Prompt-Tuning Can Be Much Better Than Fine-Tuning on Cross-lingual Understanding With Multilingual Language Models

Lifu Tu and **Caiming Xiong** and **Yingbo Zhou**
Salesforce AI Research
{ltu,cxiong,yingbo.zhou}@salesforce.com

Abstract

Pre-trained multilingual language models show significant performance gains for zero-shot cross-lingual model transfer on a wide range of natural language understanding (NLU) tasks. Previously, for zero-shot cross-lingual evaluation, pre-trained models are only fine-tuned on English data and tested on a variety of target languages. In this paper, we do cross-lingual evaluation on various NLU tasks (sentence classification, sequence labeling, question answering) using prompt-tuning and compare it with fine-tuning. The results show that prompt tuning achieves much better cross-lingual transfer than fine-tuning across datasets, with only 0.1% to 0.3% tuned parameters. Additionally, we demonstrate through the analysis that prompt tuning can have better cross-lingual transferability of representations on downstream tasks with better aligned decision boundaries.

1 Introduction

Large Multilingual language models (Pires et al., 2019; Wu and Dredze, 2019; Conneau et al., 2020) show surprisingly impressive zero-shot cross-lingual transfer on NLP tasks, even though they are only trained from monolingual corpora. Recently, large-scale benchmarks such as XTREME (Hu et al., 2020) and XGLUE (Liang et al., 2020) are introduced for cross-lingual evaluation.

In a cross-lingual transfer setting, models are only fine-tuned on the task-specific annotations in one language and evaluated in other languages. During fine-tuning, pre-trained language models are used for initialization and the entire model parameters are tuned on downstream tasks. While fine-tuning obtains strong performance, it is inefficient. Also as shown in (Hu et al., 2020), the cross-lingual transfer gap between the performance on the English test set and all other languages is large even with the best baseline XLM-R (Conneau et al., 2020).

Recently, prompt tuning, where only a small amount of additional parameters (i.e. prompts) is added and tuned, but the original model is kept frozen. Much fewer parameters or no parameters are tuned and thus the training is a lot more efficient. Prompt tuning still performs worse than fine-tuning in lots of NLP tasks(Brown et al., 2020; Shin et al., 2020; Zhong et al., 2021). More recently, Li and Liang (2021); Lester et al. (2021); Hambardzumyan et al. (2021) indicate prompt tuning is competitive with fine tuning on some of the NLU tasks. Language model capacity (e.g., 10 billion parameters) is a key ingredient for these approaches to succeed. More recently, (Liu et al., 2022) shows prompt tuning can also be comparable on several hard monolingual sequence labeling tasks such as extractive question answers.

In this paper, we aim to investigate the effect of prompt tuning in cross-lingual tasks.We freeze the entire multilingual language model and tune task prompts on the English training set for downstream tasks (sentence classification, structure prediction, question answering). Even with medium size multilingual language model (less than 1 billion parameters), prompt tuning achieves much higher performance than fine-tuning on various NLU tasks.

According to the analysis results, prompt tuning does fewer changes to sentence representations than fine-tuning and keeps good cross-lingual sentence representations. We also find that the decision boundaries of different language sentence representations after prompt tuning on English data are almost aligned well. However, these decision boundaries of different languages after fine-tuning are a large difference. These aligned decision boundaries can lead to stronger cross-lingual transfer.

This work sheds light on the strong cross-lingual ability of prompt tuning. Our results suggest prompt tuning is better than fine-tuning on cross-lingual transfer. Our contributions are summarized as follows: we show that prompt tuning can per-

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5507 - 5514

form much better as compared to fine-tuning for cross-lingual transfer; we also show prompt tuning works better in the case of the cross-lingual transfer due to the relative small robust changes it brings to the originally learned representations.

2 Prompt-Tuning for Cross-Lingual Tasks

Multilingual Language Models. In the past years, lots of pre-trained multilingual language models come out: mBERT, XLM (CONNEAU and Lample, 2019), XLM-R (Conneau et al., 2020), etc. XLM-R (Conneau et al., 2020) significantly outperforms multilingual BERT (mBERT; Devlin et al., 2019) on a variety of cross-lingual benchmarks XTREME (Hu et al., 2020). In some previous work (Luo et al., 2021; Zhang et al., 2019), XLM-R is also used for initialization to do another round of pretraining with parallel data to get the stronger cross-lingual ability. Previously, in the cross-lingual evaluation, models are fine-tuned on the English training data but evaluated on all target languages. As far as we know, we are the first to explore prompt tuning on several hard multilingual NLP tasks including structure prediction and question answering

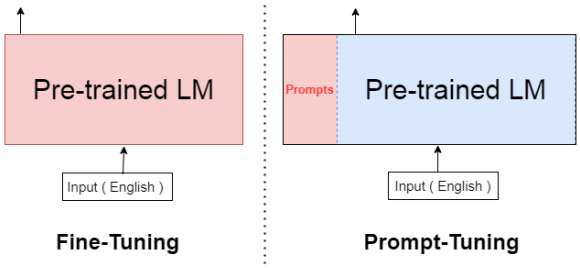


Figure 1: Two different approaches for cross-lingual evaluation when using large multilingual language model. **Left:** In fine-tuning, all model parameters are tuned on English task data. This setting is used in cross-lingual evaluation before. **Right:** In prompt tuning, only small ratio parameters are tuned. We use prefix prompts and use layer prompts in our experiments.

Prompt Tuning. Fine-tuning on large pre-trained language models leads to strong performance on downstream tasks, however, it is memory-consuming and lots of parameters need to save for each task. In prompt tuning, only a small part of the parameters (e.g., prompts or task classifier) are tuned during learning. However, it usually performs not as good as compared to fine-tuning. Recently, Lester et al. (2021) find prompt tuning can be better than fine-tuning when the model size is not extremely large (10 billion parameters). Prefix-tuning (Li and Liang, 2021) obtains comparable performance for natural language generation tasks. Liu et al. (2022) shows prompt tuning can be matched to fine-tuning on language understanding tasks even at hard sequence tagging tasks.

We investigate prompt tuning on cross-lingual understanding on a pre-trained multilingual language model. The framework is shown in Figure 1. Our setting is similar to Li and Liang (2021); Liu et al. (2022). The continuous prompts are added as prefix tokens and tuned during learning. In the implementation, the prompts are operated as past keys and values in each transformer layer. Each transformer layer has separated prompts. These continuous prompts are optimized, but multilingual language model parameters are frozen.

3 Experiments Setup

3.1 Datasets.

We perform experiments on four datasets included in XTREME: cross-lingual natural language inference (XNLI; Conneau et al., 2018), cross-lingual adversarial dataset for paraphrase identification (PAWS-X; Yang et al., 2019), part-of-speech tagging on the Universal Dependencies (UD-POS; Nivre et al., 2018), cross-lingual question answering on XQuAD (Artetxe et al., 2020) and TyDiQA-GoldP (Clark et al., 2020). Three categories of downstream tasks are included: (1) sentence classification); (2) structure prediction; (3) question answering.

3.2 Training Details.

Our frozen models are built on the top of the pre-trained XLM-R checkpoint of LARGE size with about 560M parameters. Previous work (Hu et al., 2020) shows it achieves stronger performance than mBERT[1]. All our experiments were run with Huggingface (Wolf et al., 2020). More details are in the appendix.

Prompt Length. Prompt length usually plays an important role in prompt tuning. In our experiments, we treat this as a hyper-parameter. Longer prompt length often leads to have higher performance. In our experiments, prompt length is set to 16 or 32 and tuned on the English validation set.

[1]Some preliminary results are obtained with mBERT.

Model	Sentence Classification		Structured Prediction	Question Answering	
	XNLI	PAWS-X	UD-POS	XQuAD	TyDiQA
Metrics	Acc.	Acc.	F1	F1 / EM	F1 / EM
Fine Tuning					
MBERT*	65.4	81.9	70.3	64.5 / 49.4	59.7 / 43.9
XLM-R-LARGE*	79.2	86.4	72.6	76.6 / 60.8	65.1 / 45.0
XLM-R-LARGE$^+$	79.2	-	75.0	77.2 / 61.6	64.3 / 45.8
XLM-R-LARGE (OUR)	78.8 (0.2)	87.9 (0.5)	74.4 (0.7)	77.3 (0.4) / 61.8 (0.5)	70.1 (0.6) / 51.7 (2.7)
Prompt Tuning					
XLM-R-LARGE	**79.9** (0.1)	**88.4** (0.3)	**75.4** (0.2)	**79.0** (0.2) / **64.1** (0.4)	**71.5** (0.4) / **55.1** (0.6)

Table 1: Zero-shot cross-lingual transfer evaluation results (with standard deviation) on XTREME structured prediction, question answering, and sentence classification tasks. For both fine tuning and prompt tuning, models are only fine-tuned on the English training data but evaluated on all target languages. Baseline fine-tuning results with "*" and "+" are taken from (Hu et al., 2020) and (Ruder et al., 2021) respectively. More results are shown in the Appendix.

4 Results

Tuned Parameter Sizes Comparison For the prompt tuning test results in Table 1, we did limited tuning on prompt length. The prompt length is 16, except prompt length for task XNLI is 32. With only 0.1% to 0.3% additional prompt parameters as compared to the original model, the framework already demonstrates strong cross-lingual results.

Overall Results Table 1 shows the zero-shot cross-lingual results on four different tasks. Prompt tuning performs much better than fine-tuning, especially for hard sequence task question answering. And prompt tuning is also with smaller variance.

Previously, although with parallel data or more monolingual data, cross-lingual transfer results (Zhang et al., 2019; Luo et al., 2021; Ruder et al., 2021) on question answering and structured prediction tasks improved only slightly. With prompt tuning, there is larger performance gains for question answering and structured prediction tasks. It suggests that prompt tuning is a better tuning method for cross-lingual transfer.

Cross-lingual Transfer Gap According to the above result, on average, prompt tuning achieves better performance than fine tuning. Table 2 shows the cross-lingual transfer gap of the two different tuning methods. Prompt tuning can also reduce the gap significantly.

Discussion In our preliminary experiments, for the smaller size model (e.g., mBERT), prompt tuning perform a little worse than fine tuning on English, and match the performance of fine-tuning on all languages. The language model size still matter. There is still some space for smaller size model.

	XNLI	PAWS-X	UD-POS	XQuAD
Fine Tuning	10.2	12.4	24.3	16.3
Prompt Tuning	9.7	8.7	20.7	14.5

Table 2: Cross-lingual transfer gap of the two tuning methods. The cross-lingual transfer gap is the performance difference between English test set and the average of the other languages. The smaller is better.

This also indicates potential for future work with better prompt tuning method.

5 Analysis

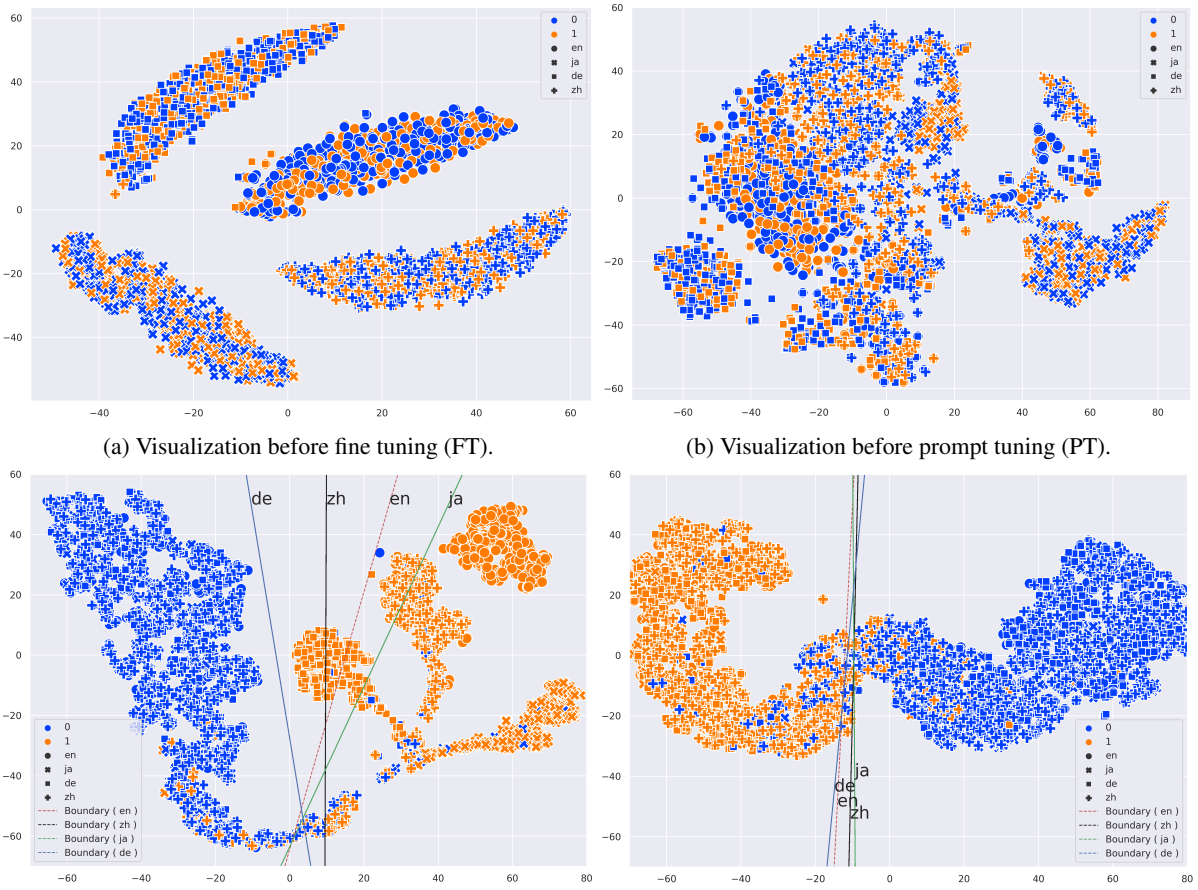
In order to perform some analysis on prompt tuning and fine tuning, we select 1000 samples for each language (en, de, es, fr, ja, ko, zh) from PAWS-X (Yang et al., 2019) dataset. For each English language sample in our selections, there is a human translated sample from the other six languages. [2]

Figure 2 shows t-SNE visualization of sample representations from frozen multilingual language model XLM-R. Samples' representations are clustered well respect to languages, however, there is weak correlation with labels.

5.1 Language Representation Changes

For each tuning method (fine-tuning and prompt-tuning), Table 3 shows the cosine similarity of representations from frozen language model and tuned model. According to the results, both of two tuned method make notable change on sentence representations. However, the average cosine similarity of fine-tuning is much smaller. It indicates that fine-tuning leads much larger changes on sentence rep-

[2]Each sample in PAWS-X dataset is a sentence pair. In the following experiments, we treat the representations at CLS token as the sample sentence representations.



(a) Visualization before fine tuning (FT).

(b) Visualization before prompt tuning (PT).

(c) Decision boundaries after fine tuning (FT).

(d) Decision boundaries after prompt tuning (PT).

Figure 2: T-SNE visualization of representations of four languages (en: English; de: German; ja: Japanese; zh: Chinese) before and after two different tuning methods on English task data. The decision boundaries after prompt tuning is aligned much better.

resentations than prompt tuning. We can also see representation changes is larger when tuning is on MNLI, while prompt tuning still has less changes on representations.

	en	de	es	fr	ja	ko	zh
Training on PAWS							
FT	25.2	26.5	24.5	25.2	18.9	15.0	22.6
PT	57.6	56.8	57.2	57.7	58.7	59.4	59.5
Training on MNLI							
FT	-16.9	-19.1	-16.3	-14.5	-16.7	-11.8	-14.9
PT	32.2	32.1	31.2	32.1	33.8	36.0	35.8

Table 3: Cosine similarity (%) of representations after tuning for each language. FT: fine-tuning; PT: prompt tuning. These checkpoints are on tuned on two English datasets: PAWS and MNLI.[3]

5.2 Cross-lingual Alignment After Tuning

We compute the averaged cosine similarity of all the 1000 translation pairs for each language pair <en , xx>, where xx is de, es, fr, ja, ko or zh. We also compute averaged cosine similarity of all the 1000*999/2 non-translations for each language pair. As shown in Table 3, both fine tuning and prompt tuning are doing well. Prompt tuning has the advantage in the sense that they change the representation more mildly, still have high cosine similarity on translation pairs. This resulted in more robust transfer and less overfitting.

5.3 Decision Boundaries

Prompt tuning keeps high cross-lingual alignment with fewer changes in the previous subsections. The general level of the learned representations' quality is still unknown, though. The learned representations quality are examined in this subsection.

Figure 2 (a) and (b) show t-SNE visualization of representations before two different tuning methods. Each dot in the two figures is a PAWS-X sample from four languages: German (de), zh (Chinese), en (English), ja (Japanese). The blue sample is a paraphrase, the orange sample is a non-paraphrase. Samples of the same language are

	en-de	en-es	en-fr	en-ja	en-ko	en-zh
Training on MNLI						
FT	81.5	85.4	83.0	71.8	68.2	73.9
FT-neg	52.6	53.1	52.8	51.5	50.6	50.0
rel-diff (%)	54.8	60.8	57.2	39.4	34.8	47.8
PT	96.4	97.3	96.6	94.8	93.8	95.0
PT-neg	91.0	91.1	90.8	90.5	90.1	90.2
rel-diff (%)	5.9	6.8	6.4	4.8	4.1	5.3
Training on PAWS						
FT	90.4	92.1	88.8	76.8	75.3	82.0
FT-neg	13.3	13.2	13.4	14.3	14.4	13.6
rel-diff (%)	580	598	563	437	423	503
PT	98.4	98.6	98.3	96.3	96.0	96.7
PT-neg	88.1	88.1	88.3	89.1	89.4	88.9
rel-diff (%)	11.7	11.9	11.3	8.1	7.4	8.8

Table 4: Cosine similarity (%) of translation pairs after tuning on two English dataset: MNLI and PAWS. "-neg" means the average cosine similarity of non-translations for each language pair. "rel-diff" means the relative difference between translation and non-translations. Two different tuning method are shown, one is fine-tuning (FT), the other is prompt tuning (PT).

grouped together. However, label information is missing from sample representations.

Figure 2 (c) and (d) shows t-SNE (van der Maaten and Hinton, 2008) visualization after fine tuning (FT) and prompt tuning (PT). After tuning, both have reasonable and nice separated representations. For each language, we also plot logistic regression decision boundary for these t-SNE embeddings. The decision boundaries for various languages vary significantly after fine tuning. The English decision boundary can not separate well on German samples. After prompt tuning, the decision boundaries of the four languages are surprisingly aligned well. This suggest that prompt tuning learns better language-independent classifier than fine tuning, although the tuning is only on English training set.

6 Related Work

Recently, several previous works show prompt tuning for multilingual language models. Winata et al. (2021) shows the multilingual skills of large pre-trained models with few examples. Zhao and Schütze (2021); Huang et al. (2022); Qi et al. (2022) shows new proposed prompt tuning methods. The goal of our work is different from theirs. We show prompt tuning is better than fine-tuning for cross-lingual evaluation. We have a conclusion that our prompt tuning achieves higher performance than fine-tuning consistently in the setting.

Previous work (Zhao and Schütze, 2021; Huang et al., 2022; Qi et al., 2022) only experimented on the sentence classification task. Hard sequence tagging tasks and question answering is not explored or the settings are in low resource regimes. We investigate cross-lingual transfer ability on various NLU tasks from XTREME (Hu et al., 2020), which is one of the important cross-lingual transfer evaluation benchmarks. Sentence classification, sequence labeling, and question answering are included.

7 Conclusion

In this work, we compared prompt tuning and fine tuning on cross-lingual understanding with multilingual languages models, finding that prompt tuning achieves a better performance. This suggest that it is promissing to use prompt tuning on cross-lingual transfer.

Limitations

In this work, we investigate the effects of prompt tuning on cross-lingual understanding and empirically demonstrate some promising outcomes. We need a lot of GPU resources to complete our experiments. The experiments on large size pretrained multilingual language models are conducted on A100s with 40G memory. Training can be accelerated by using large batches.

This is a preliminary exploration of prompt tuning on cross-lingual transfer. In this work, encoder-only models are explored on natural language understanding tasks in the paper. Future work may also involve encoder-decoder models and other tasks.

Acknowledgements

We would like to thank Salesforce AI Research team for helpful discussions, and the reviewers for insightful comments.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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 and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: Word-level Adversarial ReProgramming. 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 4921–4933, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. 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.

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*, abs/2202.11451.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. 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.

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.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. 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.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2021. VECO: Variable and flexible cross-lingual pre-training for language understanding and 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 3980–3994, Online. Association for Computational Linguistics.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, et al. 2018. Universal dependencies 2.2.

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.

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.

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 *Empirical Methods in Natural Language Processing (EMNLP)*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. 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. 2019. 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.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

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.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

A Appendix

A.1 More Training Details

For prompt tuning, we train with the Adam optimizer (Kingma and Ba, 2015) with no warmup step. Batch size is 32 for tasks, and with the exception of answering questions, which has a batch size of 8. Linear learning rate scheduler is used. We tune the learning rate in $\{5e-2, 1e-2, 5e-3, 1e-3, 5e-4, 1e-4\}$. We train all prompt tuning models for 30 epochs. Finally, tuned prompt length for MNLI is 32. It is 16 for the other tasks. We use A100s with 40G memory and all experiments can be done in few hours.

Method	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	**avg**
	88.2	77.4	82.3	82.6	81.1	83.7	82.0	75.2	79	71.0	76.7	77.5	71.4	79.1	78.6	79.1
	88.3	76.9	81.9	81.9	81.4	83.6	81.6	74.3	78.1	70.1	75.8	77.6	70.7	78.8	77.6	78.6
FT	88.1	77.5	82.4	81.8	81.3	83.4	82.6	75.0	78.9	70.3	75.6	78.1	70.8	78.5	78.3	78.8
	88.4	77.4	81.7	82.0	81.5	83.3	82.3	75.4	79.0	70.2	75.5	78.1	71.2	79.1	77.9	78.9
	88.2	77.6	82.5	81.7	80.9	83.2	81.9	75.1	78.2	69.5	76.5	77.6	71.0	78.8	78.6	78.8
	88.5	78.3	82.8	82.2	82.5	84.2	83.0	76.1	80.4	71.0	77.6	79.2	72.5	80.0	78.3	79.8
	88.7	78.7	82.9	82.1	82.8	84.3	83.2	76.1	80.4	71.0	77.6	79.2	72.5	80.0	78.3	79.8
PT	88.8	78.1	82.7	81.7	81.9	84.0	83.2	75.9	80.7	71.4	77.5	79.3	72.5	79.4	78.7	79.7
	89.1	79.2	83.2	82.1	82.4	84.1	83.0	76.2	80.8	70.7	77.7	79.5	72.5	79.9	78.4	79.9
	89.0	78.7	83.2	82.2	82.8	84.3	83.4	76.2	80.8	71.3	77.9	79.2	72.5	80.3	78.2	80.0

Table 5: XNLI accuracy scores for each language with fine-tuning (FT) and prompt tuning (PT).

Method	en	de	es	fr	ja	ko	zh	**avg**
	95.6	90.8	81.4	91.3	82.7	81.8	84.5	88.3
	95.7	90.5	91.0	91.3	81.7	81.2	84.0	87.9
FT	95.4	89.4	90.8	90.9	80.5	80.6	84.0	87.4
	95.4	90.2	90.6	90.5	80.6	80.4	83.4	87.2
	94.7	91.0	91.4	92.1	82.4	93.2	84.2	88.6
	96.2	92.3	91.4	92.1	81.3	83.2	84.8	88.8
	95.3	91.6	91.1	92.0	82.7	83.1	84.2	88.6
PT	95.4	90.9	91.4	91.8	82.1	82.8	84.7	88.4
	95.9	90.7	90.7	91.6	81.4	81.6	84.6	88.1
	95.6	91.6	90.5	91.7	82.2	81.7	83.0	88.0

Table 6: PAWS-X accuracy scores for each language with fine-tuning (FT) and prompt tuning (PT).

Method	en	es	de	el	ru	tr	ar	vi	th	zh	hi	**avg**
	75.2 / 87.2	61.2 / 80.7	62.7 / 82.5	60.2 / 78.7	63.5 / 80.1	57.8 / 74.3	58.6 / 75.5	59.9 / 79.4	59.9 / 73.2	58.7 / 68.5	56.6 / 74.7	61.3 / 77.7
	75.0 / 86.8	61.6 / 79.8	61.9 / 80.0	59.6 / 78.6	62.7 / 79.6	57.6 / 73.3	56.6 / 74.4	57.8 / 78.6	61.3 / 72.4	60 / 67.5	58.2 / 74.6	61.1 / 76.9
PT	75.5 / 87.0	64.0 / 81.3	64.8 / 80.9	62.4 / 80.0	63.8 / 80.1	57.7 / 73.8	55.9 / 72.8	60.2 / 79.5	62.3 / 73.4	61.4 / 69.6	59.8 / 76.1	62.5 / 77.7
	75.8 / 87.0	63.0 / 81.4	62.4 / 79.4	62.1 / 79.9	62.9 / 79.8	56.9 / 73.6	57.4 / 74.6	59.6 / 78.5	62.8 / 74.7	60.5 / 70.5	57.5 / 74.2	61.9 / 77.6
	76.0 / 87.4	62.8 / 80.8	65.0 / 80.2	61.2 / 78.3	63.1 / 79.5	56.3 / 72.3	57.3 / 73.9	57.6 / 77.5	62.9 / 71.6	61.0 / 68.7	58.4 / 74.2	62.0 / 76.8
	77.2 / 88.4	65.1 / 83.1	64.8 / 81.4	63.7 / 81.2	58.7 / 80.2	58.7 / 74.6	60.3 / 77.0	61.4 / 80.6	66.4 / 74.7	60.3 / 68.6	61.8 / 78.1	63.5 / 78.9
	77.4 / 88.5	64.4 / 82.3	64.8 / 81.2	63.5 / 80.8	64.7 / 80.7	58.3 / 74.1	60.3 / 76.8	61.0 / 80.3	66.6 / 75.0	61.7 / 70.2	61.5 / 77.5	64.0 / 78.9
PT	77.4 / 88.6	65.4 / 83.4	64.5 / 80.9	64.0 / 81.2	64.1 / 80.7	58.7 / 74.9	59.8 / 76.5	62.1 / 81.4	66.6 / 75.4	61.3 / 69.9	62.8 / 77.8	64.2 / 79.2
	77.1 / 88.5	64.7 / 82.9	63.9 / 80.7	62.7 / 80.5	64.7 / 80.4	59.2 / 74.6	59.7 / 76.3	60.8 / 80.7	66.6 / 74.6	61.0 / 69.1	61.6 / 77.6	64.7 / 78.7
	77.9 / 88.7	65.0 / 83.0	64.2 / 81.2	63.4 / 80.2	64.8 / 80.9	58.2 / 75.4	60.2 / 77.0	62.9 / 81.3	67.3 / 75.9	60.7 / 69.8	61.1 / 77.9	64.2 / 79.2

Table 7: XQuAD results (EM / F1) for each language with fine-tuning (FT) and prompt tuning (PT).

Method	en	ar	bn	fi	id	ko	ru	sw	te	avg
	60.5 / 74.2	51.5 / 71.5	50.4 / 68.6	51.0 / 67.6	62.5 / 78.6	49.6 / 60.9	45.3 / 67.7	44.7 / 65.7	56.7 / 75.7	46.2 / 70.0
	57.7 / 71.8	51.5 / 71.0	50.4 / 70.4	53.5 / 70.3	61.4 / 77.1	53.6 / 64.5	47.8 / 68.6	50.3 / 70.3	57.0 / 75.7	53.7 / 71.1
FT	57.7 / 73.2	51.9 / 72.5	48.7 / 66.4	53.6 / 69.7	59.8 / 77.0	50.7 / 59.6	50.7 / 68.0	49.1 / 69.1	58.0 / 77.8	53.4 / 70.4
	58.6 / 71.7	53.0 / 71.6	46.0 / 62.5	53.7 / 68.4	59.8 / 75.7	52.5 / 63.0	40.4 / 65.2	48.9 / 69.2	58.4 / 76.3	52.4 / 69.3
	59.8 / 72.3	48.3 / 70.6	52.2 / 68.6	49.7 / 67.5	60.4 / 77.7	55.1 / 65.5	38.9 / 65.2	45.4 / 66.6	56.8 / 75.4	51.8 / 69.9
	61.8 / 75.0	53.7 / 72.3	48.7 / 67.0	58.2 / 73.0	63.0 / 77.9	52.9 / 63.6	50.2 / 70.0	47.5 / 68.5	57.5 / 75.3	54.8 / 71.5
	60.7 / 74.0	53.1/72.2	45.1 / 64.5	55.9 / 71.8	63.5 / 78.3	51.8 / 61.9	52.3 / 71.0	48.9 / 68.9	58.4 / 76.2	54.4 / 71.0
PT	60.2 / 73.6	54.8 / 73.9	52.2 / 70.0	56.6 / 71.4	64.8 / 78.7	52.5 / 62.3	53.1 / 71.4	51.1 / 70.7	61.6 / 79.1	56.3 / 72.3
	62.0 / 75.3	53.6 / 73.0	46.0 / 64.9	57.3 / 71.3	63.7 / 78.6	53.3 / 62.0	52.7 / 71.8	48.1 / 69.0	58.7 / 75.5	55.0 / 71.3
	61.4 / 74.5	54.9 / 72.8	46.9 / 66.3	56.8 / 71.4	63.2 / 77.6	54.3 / 63.0	53.1 / 71.1	47.9 / 68.4	58.6 / 76.4	55.2 / 71.3

Table 8: TyDiQA-GoldP results (EM / F1) for each language with fine-tuning (FT) and prompt tuning (PT).