Is Prompt-Based Finetuning Always Better than Vanilla Finetuning? Insights from Cross-Lingual Language Understanding

Bolei Ma[⋆ 1] **Ercong Nie**[⋆ 1,2] **Helmut Schmid**[1] **Hinrich Schütze**[1,2]
[1]Center for Information and Language Processing (CIS), LMU Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany
`bolei.ma@campus.lmu.de` `nie@cis.lmu.de`

Abstract

Multilingual pretrained language models (MPLMs) have demonstrated substantial performance improvements in zero-shot cross-lingual transfer across various natural language understanding tasks by finetuning MPLMs on task-specific labelled data of a source language (e.g. English) and evaluating on a wide range of target languages. Recent studies show that prompt-based finetuning surpasses regular finetuning in few-shot scenarios. However, the exploration of prompt-based learning in multilingual tasks remains limited. In this study, we propose the **ProFiT** pipeline to investigate the cross-lingual capabilities of **Pro**mpt-based **Fi**ne**t**uning. We conduct comprehensive experiments on diverse cross-lingual language understanding tasks (sentiment classification, paraphrase identification, and natural language inference) and empirically analyze the variation trends of prompt-based finetuning performance in cross-lingual transfer across different few-shot and full-data settings. Our results reveal the effectiveness and versatility of prompt-based finetuning in cross-lingual language understanding. Our findings indicate that prompt-based finetuning outperforms vanilla finetuning in full-data scenarios and exhibits greater advantages in few-shot scenarios, with different performance patterns dependent on task types. Additionally, we analyze underlying factors such as language similarity and pretraining data size that impact the cross-lingual performance of prompt-based finetuning. Overall, our work provides valuable insights into the cross-lingual prowess of prompt-based finetuning.

1 Introduction

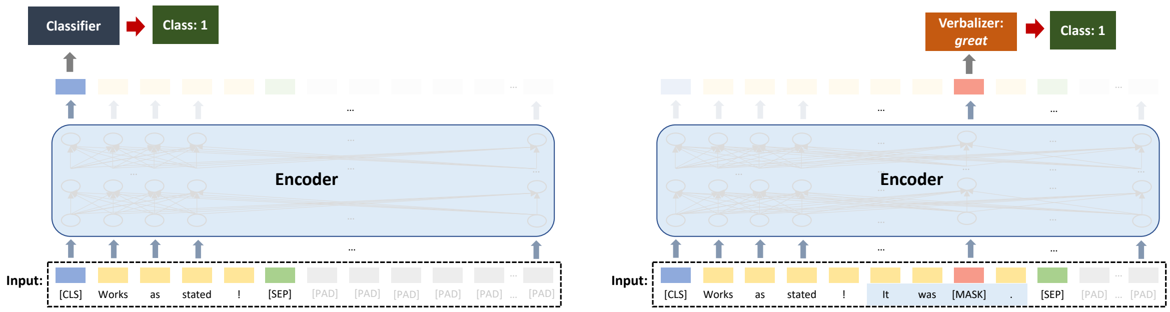
Pretrained language models (PLMs) (Devlin et al., 2019; Yang et al., 2019b; Radford et al., 2019), trained on massive amounts of unlabelled data in a self-supervised manner, have shown strong performance after finetuning on task-specific labelled data for a given downstream task, such as sentence classification (Zhuang et al., 2021), text summarization (Zhang et al., 2020), or dialogue generation (Liu et al., 2023c). *Prompt-based learning* (Brown et al., 2020; Schick and Schütze, 2021a,b,c) has recently emerged as a notable advancement, surpassing regular finetuning approaches in few-shot scenarios (Liu et al., 2023a). In prompt-based learning, downstream tasks are reformulated to resemble the types of problems tackled during the PLM's original pretraining by using a textual prompt. For example, in Figure 1(b), an input sentence of the binary sentiment analysis task "`Works as stated!`" can be reformulated with a prompt pattern $P(X) = X\circ$ "`It was [MASK].`" as "`Works as stated! It was [MASK].`" where $\circ$ is the string concatenation operator. We use a *verbalizer* which maps the class label to a *label word*. In this example, the class labels POSITIVE and NEGATIVE can be verbalized as "`great`" and "`bad`". By comparing the probabilities of the label words "`great`" and "`bad`" as fillers of the `[MASK]` token, we can predict the correct class label. In the example above, a natural language understanding (NLU) task is transformed into a masked language modeling (MLM) problem, which is the same as the PLM's pretraining objective.

The reformulated input can be used for finetuning, i.e. *prompt-based finetuning*. Figure 1 shows the difference between prompt-based finetuning and vanilla finetuning. Vanilla finetuning solely relies on the hidden embedding of the `[CLS]` token. In contrast, prompt-based finetuning makes use of both the semantic information from the task labels and the prior knowledge encoded in the pretraining phase. Recent empirical studies of few-shot learning showed advantages of prompt-based finetuning over vanilla finetuning (Gao et al., 2021; Li and Liang, 2021).

When applied to multilingual pretrained lan-

⋆ Equal Contribution.

Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023), pages 1–16
September 18–22, 2023.



(a) Vanilla finetuning (b) Prompt-based finetuning

Figure 1: The comparion of vanilla finetuning and prompt-based finetuning. [CLS], [SEP], [MASK], [PAD] are special tokens in the encoder vocabulary. The verbalizer is a function mapping from the task label set to a subset of the encoder vocabulary. Input tokens in blue represent the prompt pattern.

guage models (MPLMs), prompt-based finetuning also enables zero-shot[1] cross-lingual transfer. MPLMs such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) are pretrained on huge multilingual corpora and show strong multilinguality (Pires et al., 2019; Dufter and Schütze, 2020; Liang et al., 2021). They have become the dominant paradigm for zero-shot cross-lingual transfer, where annotated training data is available for some source language (e.g. English) but not for the target language (Wu and Dredze, 2019; Hu et al., 2020a). Zhao and Schütze (2021) proposed prompt-based finetuning for cross-lingual transfer. Their work focused on few-shot finetuning. Their experimental results for the natural language inference task showed that prompt-based finetuning performed better in few-shot cross-lingual transfer than vanilla finetuning. However, prior studies failed to examine whether prompt-based learning is also advantageous when training data is not scarce. Therefore, we conduct a comprehensive investigation on diverse cross-lingual language understanding tasks in both full-data and few-shot settings in order to shed more light on the cross-lingual capabilities of prompt-based finetuning.

In contrast to most previous research on prompting, our work is not restricted to monolingual or few-shot scenarios. Instead we explore a wide range of few-shot settings. We adopt a multilingual perspective and aim to uncover the nuances of performance variations associated with prompt-based finetuning. To this end, we implement the PROFIT pipeline and carry out an extensive set of experiments encompassing three representative cross-lingual language understanding tasks: sentiment analysis (Amazon Reviews), paragraph identification (PAWS-X), and natural language inference (XNLI). Our task selection covers single-sentence classification, sentence pair classification and inference task, considering both binary and multi-fold classifications. Our work provides insights into the effectiveness and versatility of prompt-based finetuning in cross-lingual language understanding.

Research Questions and Contributions. In this work, we analyze how the performance of prompt-based finetuning varies with the size of the labelled source language data for zero-shot cross-lingual transfer tasks. We examine a wide range of factors which could have an impact on cross-lingual transfer performance. We attempt to address the following pivotal research questions:

RQ1 *Does prompt-based finetuning outperform vanilla finetuning in the full-data scenario in different NLU tasks?*

We propose the PROFIT pipeline for systematically conducting the cross-lingual transfer experiments. We carry out zero-shot cross-lingual transfer experiments on three different NLU tasks using all the available English training data. By comparing the results of vanilla finetuning and PROFIT for different MPLMs, we find that in the full-data scenario, PROFIT still achieves better cross-lingual performance than vanilla finetuning.

RQ2 *Is prompt-based finetuning always better than vanilla finetuning?*

[1]In this paper, "zero-shot" in "zero-shot cross-lingual transfer" refers to the number of target language training data, i.e., no target language data is provided, while "few-shot" in "few-shot finetuning" refers to the source language used for finetuning, i.e., a few source language data is provided for the finetuning of the MPLM. The finetuned model is then zero-shot transferred to target language.

We investigate how the cross-lingual performance depends on the size of the English training data. Our findings substantiate that the ProFiT exhibits greater advantages in few-shot scenarios compared to full-data scenarios. The specific patterns of performance change are contingent upon the task types.

RQ3 *What underlying factors could affect the cross-lingual performance of* ProFiT*?*

We extensively analyze the factors that could influence the cross-lingual performance of ProFiT, encompassing language similarity, pretraining data size of target languages, etc.

2 Related Work

Prompt-Based Learning GPT-3 (Brown et al., 2020) has sparked research in prompt-based methods. Recent advances include automatic generation of prompt verbalizers and patterns (Schick et al., 2020; Shin et al., 2020), soft prompting (Qin and Eisner, 2021), prefix tuning (Li and Liang, 2021), P-tuning (Liu et al., 2022a), and retrieval-augmented prompting (Liu et al., 2022b). Most of these methods focus on monolingual scenarios, leaving the cross-lingual capabilities of prompt-based methods largely unexplored.

MPLMs and Zero-Shot Cross-Lingual Transfer The advances of MPLMs have positioned them as the standard approach for cross-lingual transfer. MPLMs usually adopt the architecture of some monolingual Transformer-based language model (Vaswani et al., 2017) and are jointly pretrained on large unlabelled multilingual data. For instance, mBERT (Devlin et al., 2019) is based on BERT; XLM-R (Zhuang et al., 2021) and Glot500-m (ImaniGooghari et al., 2023) are based on RoBERTa (Conneau et al., 2020). A multitude of studies have validated the robust multilinguality exhibited by MPLMs, either through probing the MPLMs themselves (Pires et al., 2019) or by identifying the key factors that contribute to their impressive multilinguality (Dufter and Schütze, 2020). Recent empirical studies have further demonstrated the remarkable cross-lingual capabilities of MPLMs by finetuning MPLMs on English training sets and then predicting on test sets of other languages (Karthikeyan et al., 2020; Turc et al., 2021). Several benchmarks have been proposed to evaluate the performance of multilingual encoders, including XTREME (Hu et al., 2020b), XTREME-R (Ruder et al., 2021), Taxi1500 (Ma et al., 2023) and XGLUE (Liang et al., 2020).

Multilingual Prompt Learning While prompting has proven successful in English, the application of prompting techniques in multilingual tasks has yet to be thoroughly explored and extensively studied. Zhao and Schütze (2021) first investigated prompt-based methods for cross-lingual transfer with different prompt forms and verbalizers. Recent follow-up studies introduced mask token augmentation (Zhou et al., 2022) and unified multilingual prompts (Huang et al., 2022) for zero-shot cross-lingual transfer. Despite the growing attention garnered by these methods in the context of few-shot scenarios across various NLP tasks, there remains a dearth of comprehensive investigations into the variations of prompt-based learning methods across different few-shot settings and full-data settings. Tu et al. (2022) focused on an alternative prompting approach for cross-lingual transfer in full-data scenarios. In contrast to prompt-based finetuning, they introduced additional prompt parameters to PLMs and exclusively updated these parameters during the finetuning process. A more recent work (Shi and Lipani, 2023) combined prompt-based finetuning and continued pretraining, but it was limited to monolingual scenarios.

In contrast to the aforementioned previous studies, our work provides a comprehensive investigation of prompt-based finetuning for cross-lingual transfer in both few-shot and full-data scenarios. Furthermore, we empirically analyze the variation of prompt-based finetuning performance across different few-shot settings.

3 Methodology

The purpose of this study is to improve the cross-lingual transfer performance of vanilla finetuning. In vanilla settings of zero-shot cross-lingual transfer, the MPLM is directly finetuned with training data in a source language (English). The finetuned model is then applied to predict the test data in target languages.

In prompt-based learning, we need a pattern-verbalizer pair (PVP) (Schick and Schütze, 2021a) consisting of (i) a *prompt pattern* which converts the input text into a cloze-style question with a mask token, and (ii) a representative word (called *verbalizer*) for each possible class. In our ProFiT approach, a PVP is combined with training data in English during finetuning. As the *training*

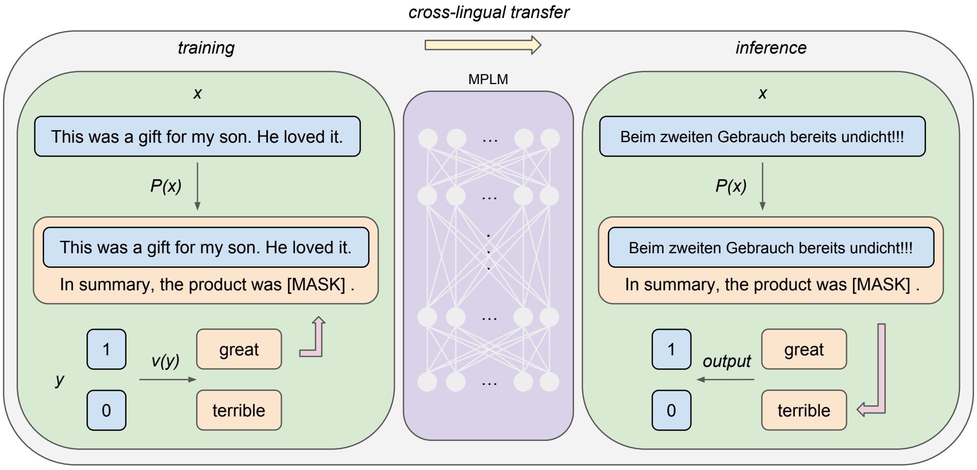


Figure 2: PROFIT pipeline of training and cross-lingual transfer with examples. X is an input sentence and $P(X)$ denotes the prompt pattern which reformulates the input into a prompt. $v(y)$ is the verbalizer which maps each class label y onto a word from the source language vocabulary.

block in Figure 2 shows, a prompt pattern such as $P(X) = X \circ$ "`In summary, the product was [MASK].`" is filled with an input example X "`This was a gift for my son. He loved it.`" A verbalizer such as {0 → "terrible", 1 → "great"} is used to map the original labels {0,1} onto words. The MPLM takes the filled pattern "`This was a gift for my son. He loved it. In summary, the product was [MASK].`", as input and returns for each of the two verbalizers "terrible" and "great" its probability of being the masked token. Thus, it uses the PVP to reformulate the sentence classification task of vanilla finetuning into a masked token prediction task.

More formally, let D={$(X_1, y_1), ..., (X_n, y_n)$} denote the set of training examples in the source language, where $X_1, ..., X_n$ are text samples and $y_1, ..., y_n$ are class labels from a label set Y. The prompt pattern $P(.)$ transforms an input sentence X into a cloze-style question with a masked token. The pretrained language model M with trainable parameters θ performs masked token prediction and returns the probabilities $p = M(P(X), \theta)$ of all candidate words for the masked token in $P(X)$. The verbalizer $v(.)$ is a bijective mapping from the set of class labels Y to a set of verbalised words V from the source language vocabulary. We predict the class $\hat{y}$ whose verbalizer $v(\hat{y})$ received the highest probability from model M:

$$\hat{y} = \arg\max_{y \in Y} p(v(y)) \quad (1)$$

We finetune the parameters θ of model M by minimizing the cross-entropy loss function ℓ on D:

$$\hat{\theta} = \arg\max_{\theta} \sum_{(X,y) \in D} \ell(v(y), M(P(X), \theta)) \quad (2)$$

The model with the finetuned parameters $\hat{\theta}$ is used to predict the class labels of the target language examples $D' = \{X'_1, ..., X'_n\}$ using the same prompt pattern and verbalizer as during finetuning (see *inference* block in Figure 2). The best label y'_i for each example X'_i is predicted according to Eq. 1.

In contrast to vanilla finetuning, prompt-based methods such as PROFIT only transform the training data with the prompt pattern P and the verbalizer v, but leave the model architecture unchanged. thus not hindering the efficiency of Vanilla much (Shi and Lipani, 2023). No extra parameters have to be trained from scratch. By reformulating the sentence classification task into a masked token prediction (MTP) task, we can better take advantage of the knowledge that the model has acquired during MTP pretraining.

In the cross-lingual setting, we simply apply the same functions P and v to the target language examples without further modifications.

4 Experimental Setups

4.1 Datasets

In order to investigate the performance on diverse NLU tasks, three representative different classification tasks on NLU are selected for evaluation in this work: sentiment analysis on Amazon product reviews (Keung et al., 2020), paraphrase identification on PAWS-X (Yang et al., 2019a), and natural language inference on XNLI (Conneau et al., 2018).

Amazon Reviews Dataset (Keung et al., 2020) contains product reviews with 5 star ratings from 1 to 5. The multilingual version of this dataset consists of test data in English and 5 other languages. We use the following prompt pattern $P(X)$ and verbalizer $v(y)$ for each review example (X, y):

- $P(X) = X \circ$ "All in all, it was [MASK]."
- $v(1) =$ "terrible", $v(2) =$ "bad", $v(3) =$ "ok", $v(4) =$"good", $v(5) =$ "great"

PAWS-X is a multilingual version of PAWS (Zhang et al., 2019), which consists of challenging paraphrase identification pairs from Wikipedia and Quora. Each data item comprises two sentences. The task is to predict whether the two sentences are paraphrases. The labels are binary: 1 for paraphrase, 0 for non-paraphrase. PAWS-X consists of datasets in English and 6 other languages. For a given sentence pair X_1 and X_2, we design the pattern and verbalizer as:

- $P(X_1, X_2) = X_1 \circ$"? [MASK], " $\circ X_2$
- $v(0) =$ "Wrong", $v(1) =$ "Right"

XNLI is a multilingual version of the MultiNLI dataset (Williams et al., 2018). The text in each data item consists of two sentences. Sentence A is the premise and sentence B is the hypothesis. The task is to predict the type of inference between the given premise and hypothesis among the three types: "entailment" (0), "neutral" (1), and "contradiction" (2). It is a kind of multi-class natural language inference task. XNLI consists of datasets in English and 14 other languages. For a given sentence pair X_1 and X_2, we design the pattern and verbalizer as:

- $P(X_1, X_2) = X_1 \circ$ "? [MASK], " $\circ X_2$
- $v(0) =$ "Yes", $v(1) =$ "Maybe", $v(2) =$ "No"

4.2 Baseline

The following baselines are considered and compared to our PROFIT approach:

MAJ The majority baseline. It always assigns the majority class from the training data.

Direct The pattern filled with the input sample is directly fed to the MPLM for prediction, without finetuning. This is the zero-shot scenario.

Vanilla The standard finetuning method which predicts the class from the hidden embedding of the `[CLS]` token without using a prompt pattern. We use the cross-entropy loss as the objective function for finetuning and AdamW for optimization with a learning rate of 1e-5 and 5 training epochs. The finetuned models are then used to predict the test data.

4.3 Multilingual Models

In order to solve the classification tasks with cross-lingual transfer, we use the pretrained multilingual BERT model (Devlin et al., 2019) "`bert-base-multilingual-cased`" (M) and the XLM-R model (Conneau et al., 2020) "`xlm-roberta-base`" (X) from the Huggingface Transformers library (Wolf et al., 2020). Both models are evaluated with the methods Vanilla and PROFIT. We repeat all our experiments 5 times with different random seeds. The details about model training and hyperparameter settings can be found in Appendix §A.1.

5 Results

5.1 Main Results

	Amazon	PAWS-X	XNLI	Avg.
MAJ	20	55.81	33.33	36.17
Direct-mBERT	20.21	45.05	35.05	33.44
Vanilla-mBERT	42.97	80.24	65.05	62.75
PROFIT-mBERT	**43.98**	**82.16**	**65.79**	**63.98**
Direct-XLM-R	21.98	51.10	35.68	36.25
Vanilla-XLM-R	54.56	82.51	73.61	70.22
PROFIT-XLM-R	**54.66**	**82.73**	**73.82**	**70.40**

Table 1: Overview of results

Table 1 gives an overview of the experimental results. PROFIT outperforms the MAJ baseline with both mBERT and XLM-R for all three classification tasks. PROFIT also outperforms the Direct and Vanilla baselines in both mBERT and XLM-R settings: When trained with mBERT, the performance is improved by **23.77%**, **37.11%** and **30.74%** compared to Direct on Amazon, PAWS-X and XNLI respectively, and by **1.01%**, **1.92%** and **0.74%** compared to Vanilla. When trained with XLM-R, the performance is improved by **32.68%**, **31.63%** and **38.14%** compared to Direct, and by **0.10%**, **0.22%** and **0.21%** compared to Vanilla respectively.

Task	Model	en	ar	bg	de	el	es	fr	hi	ja	ko	ru	sw	th	tr	ur	vi	zh	avg.
Amazon	Vanilla-M	58.92	-	-	45.69	-	48.02	47.45	-	35.07	-	-	-	-	-	-	-	**38.63**	42.97
	ProFiT-M	**59.05**	-	-	**46.66**	-	**49.30**	**48.38**	-	**37.31**	-	-	-	-	-	-	-	38.26	**43.98**
	Vanilla-X	59.61	-	-	**60.14**	-	55.24	55.66	-	51.93	-	-	-	-	-	-	-	49.82	54.56
	ProFiT-X	**60.06**	-	-	59.60	-	**55.72**	**55.89**	-	**52.34**	-	-	-	-	-	-	-	**49.75**	**54.66**
PAWS-X	Vanilla-M	93.85	-	-	84.94	-	87.11	86.55	-	73.39	72.44	-	-	-	-	-	-	77.01	80.24
	ProFiT-M	**94.21**	-	-	**86.06**	-	**88.17**	**87.91**	-	**75.79**	**75.82**	-	-	-	-	-	-	**79.22**	**82.16**
	Vanilla-X	94.33	-	-	86.92	-	88.55	89.04	-	**76.07**	74.71	-	-	-	-	-	-	79.75	82.51
	ProFiT-X	**94.90**	-	-	**87.06**	-	**88.87**	**88.86**	-	75.53	**75.40**	-	-	-	-	-	-	**80.63**	**82.73**
XNLI	Vanilla-M	82.57	65.12	68.97	71.40	66.30	74.22	73.68	60.02	-	-	68.95	50.24	53.15	62.02	57.96	69.80	68.91	65.05
	ProFiT-M	82.57	**65.55**	**69.47**	**71.57**	**67.43**	**75.10**	**74.57**	**60.57**	-	-	**69.55**	**51.13**	**54.58**	**62.64**	**58.04**	**70.74**	**70.08**	**65.79**
	Vanilla-X	84.91	**71.86**	77.78	76.86	75.96	79.25	78.21	69.92	-	-	**75.79**	**65.21**	72.02	73.12	66.07	74.71	73.72	73.61
	ProFiT-X	**84.97**	71.81	**77.92**	**77.35**	**76.11**	**79.31**	**78.75**	**70.10**	-	-	75.43	65.13	**72.39**	**73.23**	**66.95**	**75.05**	**73.92**	**73.82**

Table 2: Detailed cross-lingual performance results on three classification tasks. When calculating the average (avg.), due to the aim of zero-shot cross-lingual transfer, the performance results of the source language English are not taken into account. Model M stands for mBERT, and X for XLM-R.

While ProFiT outperforms all baselines on all three tasks, the degree of improvement differs. The improvements of ProFiT over Vanilla when trained with mBERT (**+1.23%**) are larger than the improvements when trained with XLM-R (**+0.18%**).

We further conducted T-tests for results of Vanilla and ProFiT with different random seeds (see §A.1 for the seeds). Table 3 shows the T-test results with p values for each task with mBERT and XLM-R models. We can see that the p values of all three tasks with mBERT model are under 0.05, indicating that the performance gain of ProFiT is significant with mBERT, while the p values of all three tasks with XLM-R model are bigger than 0.05, showing no significant performance difference.

Model	Amazon	PAWS-X	XNLI
mBERT	0.005	0.003	0.005
XLM-R	0.40^*	0.46^*	0.44^*

Table 3: T-Test results (p) for results of Vanilla and ProFiT with different random seeds. Insignificant results with a p value > 0.05 are marked with *.

One reason for the performance difference of the two models could be that the XLM-R model was pretrained on far more data than mBERT and is also much bigger, so that the Vanilla performance with XLM-R finetuning is much better than with mBERT in cross-lingual context (Conneau et al., 2020; Lauscher et al., 2020), leaving less space for improvement.

A detailed overview of the cross-lingual performance of ProFiT compared to Vanilla for each target language is presented in Table 2. Although the overall performance of ProFiT is better than Vanilla for all three tasks in both mBERT and XLM-R settings, individual differences between languages can be noticed. On Amazon, with mBERT, the improvement in Japanese (ja) (**+2.24%**) is far greater than on average, whereas Chinese (zh) shows no improvement (**-0.37%**); with XLM-R, ProFiT performs slightly worse than Vanilla on both Chinese with **-0.07%** and German (de) with **-0.54%**. On PAWS-X, Korean (ko) shows a larger improvement (**+3.38%**) than average with mBERT, and with XLM-R, whereas French (fr) (**-0.18%**) and Japanese (**-0.54%**) show a slightly worse performance than Vanilla. On XNLI, we find improvements for all languages with mBERT, and with XLM-R, Arabic (ar) (**-0.06%**), Russian (ru) (**-0.36%**), and Swahili (sw) (**-0.08%**) show slightly worse performance than Vanilla.

We conclude that the performance gain of ProFiT over Vanilla depends on the models and languages. In §6, we will further investigate how linguistic factors influence cross-lingual transfer performance.

5.2 Few-shot Ablations

Previous studies show that the prompt framework is more effective than finetuning when training data is scarce (Zhao and Schütze, 2021; Qi et al., 2022). We investigated how the performance changes as the number of training samples K increases in few-shot settings. The training and validation data are randomly sampled with $K \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$

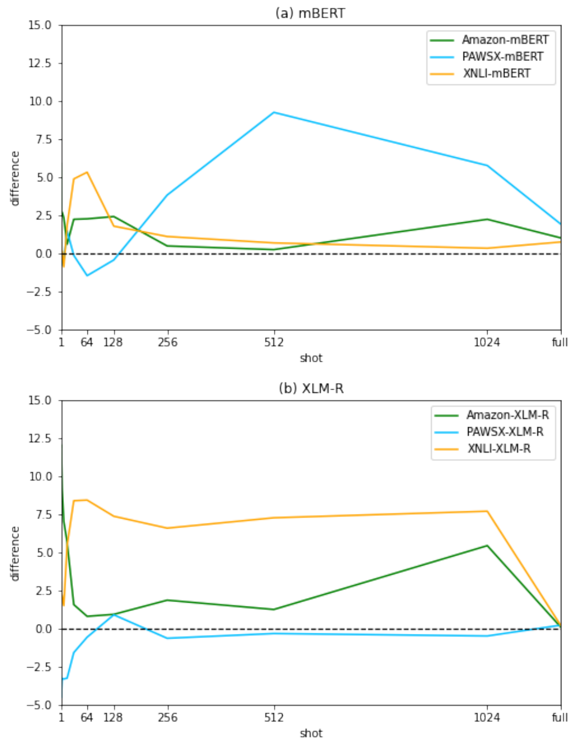


Figure 3: Performance difference between PROFIT and Vanilla in different few-shot settings and full training setting on three NLU tasks with both mBERT and XLM-R models.

shots per class from the English training data.

The detailed results of few-shot ablations can be found in Table 9, Table 10 and Table 11 in Appendix §A.4. Figure 3 shows the performance changes on all three tasks with both mBERT and XLM-R models. On the Amazon task, the performance improvement for smaller numbers of shots is greater than for full training. As the number of shots increases, the improvement decreases accordingly. This implies that on the sentiment analysis task, PROFIT is most valuable with small training data. On XNLI, the improvement of PROFIT over Vanilla is first small with in small shots. It then gets greater, as K increases, and drops again, as bigger K towards full data size shows up. We conclude that on NLI tasks such as XNLI, PROFIT is most effective in few-shot settings with a certain number of K. On PAWS-X, no obvious difference in few-shot settings can be found with mBERT in small shots, but in bigger shots there is greater improvement with $K \in \{256, 512, 1024\}$; however, with XLM-R, PROFIT shows almost no performance improvement over Vanilla.

Overall, sentiment analysis exhibits a clearer performance improvement for smaller numbers of shots, whereas the language inference and paraphrase tasks show greater performance enhancements in few-shot scenarios with larger K. This might be due to difficulties with pairwise inputs in these tasks, where we aim to identify the relationship between a pair of sentences. When it comes to transferring knowledge of sentence relationships, more examples are needed for successful learning than in sentiment analysis tasks where semantic information from comparable cross-lingual sentences can be directly transferred.

6 Cross-Lingual Analysis

In previous empirical studies of cross-lingual transfer learning (Lauscher et al., 2020; Nie et al., 2023), several key factors were identified to exert great effect on the cross-lingual performance, including (1) the size of the pretraining corpus for the target language and (2) the similarity between the source and target languages. We analyze how these two factors influence PROFIT's effectiveness for the languages on three tasks.

The pretraining corpus size of the target languages can be simply measured by the log_2 of the number of articles in Wikipedia[2].

For measuring the similarity between languages, we employ methods from recent studies of language representations. In these studies, languages are encoded as vectors according to their various linguistic and typological features. With these language vectors, a range of distance metrics, such as Euclidean distance and cosine similarity, can be used to measure the similarity between languages. Littell et al. (2017) proposed LANG2VEC which encodes languages using 5 vectors, with each vector representing a specific language feature. Östling and Kurfalı (2023) measured the lexical similarity by calculating language vectors based on the ASJP word list database (Wichmann et al., 2022). Liu et al. (2023b) recently proposed a novel language similarity metric from the perspective of conceptualization across multiple languages. In our work, we compute two similarity metrics: (i) a comprehensive linguistic similarity metric based on LANG2VEC (Littell et al., 2017) and (ii) a lexical similarity metric based on the ASJP word list database (Östling and Kurfalı, 2023).

The LANG2VEC approach provides information-rich vector representations of languages from dif-

[2] https://meta.wikimedia.org/wiki/List_of_Wikipedias

lang	Typological & Phylogenetic Sim.						Lexical Sim.			Size	Task Performance					
	SYN	PHO	INV	FAM	GEO	**$\mathbf{Sim}_1$**	*UMAP*	*SVD*	**$\mathbf{Sim}_2$**		amazon-M	amazon-X	pawsx-M	pawsx-X	xnli-M	xnli-X
ar	65.47	70.06	75.88	0.00	97.04	**61.69**	-1.90	4.87	**1.49**	**20.20**	-	-	-	-	65.55	71.81
bg	78.78	90.45	70.02	13.61	99.01	**70.38**	8.65	33.21	**20.93**	**18.15**	-	-	-	-	69.47	77.92
de	79.05	83.62	77.62	54.43	99.76	**78.90**	83.42	76.83	**80.13**	**21.42**	46.66	59.60	86.06	87.06	71.57	77.35
el	73.19	95.35	64.75	14.91	98.95	**69.43**	1.24	24.81	**13.03**	**17.76**	-	-	-	-	67.43	76.11
es	84.97	85.81	64.99	9.62	99.59	**69.00**	1.61	28.30	**14.96**	**20.83**	49.30	55.72	88.17	88.87	75.10	79.31
fr	76.83	75.26	73.64	9.62	99.93	**67.06**	1.34	31.76	**16.55**	**21.27**	48.38	55.89	87.91	88.86	74.57	78.75
hi	58.79	85.81	76.53	12.60	91.10	**64.97**	1.20	21.11	**11.16**	**17.26**	-	-	-	-	60.57	70.10
ja	49.63	64.44	65.92	0.00	85.65	**53.13**	-	-	-	**20.39**	37.31	52.34	75.79	75.53	-	-
ko	55.66	74.62	71.04	0.00	86.93	**57.65**	-0.22	12.42	**6.10**	**19.28**	-	-	75.82	75.40	-	-
ru	75.74	90.45	63.17	16.67	95.81	**68.37**	8.63	32.60	**20.62**	**20.87**	-	-	-	-	69.55	75.43
sw	42.26	90.91	76.16	0.00	91.50	**60.17**	-9.05	-7.18	**-8.12**	**16.23**	-	-	-	-	51.13	65.13
th	65.20	81.82	78.88	0.00	85.25	**62.23**	-0.21	3.82	**1.81**	**17.25**	-	-	-	-	54.58	72.39
tr	43.36	85.81	68.49	0.00	98.25	**59.18**	-7.80	-1.56	**-4.68**	**19.00**	-	-	-	-	62.64	73.23
ur	50.01	0.00	71.56	12.60	92.54	**45.34**	1.35	24.92	**13.14**	**17.54**	-	-	-	-	58.04	66.95
vi	64.92	78.33	74.76	0.00	85.25	**60.65**	0.86	-18.50	**-8.82**	**20.29**	-	-	-	-	70.74	75.05
zh	73.49	78.33	74.91	0.00	88.42	**63.03**	-	-	-	**20.37**	38.26	49.75	79.22	80.63	70.08	73.92

Table 4: Overview of language features and task performances with PROFIT for correlation analysis. Language features include typological & phylogenetic similarities (**$\mathbf{Sim}_1$**), lexical similarities (**$\mathbf{Sim}_2$**), and target language size (**Size**). Task performance contains the PROFIT results on the three datasets with both mBERT and XLM-R models.

ferent linguistic and ethnological perspectives. We adopt five linguistic categories: syntax (SYN), phonology (PHO), phonological inventory (INV), language family (FAM), and geography (GEO). SYN, PHO and INV are typological categories, and FAM and GEO are phylogenetic categories. Given these vectors, we calculate 5 different cosine similarity metrics between English and each target language.

The lexical similarity metric is based on a mean normalized pairwise Levenshtein distance matrix from ASJP. The language vectors used for calculating the lexical similarity are reduced in dimensionality. Two dimensionality reduction methods are employed for calculating the lexical similarity: Uniform Manifold Approximation and Projection (*UMAP*) (McInnes et al., 2018) and Singular Value Decomposition (*SVD*) (Stewart, 1993).

The final typological and phylogenetic similarity score **$\mathbf{Sim}_1$** for each language pair is calculated by averaging the 5 similarities of LANG2VEC. Similarly, the lexical similarity score **$\mathbf{Sim}_2$** is calculated by averaging the similarities of the *UMAP* and *SVD* vectors. More formally, as Eq. 3 shows, let f denote a feature from the feature set $\mathcal{F}_n$ for metric n, and let v_f denote the corresponding feature vector. The sim_1 and sim_2 scores for the source language English (e) and some target language j are then calculated by:

$$sim_n(e,j) = \frac{1}{|\mathcal{F}_n|} \sum_{f \in \mathcal{F}_n} \frac{v_f(e) \cdot v_f(j)}{\|v_f(e)\|_2 \, \|v_f(j)\|_2} \quad (3)$$

Table 4 shows a list of language features (typological & phylogenetic similarities, lexical similarities, and target language size) and task performances with PROFIT for the following correlation analysis. The language similarities, namely the typological & phylogenetic similarities (**$\mathbf{Sim}_1$**) and lexical similarities (**$\mathbf{Sim}_2$**) refer to the similarity between each language and English, based on the above introduced language vectors. Sim_1 and Sim_2 are calculated by Eq. 3. *ja* and *zh* are not included in Östling and Kurfalı (2023)'s original language sets, thus these two values are missing for the lexical similarities. The target language size (**Size**) is calculated by the log_2 of the number of articles in Wikipedia.

Based on the obtained language features and experimental results of task performance with PROFIT, we did a correlation analysis. Table 5 shows the results of the two correlation tests on each task.

According to the results of Pearson and Spearman tests and the p values, the two factors, namely, both the size of pretraining data for the target language and the similarity of typological and phylogenetic features of languages (sim_1) have a significant positive correlation with the improvement

Task	Model	Stat.	sim$_1$		sim$_2$		Size	
			$corr.$	p	$corr.$	p	$corr.$	p
Amazon	ProFiT-M	P	0.73	0.16*	-0.95	0.21*	0.81	0.09*
		S	0.70	0.19*	-1.00	0.00	0.50	0.39*
	ProFiT-X	P	0.80	0.10*	1.00	0.01	0.92	0.03
		S	0.80	0.10*	1.00	0.00	1.00	1e-24
PAWS-X	ProFiT-M	P	0.82	0.05	0.31	0.69*	0.82	0.04
		S	0.83	0.04	0.20	0.80*	0.60	0.21*
	ProFiT-X	P	0.83	0.04	0.34	0.66*	0.84	0.04
		S	0.77	0.07*	0.20	0.80*	0.71	0.11*
XNLI	ProFiT-M	P	0.57	0.03	0.43	0.14*	0.86	9e-05
		S	0.59	0.03	0.53	0.06*	0.90	1e-05
	ProFiT-X	P	0.72	4e-03	0.43	0.14*	0.70	5e-03
		S	0.77	1e-03	0.63	0.02	0.72	4e-03

Table 5: Correlations between task performance and language similarities (sim$_1$ & sim$_2$) and target language size. P stands for Pearson test and S for Spearman test. Insignificant results with a p value > 0.05 are marked with *.

of cross-lingual performance especially on XNLI, with both ProFiT-M and ProFiT-X models. Only the correlations calculated with the similarity of lexical features (sim$_2$) show some insignificant results. Furthermore, on XNLI, the correlation with language similarity is stronger with ProFiT-X, while the correlation with target data size is stronger with ProFiT-M. We argue that the XLM-R model is bigger than mBERT, so that the linguistic features have more effect on the performance, while for the smaller model mBERT the data size plays a greater role, which further reveals our findings in §5.1 that the applied pretrained model for finetuning has an impact on the ProFiT performance.

On PAWS-X and Amazon, we find weak correlations with the proposed factors, which could result from the limitation of languages in test data: XNLI comprises 15 different languages, whereas PAWS-X and Amazon only contain 7 and 6 languages in the test set, respectively. Thus weaker correlations have been found.

To sum up, language similarity and size are two factors that impact the cross-lingual performance in our study, and we find significant correlations when the test set contains a larger amount of languages.

7 Conclusion

In our work, we introduce ProFiT for zero-shot cross-lingual transfer, a pipeline which reformulates input examples into cloze-style prompts and applies the input examples with the prompts and its verbalizers as masked token to finetuning, changing the sentence classification task of vanilla finetuning into a masked token prediction task. We finetune the multilingual pretrained language model (MPLM) on source language prompts and apply it to target language data. We use ProFiT with the two MPLMs mBERT and XML-R, and evaluate its efficacy on three different types of multilingual classification tasks in natural language understanding – multi-class sentiment classification, binary paraphrase identification, and multi-class natural language inference. Our experiments show that ProFiT outperforms vanilla finetuning with both mBERT and XML-R on all three tasks. We further discovered that the performance improvement of ProFiT is generally more obvious in few-shot scenarios. Additionally, we demonstrate that the similarity of the source and target language and the size of the target language pretraining data significantly correlate with the cross-lingual transfer performance of ProFiT, especially on a big dataset with a variety of test languages.

Limitations

This study presents the ProFiT pipeline, which aims to enhance zero-shot cross-lingual transfer performance. Our approach was evaluated on various multilingual datasets and showed improved performance. However, due to the limitations of the datasets, only a few languages could be evaluated, thus making it difficult to draw a typological conclusion for all languages. Besides, our exploration in using the prompt-based learning method for cross-lingual language understanding is restricted to single-sentence and sentence pair classifications. As future work, our investigation should be extended to more types of language understanding tasks, such as sequence labelling tasks, e.g. slot detection, named entity recognition, etc.

Ethics Statement

This research was conducted in accordance with the ACM Code of Ethics. All the datasets that we use are publicly available. We report only aggregated results in the main paper. We have not intended or do not intend to share any Personally Identifiable Data with this paper.

Acknowledgements

We thank the anonymous reviewers for their efforts and helpful advice. This work was partly supported by Munich Center for Machine Learning (MCML) and China Scholarship Council (CSC).

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

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.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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 3816–3830, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020a. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020b. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

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. *arXiv preprint arXiv:2202.11451*.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Kaliyaperumal Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

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.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2021. Locating language-specific information in contextualized embeddings. *arXiv preprint arXiv:2109.08040*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022a. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Yanchen Liu, Timo Schick, and Hinrich Schütze. 2022b. Semantic-oriented unlabeled priming for large-scale language models. *arXiv preprint arXiv:2202.06133*.

Yihong Liu, Haotian Ye, Leonie Weissweiler, Philipp Wicke, Renhao Pei, Robert Zangenfeind, and Hinrich Schütze. 2023b. A crosslingual investigation of conceptualization in 1335 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12969–13000, Toronto, Canada. Association for Computational Linguistics.

Yongkang Liu, Shi Feng, Daling Wang, Yifei Zhang, and Hinrich Schütze. 2023c. PVGRU: Generating diverse and relevant dialogue responses via pseudo-variational mechanism. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3295–3310, Toronto, Canada. Association for Computational Linguistics.

Chunlan Ma, Ayyoob ImaniGooghari, Haotian Ye, Ehsaneddin Asgari, and Hinrich Schütze. 2023. Taxi1500: A multilingual dataset for text classification in 1500 languages. *arXiv preprint arXiv:2305.08487*.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29).

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. 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.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021a. 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.

Timo Schick and Hinrich Schütze. 2021b. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021c. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Zhengxiang Shi and Aldo Lipani. 2023. Don't stop pre-training? make prompt-based fine-tuning powerful learner.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Gilbert W Stewart. 1993. On the early history of the singular value decomposition. *SIAM review*, 35(4):551–566.

Lifu Tu, Caiming Xiong, and Yingbo Zhou. 2022. Prompt-tuning can be much better than fine-tuning on cross-lingual understanding with multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5478–5485, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *arXiv preprint arXiv:2106.16171*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Søren Wichmann, Eric W. Holman, and Cecil H. 2022. The asjp database. Version 20.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019a. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling.

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.

Meng Zhou, Xin Li, Yue Jiang, and Lidong Bing. 2022. Enhancing cross-lingual prompting with mask token augmentation. *arXiv preprint arXiv:2202.07255*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Robert Östling and Murathan Kurfalı. 2023. Language embeddings sometimes contain typological generalizations.

A Appendix

A.1 Training Details

During training, we used the same hyperparameters for Vanilla and PROFIT to keep the variables consistent for comparison. The chosen hyperparameters for both full-shot training and few-shot training are documented in Table 6. To avoid random effects on training, we trained each experiment with 5 different random seeds $\{10, 42, 421, 510, 1218\}$ and take the average results.

Hyperparameter	Full	Few-shot
EPOCHS	5	50
LEARNING_RATE	1e-5	1e-5
BATCH_SIZE	8	1
GRADIENT_ACCUMULATION_STEPS	4	2
MAX_SEQ_LENGTH	128	128
EARLY_STOPPING_PATIENCE	-	3

Table 6: Hyperparameters

A.2 Dataset Statistics

In Table 7 we show a basic statistic view of the Amazon Review (Keung et al., 2020) , PAWS-X (Zhang et al., 2019) and XNLI (Williams et al., 2018) datasets. We use the original train-dev-test split from the datasets. For training and validation we use the English train and dev dataset, and for test we use the test sets of all languages. The test data size for each target language is the same in all tasks.

Task	Size			#Labels
	| Train |	| Dev |	| Test |	
Amazon	200 000	5 000	5 000	5
PAWS-X	49 401	2 000	2 000	2
XNLI	392 702	2 490	5 010	3

Table 7: Overview of the three datasets. Train and dev data size refers to the number of samples for English. Test data size refers to the number of samples for each target language.

A.3 Reproducibility

The code for data processing and model training is available at the following Github repository: `https://github.com/boleima/ProFiT`.

A.4 Detailed Results

We present the detailed results of few-shot training performance of Vanilla and PROFIT for all three tasks in Table 9 (Amazon Review), Table 10 (PAWS-X) and Table 11 (XNLI), as well as the T-test results for all tasks in few-shot conditions in Table 8.

Shot	Amazon		PAWS-X		XNLI	
	M	X	M	X	M	X
1	0.001	0.001	0.50	0.56*	0.01	0.12*
2	0.10	0.01	0.22*	0.08*	0.89*	0.18*
4	0.09*	0.02	0.80*	0.10*	0.05	0.07*
8	0.23*	0.04	0.83*	0.04	0.86*	0.14*
16	0.78*	0.11*	0.30*	0.05	0.27*	0.03
32	0.06*	0.16*	1.00*	0.58*	0.11*	0.01
64	0.03	0.18*	0.02	0.80*	0.09*	0.002
128	0.07*	0.11*	0.15*	0.82*	0.34*	0.01
256	0.73*	0.21*	0.12*	0.78*	0.07*	0.02
512	0.86*	0.01	0.04	0.90*	0.61*	0.004
1028	0.003	0.31*	0.03	0.55*	0.74*	0.03
full	0.005	0.40*	0.003	0.46*	0.005	0.44*

Table 8: T-Test results (p) for results of Vanilla and PROFIT in different few-shot conditions. M stands for mBERT and X stands for XLM-R. Insignificant results with a p value > 0.05 are marked with *.

Shot	Model	en	de	es	fr	ja	zh	avg.
1	Vanilla-M	22.30	20.66	19.82	20.02	20.14	20.08	20.14
	ProFiT-M	**28.52**	**26.05**	**26.98**	**26.18**	**25.96**	**25.01**	**26.04**
	Vanilla-X	21.98	22.15	21.69	21.79	21.42	21.52	21.71
	ProFiT-X	**37.09**	**29.86**	**35.06**	**36.10**	**33.13**	**34.00**	**33.63**
2	Vanilla-M	24.37	23.14	23.00	22.70	21.27	21.36	22.29
	ProFiT-M	**27.63**	**25.78**	**26.04**	**25.05**	**23.24**	**23.73**	**24.77**
	Vanilla-X	21.31	21.08	21.52	20.67	20.76	21.41	21.09
	ProFiT-X	**35.63**	**31.82**	**33.46**	**34.40**	**33.35**	**32.70**	**33.14**
4	Vanilla-M	27.04	24.94	23.95	23.93	23.86	22.20	23.78
	ProFiT-M	**30.63**	**26.87**	**27.67**	**26.34**	**25.44**	**26.05**	**26.47**
	Vanilla-X	29.74	29.96	29.67	30.87	26.12	28.89	29.10
	ProFiT-X	**40.23**	**37.91**	**38.60**	**38.75**	**38.84**	**37.11**	**38.24**
8	Vanilla-M	29.95	26.82	26.75	26.91	24.18	25.70	26.07
	ProFiT-M	**32.67**	**29.07**	**30.20**	**29.38**	**26.24**	**27.12**	**28.40**
	Vanilla-X	32.02	32.84	33.02	32.60	28.84	31.51	31.76
	ProFiT-X	**42.23**	**35.63**	**40.55**	**39.79**	**39.65**	**38.33**	**38.79**
16	Vanilla-M	33.92	30.87	32.01	30.29	28.94	28.36	30.09
	ProFiT-M	**35.27**	**31.66**	**32.10**	**31.37**	**29.70**	**28.58**	**30.68**
	Vanilla-X	38.97	39.42	38.70	38.84	34.61	35.72	37.45
	ProFiT-X	**44.78**	**44.40**	**43.89**	**43.55**	**42.57**	**41.26**	**43.13**
32	Vanilla-M	36.73	31.26	31.64	31.69	28.94	29.08	30.52
	ProFiT-M	**37.90**	**33.44**	**34.68**	**33.72**	**31.18**	**30.77**	**32.76**
	Vanilla-X	44.92	45.42	44.45	44.78	42.16	41.85	43.73
	ProFiT-X	**47.51**	**47.12**	**46.67**	**45.78**	**44.24**	**42.70**	**45.30**
64	Vanilla-M	39.85	33.76	35.20	34.65	30.98	29.90	32.90
	ProFiT-M	**41.62**	**36.25**	**37.84**	**36.15**	**32.97**	**32.56**	**35.15**
	Vanilla-X	48.06	**48.48**	46.77	**47.34**	44.01	42.05	45.73
	ProFiT-X	**49.42**	48.16	**47.99**	46.93	**45.58**	**44.00**	**46.53**
128	Vanilla-M	43.29	35.52	37.50	36.38	32.36	31.51	34.65
	ProFiT-M	**44.19**	**38.39**	**39.84**	**38.74**	**34.62**	**33.71**	**37.06**
	Vanilla-X	50.40	50.75	48.37	48.12	46.26	44.80	47.66
	ProFiT-X	**50.75**	**51.24**	**49.75**	**49.22**	**47.39**	**45.35**	**48.59**
256	Vanilla-M	**45.64**	37.15	39.23	38.20	**33.54**	**32.86**	36.20
	ProFiT-M	45.39	**37.71**	**39.99**	**40.31**	32.55	32.82	**36.68**
	Vanilla-X	51.21	50.92	47.15	47.85	46.01	44.23	47.23
	ProFiT-X	**51.40**	**52.18**	**50.22**	**49.81**	**47.65**	**45.60**	**49.09**
512	Vanilla-M	**47.66**	**37.57**	39.90	39.16	**33.82**	**33.64**	36.82
	ProFiT-M	47.64	37.48	**40.63**	**40.99**	32.76	33.40	**37.05**
	Vanilla-X	51.90	51.69	49.21	49.67	46.23	43.96	48.15
	ProFiT-X	**52.94**	**52.79**	**50.21**	**50.06**	**48.16**	**45.82**	**49.41**
1024	Vanilla-M	49.26	38.47	41.24	39.88	33.52	33.79	37.38
	ProFiT-M	**49.63**	**41.47**	**43.54**	**41.97**	**36.52**	**34.54**	**39.61**
	Vanilla-X	51.33	48.55	45.06	44.91	42.85	41.79	44.63
	ProFiT-X	**54.55**	**53.15**	**51.98**	**51.18**	**47.98**	**46.08**	**50.07**
full	Vanilla-M	58.92	45.69	48.02	47.45	35.07	**38.63**	42.97
	ProFiT-M	**59.05**	**46.66**	**49.30**	**48.38**	**37.31**	38.26	**43.98**
	Vanilla-X	59.61	**60.14**	55.24	55.66	51.93	**49.82**	54.56
	ProFiT-X	**60.06**	59.60	**55.72**	**55.89**	**52.34**	49.75	**54.66**

Table 9: Few-shot performance on Amazon

Shot	Model	en	de	es	fr	ja	ko	zh	avg.
1	Vanilla-M	**54.38**	53.29	54.22	54.25	53.37	54.01	53.20	53.72
	ProFiT-M	53.21	**54.18**	**54.44**	**54.34**	**55.31**	**54.35**	**53.80**	**54.40**
	Vanilla-X	**51.95**	**51.75**	**51.57**	**51.62**	**51.95**	**51.73**	**51.80**	**51.74**
	ProFiT-X	50.19	48.53	50.68	46.83	50.80	44.55	49.91	48.55
2	Vanilla-M	**53.54**	**53.60**	**53.81**	**54.18**	**54.43**	**54.54**	**53.77**	**54.06**
	ProFiT-M	52.38	53.04	53.34	53.13	54.35	53.90	51.82	53.26
	Vanilla-X	**54.95**	**54.73**	**54.30**	**54.57**	**54.25**	**54.05**	**54.32**	**54.37**
	ProFiT-X	51.59	50.25	51.65	48.86	51.31	46.30	50.70	49.85
4	Vanilla-M	**53.93**	**53.11**	53.38	**53.94**	53.85	**54.28**	**53.71**	**53.71**
	ProFiT-M	52.40	53.07	**53.64**	53.41	**54.79**	53.53	51.20	53.27
	Vanilla-X	53.15	**54.45**	**53.99**	**53.90**	**53.81**	**53.79**	**53.64**	**53.93**
	ProFiT-X	**53.54**	51.25	53.00	49.05	53.46	45.29	51.83	50.65
8	Vanilla-M	**54.30**	53.50	**53.51**	**54.02**	**54.03**	**53.94**	**54.15**	**53.86**
	ProFiT-M	52.81	**54.12**	53.42	53.31	53.98	53.51	51.93	53.38
	Vanilla-X	**54.60**	**55.13**	**54.68**	**54.80**	**55.46**	**55.10**	**55.14**	**55.05**
	ProFiT-X	53.18	52.65	53.03	51.22	52.48	48.83	52.21	51.74
16	Vanilla-M	**54.08**	50.86	52.04	52.66	51.77	52.27	51.23	51.81
	ProFiT-M	52.81	**53.08**	**53.80**	**53.20**	**53.51**	**53.95**	**52.09**	**53.27**
	Vanilla-X	**54.45**	**54.84**	**54.45**	**54.54**	**54.96**	**54.56**	**54.78**	**54.69**
	ProFiT-X	53.73	51.58	53.24	49.95	53.21	48.28	52.31	51.43
32	Vanilla-M	**54.03**	52.94	53.48	**53.65**	53.13	53.58	**53.08**	**53.31**
	ProFiT-M	52.99	**52.97**	**53.75**	53.14	**53.57**	**54.16**	51.42	53.17
	Vanilla-X	52.44	**53.95**	52.96	**53.21**	53.46	**54.05**	**53.94**	**53.60**
	ProFiT-X	**53.63**	51.96	**53.44**	50.51	**53.61**	49.84	52.73	52.01
64	Vanilla-M	**55.44**	**55.42**	**55.46**	**55.97**	**54.80**	**55.92**	**56.41**	**55.66**
	ProFiT-M	53.95	54.59	54.05	54.48	54.51	54.95	52.61	54.20
	Vanilla-X	55.20	**55.35**	54.69	**54.95**	**55.84**	**55.09**	**55.39**	**55.22**
	ProFiT-X	**56.60**	54.95	**55.90**	54.59	55.63	51.51	55.29	54.64
128	Vanilla-M	**56.63**	**56.29**	**56.69**	**56.43**	55.31	55.70	**55.75**	**56.03**
	ProFiT-M	55.54	55.76	55.28	55.26	**55.88**	**55.75**	55.61	55.59
	Vanilla-X	54.61	54.99	54.44	54.80	55.24	**55.14**	54.98	54.93
	ProFiT-X	**58.66**	**56.28**	**57.95**	**54.91**	**56.09**	52.39	**57.35**	**55.83**
256	Vanilla-M	58.66	56.00	56.38	56.93	55.36	55.77	55.65	56.02
	ProFiT-M	**61.84**	**60.51**	**60.65**	**60.90**	**58.56**	**58.70**	**59.70**	**59.84**
	Vanilla-X	59.30	**58.23**	58.79	**58.54**	57.18	**57.54**	**57.70**	**57.99**
	ProFiT-X	**59.94**	57.75	**59.58**	57.86	**57.28**	54.31	57.35	57.35
512	Vanilla-M	64.23	59.38	60.00	60.15	56.90	56.84	56.79	58.34
	ProFiT-M	**73.47**	**69.74**	**70.23**	**70.20**	**63.84**	**64.56**	**66.97**	**67.59**
	Vanilla-X	**77.03**	**71.28**	72.09	**72.46**	**63.43**	**63.79**	66.53	**68.26**
	ProFiT-X	76.94	71.01	**72.29**	71.24	63.19	63.28	**66.61**	67.94
1024	Vanilla-M	74.43	68.44	69.47	70.01	61.95	61.13	64.69	65.95
	ProFiT-M	**81.06**	**74.58**	**76.08**	**76.15**	**66.05**	**66.76**	**70.64**	**71.71**
	Vanilla-X	86.33	**79.23**	80.86	**80.74**	**69.25**	**68.18**	**73.26**	**75.25**
	ProFiT-X	**87.84**	78.94	**81.53**	80.58	67.68	68.01	71.85	74.76
full	Vanilla-M	93.85	84.94	87.11	86.55	73.39	72.44	77.01	80.24
	ProFiT-M	**94.21**	**86.06**	**88.17**	**87.91**	**75.79**	**75.82**	**79.22**	**82.16**
	Vanilla-X	94.33	86.92	88.55	**89.04**	**76.07**	74.71	79.75	82.51
	ProFiT-X	**94.90**	**87.06**	**88.87**	88.86	75.53	**75.40**	**80.63**	**82.73**

Table 10: Few-shot performance on PAWS-X

Shot	Model	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg.
1	Vanilla-M	33.58	32.97	32.97	33.46	32.70	33.33	33.43	32.44	32.93	32.85	33.12	33.05	32.96	33.00	32.99	33.02
	ProFiT-M	**37.58**	**34.93**	**33.56**	**35.95**	**35.02**	**34.25**	**36.38**	**33.93**	**36.76**	**34.62**	**33.83**	**34.07**	**34.22**	**36.43**	**37.41**	**35.10**
	Vanilla-X	33.73	33.07	32.86	33.51	32.66	33.40	33.54	32.50	33.04	33.15	33.18	33.14	33.00	33.08	33.04	33.08
	ProFiT-X	**39.26**	**34.61**	**34.85**	**36.28**	**36.88**	**33.59**	**34.92**	**39.76**	**34.47**	**36.53**	**36.33**	**36.56**	**37.03**	**37.40**	**36.61**	**36.13**
2	Vanilla-M	34.67	**34.98**	36.21	**36.15**	**35.46**	**36.91**	**36.42**	34.34	35.42	**34.67**	**34.20**	**35.40**	34.04	36.18	35.61	**35.43**
	ProFiT-M	**38.38**	34.85	**34.02**	35.07	35.20	33.44	35.70	**35.63**	**35.65**	34.50	33.78	34.35	**34.67**	**36.57**	**37.12**	35.04
	Vanilla-X	34.84	34.33	35.51	35.62	34.99	**36.25**	35.86	34.14	35.09	34.39	33.95	34.87	33.76	35.55	35.02	34.95
	ProFiT-X	**39.22**	**36.54**	**36.73**	**38.48**	**37.83**	34.21	**37.91**	**38.87**	**35.56**	**37.16**	**38.42**	**38.01**	**37.75**	**38.25**	**36.98**	**37.34**
4	Vanilla-M	37.91	**35.47**	**36.12**	36.20	35.03	**36.22**	36.09	34.60	35.60	**35.01**	**34.35**	**35.49**	**34.49**	36.28	35.74	35.48
	ProFiT-M	**38.04**	35.43	34.64	**36.67**	**36.50**	33.66	**36.63**	**36.07**	**36.83**	34.87	33.42	35.41	34.44	**37.06**	**37.07**	**35.62**
	Vanilla-X	37.55	34.31	35.08	35.11	34.09	**35.06**	34.85	33.74	**34.53**	34.09	33.58	34.39	33.71	35.09	34.56	34.44
	ProFiT-X	**38.79**	**36.03**	**35.23**	**37.49**	**37.36**	33.50	**36.54**	**38.79**	34.21	**37.11**	**37.79**	**36.47**	**37.58**	**37.96**	**36.22**	**36.59**
8	Vanilla-M	**40.83**	**37.39**	**38.56**	**38.69**	**37.77**	**39.25**	**39.06**	36.38	37.72	**37.54**	**36.46**	**38.07**	**36.28**	**38.22**	37.76	**37.80**
	ProFiT-M	38.71	36.59	35.73	37.20	37.33	34.88	38.05	**38.22**	**38.32**	35.37	35.40	36.48	35.99	38.20	**38.93**	36.91
	Vanilla-X	40.84	36.52	37.57	37.97	36.85	**38.50**	**38.35**	35.70	37.00	36.77	35.57	37.33	35.57	37.56	36.95	37.01
	ProFiT-X	**41.58**	**37.81**	**37.61**	**39.74**	**39.06**	35.07	37.65	**39.78**	**37.26**	**38.64**	**40.32**	**38.79**	**38.65**	**40.33**	**38.54**	**38.52**
16	Vanilla-M	42.42	39.56	40.71	40.36	39.63	**41.49**	41.14	37.86	39.60	**38.27**	37.35	38.77	37.44	40.76	40.25	39.51
	ProFiT-M	**44.52**	**42.10**	**41.96**	**40.85**	**42.18**	40.63	**43.98**	**41.17**	**43.10**	36.50	**38.83**	**41.71**	**38.95**	**43.40**	**43.14**	**41.32**
	Vanilla-X	42.65	39.37	40.33	40.09	39.15	**41.12**	40.73	37.72	39.44	38.02	37.34	38.63	37.19	40.73	40.01	39.28
	ProFiT-X	**49.72**	**42.15**	**43.51**	**47.38**	**46.22**	40.19	**44.09**	**45.59**	**43.14**	**44.81**	**46.16**	**45.39**	**44.43**	**47.35**	**45.69**	**44.72**
32	Vanilla-M	46.18	40.39	41.17	41.25	40.39	42.65	41.88	38.69	40.77	**38.29**	38.47	39.62	38.82	41.18	40.89	40.32
	ProFiT-M	**49.02**	**45.64**	**46.01**	**44.64**	**47.57**	**45.00**	**48.32**	**45.06**	**46.37**	38.28	**43.39**	**43.68**	**43.88**	**47.18**	**47.78**	**45.20**
	Vanilla-X	46.11	39.69	40.44	40.57	39.81	42.05	41.28	38.30	40.25	37.71	37.99	39.05	38.17	40.27	40.00	39.68
	ProFiT-X	**52.27**	**46.87**	**48.41**	**49.79**	**49.12**	**45.55**	**48.85**	**48.42**	**48.10**	**45.90**	**49.20**	**47.88**	**46.58**	**49.84**	**48.55**	**48.08**
64	Vanilla-M	52.10	45.26	46.64	48.10	46.32	49.44	48.57	42.71	45.45	39.13	40.24	42.19	42.41	47.23	46.91	45.04
	ProFiT-M	**55.04**	**50.28**	**51.76**	**52.60**	**52.90**	**50.46**	**53.85**	**49.57**	**51.68**	**42.26**	**46.38**	**49.01**	**48.85**	**52.89**	**52.57**	**50.36**
	Vanilla-X	51.86	44.99	46.39	47.86	45.84	48.92	48.47	42.99	45.25	39.04	40.35	42.43	42.51	47.08	46.70	44.92
	ProFiT-X	**59.35**	**50.75**	**53.38**	**55.47**	**55.32**	**50.92**	**55.71**	**53.11**	**52.67**	**51.31**	**53.99**	**52.95**	**51.30**	**55.51**	**54.41**	**53.34**
128	Vanilla-M	58.61	51.91	54.23	54.89	54.32	**56.27**	55.30	49.05	52.87	43.18	46.02	49.56	48.28	54.02	54.06	51.71
	ProFiT-M	**59.12**	**53.87**	**55.09**	**56.44**	**55.33**	55.00	**56.09**	**52.36**	**54.71**	**45.25**	**49.41**	**52.44**	**51.35**	**55.62**	**55.98**	**53.50**
	Vanilla-X	58.27	51.41	53.86	54.61	53.85	55.90	54.89	48.68	52.21	42.87	46.23	49.26	47.89	53.55	53.90	51.36
	ProFiT-X	**64.78**	**56.50**	**60.23**	**60.77**	**60.55**	**59.51**	**61.20**	**57.41**	**59.13**	**55.12**	**58.44**	**58.15**	**55.36**	**60.24**	**59.68**	**58.73**
256	Vanilla-M	61.88	53.54	56.61	57.25	56.20	**58.77**	57.91	51.31	55.45	44.97	46.97	52.75	50.07	56.51	56.76	53.94
	ProFiT-M	**62.30**	**54.82**	**56.96**	**57.92**	**56.48**	58.69	**58.39**	**53.58**	**57.09**	**45.55**	**49.06**	**53.64**	**52.41**	**57.81**	**58.06**	**55.03**
	Vanilla-X	61.68	53.30	56.19	57.01	55.91	58.47	57.74	51.13	55.22	44.86	46.68	52.77	49.79	56.24	56.33	53.69
	ProFiT-X	**66.55**	**58.08**	**62.26**	**62.24**	**61.23**	**62.88**	**63.44**	**58.56**	**60.42**	**54.77**	**59.95**	**59.95**	**56.59**	**62.28**	**61.18**	**60.27**
512	Vanilla-M	64.94	56.75	59.66	60.73	58.53	**61.99**	60.89	53.69	58.94	46.24	48.58	55.50	52.56	59.71	59.89	56.69
	ProFiT-M	**65.39**	**57.36**	**60.18**	**61.03**	**58.95**	61.59	**61.04**	**55.07**	**59.52**	**47.23**	**50.48**	**55.98**	**54.08**	**60.25**	**60.41**	**57.37**
	Vanilla-X	64.92	56.33	59.53	60.47	58.11	61.92	60.59	53.36	58.53	45.92	47.99	55.25	52.15	59.32	59.49	56.35
	ProFiT-X	**70.13**	**61.99**	**66.33**	**65.47**	**64.91**	**67.43**	**66.72**	**60.53**	**64.80**	**57.27**	**63.16**	**63.35**	**58.78**	**65.31**	**64.74**	**63.63**
1024	Vanilla-M	65.90	56.85	59.73	61.10	58.40	**62.73**	**62.07**	54.57	59.38	46.46	48.46	56.19	**54.21**	60.32	60.51	57.21
	ProFiT-M	**66.77**	**57.83**	**59.94**	**61.53**	**59.42**	62.05	61.99	**55.37**	**59.54**	**47.44**	**49.10**	**56.40**	53.91	**60.48**	**60.62**	**57.54**
	Vanilla-X	65.67	56.88	59.61	60.95	57.99	62.47	61.93	54.48	59.30	46.36	48.21	56.01	54.29	60.15	60.25	57.06
	ProFiT-X	**71.51**	**63.04**	**67.62**	**66.26**	**66.27**	**68.64**	**67.72**	**62.02**	**65.86**	**58.12**	**64.33**	**64.41**	**60.46**	**66.36**	**65.50**	**64.76**
full	Vanilla-M	82.57	65.12	68.97	71.40	66.30	74.22	73.68	60.02	68.95	50.24	53.15	62.02	57.96	69.80	68.91	65.05
	ProFiT-M	82.57	**65.55**	**69.47**	**71.57**	**67.43**	**75.10**	**74.57**	**60.57**	**69.55**	**51.13**	**54.58**	**62.64**	**58.04**	**70.74**	**70.08**	**65.79**
	Vanilla-X	84.91	**71.86**	77.78	76.86	75.96	79.25	78.21	69.92	**75.79**	**65.21**	72.02	73.12	66.07	74.71	73.72	73.61
	ProFiT-X	**84.97**	71.81	**77.92**	**77.35**	**76.11**	**79.31**	**78.75**	**70.10**	75.43	65.13	**72.39**	**73.23**	**66.95**	**75.05**	**73.92**	**73.82**

Table 11: Few-shot performance on XNLI