51	20.86	12.58	27.36	21.24	15.92	14.67	20.68
CA		85.08	12.99	29.72	27.90	21.78	29.76	10.27	19.89	10.95	26.92	20.96	14.12	13.25	19.88
PCT		87.68	37.93	75.20	72.79	65.72	75.31	39.05	67.60	58.10	70.96	72.08	59.42	65.83	63.33
CoLAP w/ XRCL		85.68	46.43	72.48	73.06	63.08	74.60	41.69	67.74	51.53	67.90	68.69	58.52	58.41	62.01
CoLAP w/ XCCL		85.68	46.59	72.32	72.88	61.56	74.50	40.98	65.52	51.70	65.60	69.07	57.01	58.05	61.31
FT	100	85.08	25.79	45.00	53.14	45.55	47.94	30.87	43.14	37.34	43.36	49.82	40.16	43.21	42.11
CA		85.08	23.95	45.24	53.66	45.34	46.41	26.54	42.97	34.13	43.52	49.82	39.39	41.72	41.06
PCT		87.68	51.72	83.52	83.39	77.37	86.86	54.69	78.19	70.71	80.96	80.52	73.54	74.58	74.67
CoLAP w/ XRCL		85.68	60.63	80.60	81.20	75.16	87.30	56.31	78.15	66.00	78.20	78.20	73.35	72.50	73.97
CoLAP w/ XCCL		85.68	61.59	81.90	81.26	75.55	87.10	55.61	78.06	65.50	79.20	79.40	72.74	72.30	74.18
FT	250	85.08	89.04	89.92	90.00	90.00	89.10	88.96	90.00	90.00	89.92	90.00	90.00	90.00	89.75
CA		85.08	84.79	89.92	89.92	90.00	86.62	82.62	90.00	90.00	89.92	89.92	90.00	90.00	88.64
PCT		87.68	73.00	89.92	89.92	89.84	89.92	83.58	90.00	89.84	89.92	89.92	90.00	89.90	87.98
CoLAP w/ XRCL		85.68	83.42	91.92	92.00	91.92	91.92	85.50	92.00	91.75	91.92	91.92	92.00	91.92	90.68
CoLAP w/ XCCL		85.68	82.38	91.92	92.00	92.00	91.92	84.06	92.00	90.98	91.92	91.92	91.76	91.81	90.39

Table 13: Average accuracy per language on the MultiTACRED dataset for Mistral 7B.

Dataset	Label	Exemplar
XNLI	Entailment	Premise: "You look like f–ing hell , Brando , the Star reports he said , and he advised the actor to lose maybe a hundred pounds , pallie ." Hypothesis: "Brando looks like f–ing hell , the Star reported him as saying ."
	Neutral	Premise: "3) Their many enemies want to legislate them out of existence ." Hypothesis: "The Democrats have enemies that want to legislate them out of existence ."
	Neutral	Premise: "Only three now , sir ." Hypothesis: "3 pm now , sir ."
	Contradiction	Premise: "That would certainly help i 'm sure." Hypothesis: "No , that does not help me ."
	Contradiction	Premise: "They ate them , uncle ." Hypothesis: "The people were unharmed ."
AmNLI	Entailment	Premise: "That was my messed up." Hypothesis: "I made a mistake."
	Neutral	Premise: "They asked a few questions and I answered them and they said, Get your baggage and leave there immediately, and come to the address you were supposed to when you arrived in Washington." Hypothesis: "They told me to pick up by white suitcase."
	Neutral	Premise: "That was, that was a pretty scary day." Hypothesis: "It was really scary when the tornado came into town."
	Contradiction	Premise: "Yeah yeah i you know i i wouldn't even mind so much if they had a um corporation that is financed" Hypothesis: "It would make me angry to find out that they had financed the corporation."
	Contradiction	Premise: "Uh-huh all right bye now" Hypothesis: "Let's keep talking."
MultiTACRED	per:origin	"Flower has been exiled from the country since 2003 when he and teammate <E1>Henry Olonga</E1> wore black armbands to protest the ' death of democracy ' in <E2>Zimbabwe</E2>."
	org:city_of_headquarters	"Shares of <E1>Millipore Corp</E1> advanced again Wednesday after the <E2>Billerica maker</E2> of life sciences equipment said it had begun an effort to auction itself off to the highest bidder."
	org:founded	"There are 122 centrally-administered SOEs under the <E1>SASAC</E1> after China National Service Corporation for Chinese Personnel Working Abroad merged with China National Pharmaceutical Group Corporation on <E2>Monday</E2>."
	per:country_of_birth	"Bin Laden 's latest video rant sounds very much like <E1>Adam Gadahn</E1> a <E2>US born</E2> loner and leftist who converted to Islam and joined al Qaeda."
	per:stateorprovinces_of_residence	"For instance , the <E1>Thomas More Law Center</E1> in Ann Arbor , <E2>Mich.</E2> , which is appealing a health care ruling it lost in Detroit , is known for its unsuccessful defense of a Pennsylvania school district that hoped to teach " intelligent design " as an alternative to evolution."

Table 14: Exemplars with highest similarity scores for XNLI, AmNLI, and MultiTACRED.