Soft Prompt Tuning for Cross-Lingual Transfer: When Less is More

Fred Philippy[1,2], **Siwen Guo**[1], **Shohreh Haddadan**[1],
Cedric Lothritz[2], **Jacques Klein**[2], **Tegawendé F. Bissyandé**[2]

[1] Zortify S.A., Luxembourg
[2] University of Luxembourg, Luxembourg
{fred, siwen}@zortify.com, shohreh.haddadan@gmail.com
{cedric.lothritz, jacques.klein, tegawende.bissyande}@uni.lu

Abstract

Soft Prompt Tuning (SPT) is a parameter-efficient method for adapting pre-trained language models (PLMs) to specific tasks by inserting learnable embeddings, or soft prompts, at the input layer of the PLM, without modifying its parameters. This paper investigates the potential of SPT for cross-lingual transfer. Unlike previous studies on SPT for cross-lingual transfer that often fine-tune both the soft prompt and the model parameters, we adhere to the original intent of SPT by keeping the model parameters frozen and only training the soft prompt. This does not only reduce the computational cost and storage overhead of full-model fine-tuning, but we also demonstrate that this very parameter efficiency intrinsic to SPT can enhance cross-lingual transfer performance to linguistically distant languages. Moreover, we explore how different factors related to the prompt, such as the length or its reparameterization, affect cross-lingual transfer performance.

1 Introduction

Fine-tuning pre-trained language models (PLMs) on task-specific labeled data requires large amounts of computational resources and may cause catastrophic forgetting of the pre-trained knowledge (Goodfellow et al., 2015). In multilingual settings, this may lead to poor cross-lingual transfer performance (Vu et al., 2022).

To address these challenges, Lester et al. (2021) introduced Soft Prompt Tuning (SPT), a method that inserts learnable embeddings, or soft prompts, at the PLM's input layer. The PLM then makes predictions using the output of its pre-trained language modeling head. The key advantage of SPT lies in its ability to leverage the pre-existing knowledge within PLMs while reducing the reliance on extensive task-specific fine-tuning. SPT has been shown to achieve remarkable results in various monolingual downstream tasks, especially in few-shot settings.

Motivated by this success, some recent works have also explored the use of SPT for cross-lingual transfer, where the goal is to leverage a multilingual language model (MLLM) to transfer knowledge from a high-resource to a low-resource language. However, these works have not fully exploited the potential of SPT. Some have appended a newly initialized classifier to the model (Tu et al., 2022; Park et al., 2023), hindering the suitability of SPT for few-shot learning. Others have fine-tuned the entire model along with the prompt (Zhao and Schütze, 2021; Huang et al., 2022), which reduces the computational efficiency of SPT.

This is especially problematic given the growing size of state-of-the-art language models. Therefore, we explore the impact on SPT's cross-lingual transfer performance when adhering to the original methodology of Lester et al. (2021), which involves fine-tuning only the soft prompt while keeping all model parameters frozen. Specifically, this paper contributes to the field of cross-lingual SPT by:

- Investigating the impact of model freezing on the cross-lingual transfer performance of few-shot SPT.
- Demonstrating that by freezing the model, SPT achieves enhanced cross-lingual transfer, especially to languages linguistically distant from the source language.
- Exploring further non-linguistic factors that influence the cross-lingual transfer performance of SPT, in particular prompt length and prompt reparameterization.

In this study, we conduct experiments on a topic classification dataset in 52 different languages and using 4 different models in few-shot settings. We believe that our findings can improve the existing methods that aim to enhance cross-lingual SPT, particularly in the context of current state-of-the-art models with billions of parameters where parameter efficiency is crucial.

Proceedings of the 1st Workshop on Modular and Open Multilingual NLP (MOOMIN 2024), pages 7–15

2 Related Work

Lester et al. (2021) proposed SPT, a method to leverage a PLM's pre-trained language modeling head without appending a new classifier. SPT relies on a soft prompt, which is a set of learnable embeddings that are concatenated with the input sequence, and keeps all other model parameters frozen. Since then, several recent works have explored the use of soft prompts for MLLMs. Zhao and Schütze (2021) first show that SPT outperforms fine-tuning in few-shot scenarios for cross-lingual transfer. Huang et al. (2022) introduce a method to train a language-agnostic soft prompt. However, unlike our study, none of these works on cross-lingual SPT employ model parameter freezing, leading to a reduced efficiency in their methods. In contrast, Tu et al. (2022) and Park et al. (2023) perform model freezing and, in corroboration with Zhao and Schütze (2021), also show that SPT outperforms fine-tuning for cross-lingual transfer. However, they append a newly initialized classification head to the model instead of using the PLM's pre-trained language modeling head, which diverges from the original idea of SPT. This setup is unsuitable for few-shot learning, requiring experiments to be conducted in full-data settings. In addition, prior studies often focus on smaller ranges of languages, which impedes making conclusive observations about SPT's cross-lingual tendencies across different languages and language families.

3 Experimental Setup

Besides adhering to the original setup of SPT, enabling parameter-efficient and data-efficient training, our study also sets itself apart in its objectives from the existing literature. Rather than simply demonstrating superior cross-lingual transfer performance of SPT over fine-tuning, our research aims to show that the minimal impact on the MLLM's representation space not only generally enhances transfer performance but is particularly effective for linguistically distant languages.

We provide more specific details on our experimental setup in Appendix A.

3.1 Soft Prompt

Following Lester et al. (2021), we append a soft prompt to the input sequence which is passed through an autoregressive language model, generating the logits for the next token in the input sequence. Each class is linked to a token from the model's vocabulary, enabling us to map the token with the highest logit to the predicted class. Such a mapping is referred to as the *verbalizer* (Figure 1).

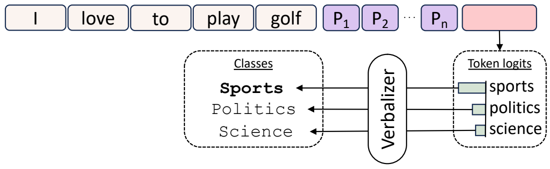

Figure 1: A simplified illustration of SPT (Lester et al., 2021). $P_1, \ldots, P_n$ denote the soft prompt tokens, with each token corresponding to a trainable embedding. Essentially, for a model with an embedding dimension d, a soft prompt of length n forms a $d \times n$ matrix.

3.2 Implementation Details

Models With the recent advancement and popularity of autoregressive language models for various tasks, our research is conducted using two types of MLLMs based on this architecture: XGLM (Lin et al., 2022) and BLOOM (Scao et al., 2022). For both models we use 2 different sizes: XGLM-564M and XGLM-1.7B for XGLM, and BLOOM-560M and BLOOM-1.1B for BLOOM.

Data In our study, we use SIB-200 (Adelani et al., 2023), a topic classification dataset containing seven distinct topics and covering a diverse range of 200 languages and dialects. We chose this dataset for its broader, more diverse language range compared to prior studies on cross-lingual SPT, covering almost all languages our models support, enabling more comprehensive observations.

Technical Details We compare two different settings: tuning the soft prompt with model freezing (*w/ MF*) and without model freezing (*w/o MF*). We perform few-shot fine-tuning only using English samples. The final cross-lingual transfer performance is then evaluated on the test sets of all languages supported by the respective model (30 for XGLM, 38 for BLOOM), using accuracy as the metric. We repeat each experiment 4 times with different random seeds and report the mean.

4 Results

We provide the full results across all models and languages in Appendix D. The results reveal that model freezing not only **boosts cross-lingual transfer performance** (Figure 2) but additionally is a step towards **closing the transfer gap** between linguistically distant and similar languages. This

		DATA	SYN	GEO	INV	GEN	PHON	FEA
BLOOM-560M	*w/o MF*	0,6781	0,6457	0,2294	0,3779	0,5081	**0,4343**	0,4221
	w/ MF	**0,6080**	**0,5742**	**0,2034**	**0,2629**	**0,3676**	0,4482	**0,3165**
BLOOM-1.1B	*w/o MF*	0,6788	0,6403	0,1693	0,4605	0,5679	0,5272	-0,4685
	w/ MF	**0,4856**	**0,4177**	**0,0290**	**0,2930**	**0,3711**	**0,4283**	**0,3002**
XGLM-564M	*w/o MF*	0,2672	0,6767	0,4694	0,4016	0,3203	0,4756	0,5949
	w/ MF	**0,2453**	**0,6574**	**0,2551**	**0,3410**	**0,2201**	**0,3285**	**0,5185**
XGLM-1.7B	*w/o MF*	0,2636	0,6722	**0,2566**	0,3623	0,2924	0,3213	0,5315
	w/ MF	**0,2560**	**0,6694**	0,2949	**0,3155**	**0,2786**	**0,2779**	**0,4922**

Table 1: Pearson correlation between (8-shot) cross-lingual transfer performance and 6 different linguistic similarity metrics, namely syntactic (SYN), geographic (GEO), inventory (INV), genetic (GEN), phonological (PHON) and featural (FEA) distance, as well as the language-specific pre-training corpus size (DATA).

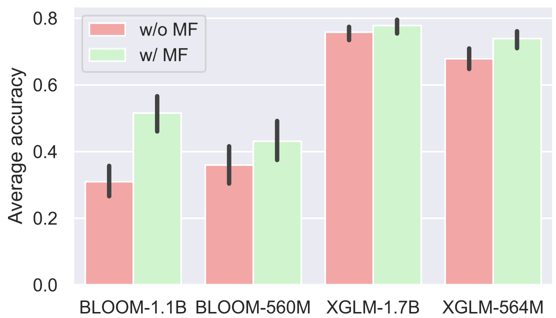

Figure 2: Average cross-lingual transfer performance of SPT with and without model freezing (MF) for different models across all languages supported by the respective model.

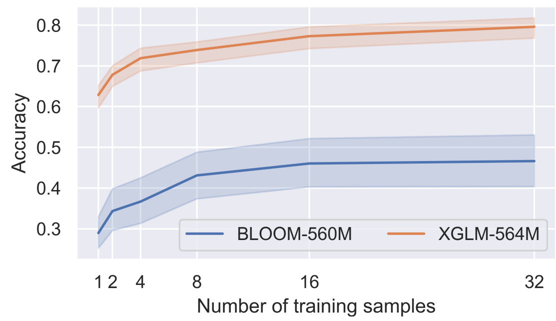

Figure 3: Average cross-lingual transfer performance of SPT with model freezing for different number of training samples per class.

can be seen in Table 1, which shows that the correlation strength between transfer performance and language similarity between source and target languages, measured using 6 different similarity metrics[1] (Littell et al., 2017), decreases when freezing model parameters. This suggests that the parameter efficiency of SPT mitigates the bias of cross-lingual transfer towards linguistically similar languages. In other words, by fine-tuning fewer parameters, cross-lingual transfer, especially to linguistically distant languages, is enhanced. This improvement over full-model fine-tuning may be attributed to the reduced impact on the MLLM's representation space during fine-tuning (Philippy et al., 2023).

Figure 3 also shows that, despite the limited number of tunable parameters when freezing all model parameters, additional training samples further boost cross-lingual transfer performance.

Parameter efficiency Besides better cross-lingual transfer performance, model freezing during SPT also provides parameter efficiency as fine-tuning is restricted to a number of soft prompt tokens, resulting in only a few thousand parameters in total. This is less than 0.01% of the parameters fine-tuned in previous studies (Zhao and Schütze, 2021; Huang et al., 2022).

For illustration, the storage requirement for a copy of the XGLM-1.7B model is approximately 3.2 GB, whereas a prompt needs less than 100KB. With respect to training duration, our observations indicate that the time required for training only the soft prompt is less than half compared to when training all model parameters. This benefit becomes even more pronounced when considering the increasing sizes of state-of-the-art models.

5 Impact of Prompt Length and Reparameterization

5.1 Prompt Length

Using the same configuration as described in Section 3.2, we compare the transfer performance of prompts with different lengths under the 8-shot setting. We consider prompt lengths in

[1] See Appendix B for more details.

$\{1, 2, 5, 10, 20, 30\}$ and report the results for all 4 models. Figure 4 shows that **if a soft prompt is too long, cross-lingual transfer performance degrades**.

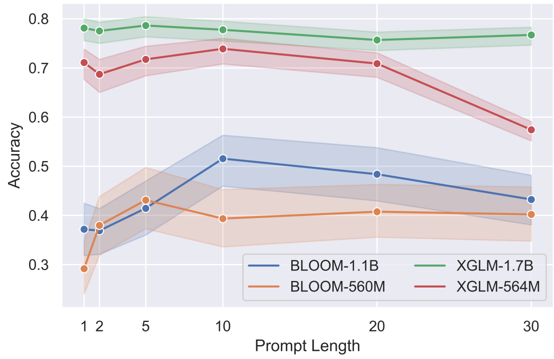

Figure 4: Average cross-lingual transfer performance, measured as accuracy, across different prompt lengths for different models.

5.2 Reparameterization

Direct fine-tuning of soft prompt embeddings may lead to unstable training and potentially reduces performance. To address this issue, previous works have proposed reparameterizing prompt embeddings using different architectures, such as an LSTM (Liu et al., 2021) or MLP (Li and Liang, 2021), which are fine-tuned along with the prompt embeddings. Liu et al. (2022) argue that reparameterization can also have negative effects depending on the task or dataset.

Motivated by this observation, we investigate the effect of reparameterization on cross-lingual transfer performance. We adopt the approach proposed by Razdaibiedina et al. (2023), which uses an MLP with a residual connection and a "bottleneck" layer for reparameterization. We provide further details on this method in Appendix C.

Our analysis reveals that BLOOM is significantly more affected by reparameterization than XGLM (Figure 5 in Appendix C). For both models, the **impact of reparameterization differs across languages** — being detrimental for some and advantageous for others. Notably, for BLOOM, Atlantic-Congo languages such as Yoruba, Twi, Kinyarwanda, Akan, Fon and Swahili experience the most significant performance decline due to reparameterization, with drops between 24% to 31%. Conversely, Indo-Aryan languages like Urdu, Hindi, Bengali, and Nepali, along with Dravidian languages like Malayalam and Tamil see the most significant improvements, with gains of up to 29%. For XGLM, the outcomes are more balanced. Nonetheless, we observe that the languages that benefit most from reparameterization either use Latin script, such as Haitian, German, and Turkish, or are Dravidian languages such as Telugu and Tamil.

Hence, we recommend that in cross-lingual settings, the decision to use or abstain from reparameterization should not be made uniformly. Instead, it should be tailored based on the specific target languages or language families in consideration.

6 Discussion

Previous works on SPT for cross-lingual transfer in few-shot settings suffers from two major drawbacks: 1) fine-tuning all model parameters along with the prompt reduces the computational efficiency of SPT; 2) a bias towards target languages that are linguistically closer to the source language. Our study tackles these issues by showing that by simply keeping model parameters frozen during SPT, we can make progress in addressing both these challenges.

Through our experiments, which covered a wider and more diverse range of languages than prior work on cross-lingual SPT, we observed intriguing effects of non-linguistic variables (such as model freezing, prompt length, and reparameterization) on the transfer performance for individual languages. Additionally, our results reveal language-specific differences that invite further inquiry into the possibility of tailoring prompts to the target language (e.g., applying prompt reparameterization or not depending on the linguistic distance between the target language family and the source language) rather than using a single prompt for universal transfer across languages. We believe that our findings will benefit future work on cross-lingual SPT and potentially improve the existing techniques (Huang et al., 2022), becoming more valuable as we adopt larger state-of-the-art models with billion- and trillion-scale parameters (Lester et al., 2021).

7 Conclusion

The objective of our study was to examine the impact of model freezing on the cross-lingual transfer performance of SPT. Our results demonstrate that SPT, a method that adjusts less than 0.01% of parameters compared to full-model fine-tuning, achieves comparable or superior performance for most target languages, particularly for those that are linguistically more distant. Furthermore, we found that shorter prompts enhance SPT's cross-lingual transfer performance, and that some target language families benefit from reparameterization while others are adversely affected by it.

Limitations

Our approach enhances transfer performance for several languages, especially those that are linguistically more distant. However, we also notice that it lowers the performance for some languages that are linguistically more similar. This limitation motivates us to pursue future research that aims to achieve balanced performance across languages

Another limitation of our approach is the instability of few-shot fine-tuning, which compromises the robustness of our method's evaluation. To mitigate this issue, we ran all experiments four times with different random seeds and reported the mean and variance of the results. However, we acknowledge that more research is needed to address the challenges of few-shot fine-tuning.

Ethics Statement

In this paper, we aim to improve the performance of MLLMs on low-resource languages, which often suffer from a lack of data and attention in NLP research. We believe that this is an important and ethical goal, as it enables NLP advances to benefit a broader range of language communities.

In addition, this paper aims to promote parameter efficiency, which is a crucial factor for reducing the computational and environmental costs of training and fine-tuning state-of-the-art language models. We believe that this aspect will enhance the accessibility and affordability of these models for researchers and practitioners who face computational constraints.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects. ArXiv:2309.07445 [cs].

Chris Collins and Richard Kayne. 2011. *Syntactic Structures of the World's Languages*. New York University, New York.

Matthew S. Dryer and Martin Haspelmath. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2015. An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks. ArXiv:1312.6211 [cs, stat].

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2015. *Glottolog 2.6*. Max Planck Institute for the Science of Human History, Jena.

Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei, and Houfeng Wang. 2022. Zero-shot Cross-lingual Transfer of Prompt-based Tuning with a Unified Multilingual Prompt. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11488–11497, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. 2015. *Ethnologue: Languages of the World, Eighteenth edition*. SIL International, Dallas, Texas.

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.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot Learning with Multilingual Generative Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. 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.

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.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT Understands, Too. ArXiv:2103.10385 [cs].

Steven Moran, Daniel McCloy, and (eds.). 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. ArXiv:2207.04672 [cs].

Nohil Park, Joonsuk Park, Kang Min Yoo, and Sungroh Yoon. 2023. On the Analysis of Cross-Lingual Prompt Tuning for Decoder-based Multilingual Model. ArXiv:2311.07820 [cs].

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Identifying the Correlation Between Language Distance and Cross-Lingual Transfer in a Multilingual Representation Space. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 22–29, Dubrovnik, Croatia. Association for Computational Linguistics.

Anastasiia Razdaibiedina, Yuning Mao, Madian Khabsa, Mike Lewis, Rui Hou, Jimmy Ba, and Amjad Almahairi. 2023. Residual Prompt Tuning: improving prompt tuning with residual reparameterization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6740–6757, Toronto, Canada. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *CoRR*, abs/2211.05100. ArXiv: 2211.05100.

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.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming Catastrophic Forgetting in Zero-Shot Cross-Lingual Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates. 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.

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.

A Reproducibility

We provide the code used for our experiments here: https://github.com/fredxlpy/cross_lingual_prompt_tuning.

A.1 Dataset

Our experiments are based on the SIB-200 dataset (Adelani et al., 2023). The dataset is based on the FLORES-200 benchmark (NLLB Team et al., 2022), and consists of 701 training, 99 validation and 204 test samples in each of the 203 languages. The task is to classify each sample into one of the 7 potential categories: `science/technology`, `travel`, `politics`, `sports`, `health`, `entertainment`, and `geography`.

A.2 Models

We provide additional information about the models used in our study in Table 2.

Model	Layers	Parameters	Hidden size	Vocab size
BLOOM-560M	24	560M	1.024	250.880
BLOOM-1.1B		1.1B	1.536	
XGLM-564M		564M	1.024	256.008
XGLM-1.7B		1.7B	2.048	

Table 2: Technical details of the models used in our study.

A.3 Technical Details

		Batch size	Learning rate	Prompt length
w/ MF	XGLM 564M	8	0.1	10
	XGLM 1.7B			
	BLOOM 560M			5
	BLOOM 1.1B			10
w/o MF	XGLM 564M	8	5e-6	10
	XGLM 1.7B			
	BLOOM 560M		1e-6	5
	BLOOM 1.1B			10

Table 3: Hyperparameters used in all of our experiments.

We conducted all of our experiments using the *Transformers* library (Wolf et al., 2020). In a k-shot setting, we fine-tune on k samples per class from the English train set and use $\frac{k}{4}$ samples per class for validation. We train all models and prompts for 20 epochs and select the best checkpoint on the development set. The different hyperparameters used in our experiments are provided in Table 3.

A.4 Soft Prompt

We follow the approach of Lester et al. (2021) and freeze all model parameters and only fine-tune the soft prompt.

In order to map the tokens predicted by the model to the respective class, we define a verbalizer $F : T \rightarrow C$, where $T = \{t_1, \ldots, t_K\}$ is a subset of the model's vocabulary V and $C = \{1, \ldots, K\}$ are the respective classes.

We append a prompt $p = \{p_1, \ldots, p_m\}$ to an input sequence $x = \{x_1, \ldots, x_n\}$ and pass $\{x_1, \ldots, x_n, p_1, \ldots, p_m\}$ through the autoregressive language model which outputs the logits for the next token in the input sequence $\{l_1, \ldots, l_{|V|}\}$.

The predicted token is then $F\,(argmax_{i \in T} l_i)$

A.5 Computing Resources

We conduct all our experiments on 4 A100 40GB GPUs, using 4 different random seeds, in parallel. All experiments could be run in a few hours.

B Language Distance Metrics

We consider six types of lang2vec[2] (Littell et al., 2017) distances:

- **Syntactic Distance** (SYN) captures the similarity of syntactic structures across languages. It is computed as the cosine distance between syntax feature vectors, which are derived from the World Atlas of Language Structures[3] (WALS) (Dryer and Haspelmath, 2013), Syntactic Structures of World Languages[4] (SSWL) (Collins and Kayne, 2011) and Ethnologue[5] (Lewis et al., 2015).

- **Geographic Distance** (GEO) reflects the spatial proximity of languages. It is calculated as the shortest distance between two languages on the surface of the earth's sphere (i.e., orthodromic distance).

- **Inventory Distance** (INV) measures the difference in sound inventories across languages. It is computed as the cosine distance between inventory feature vectors, which are obtained

[2] https://github.com/antonisa/lang2vec
[3] https://wals.info
[4] http://sswl.railsplayground.net/
[5] https://www.ethnologue.com/

from the PHOIBLE[6] database (Moran et al., 2019).

- **Genetic Distance** (GEN) indicates the historical relatedness of languages. It is based on the Glottolog[7] (Hammarström et al., 2015) tree of language families and is obtained by computing the distance between two languages in the tree.

- **Phonological Distance** (PHON) captures the similarity of sound patterns across languages. It is computed as the cosine distance between phonological feature vectors, which are sourced from WALS and Ethnologue.

- **Featural Distance** (FEA) is the cosine distance between feature vectors from a combination of the 5 above-listed linguistic features.

The values for each distance type range from 0 to 1, where 0 indicates the minimum distance and 1 indicates the maximum distance.

C Prompt Reparameterization

We follow the residual reparameterization method of Razdaibiedina et al. (2023) to examine the impact of soft prompt reparameterization. This method employs a multi-layer perceptron (MLP) architecture for the reparameterization network, which consists of a *down-projection* layer and an *up-projection* layer with parameter $W_{down} \in \mathbb{R}^{d \times m}$ and $W_{up} \in \mathbb{R}^{m \times d}$ respectively, where d denotes the model embedding size and m denotes the hidden representation dimension between both layers (*bottleneck size*). A ReLU layer is applied to the hidden representation, and a normalization layer is applied to the output of the *up-projection* layer before summing it with the initial input embedding via a residual connection. We fine-tune the soft prompt and its reparameterization network with a bottleneck size of 500 for BLOOM-560M and 200 for XGLM-564M and report the impact of reparameterization across all target languages in Figure 5. Except for the reparameterization, we adopt the same implementation settings as described in Section 3.

D Full Results

The full results discussed in Section 4 are provided in Table 4.

[6] https://phoible.org/
[7] https://glottolog.org

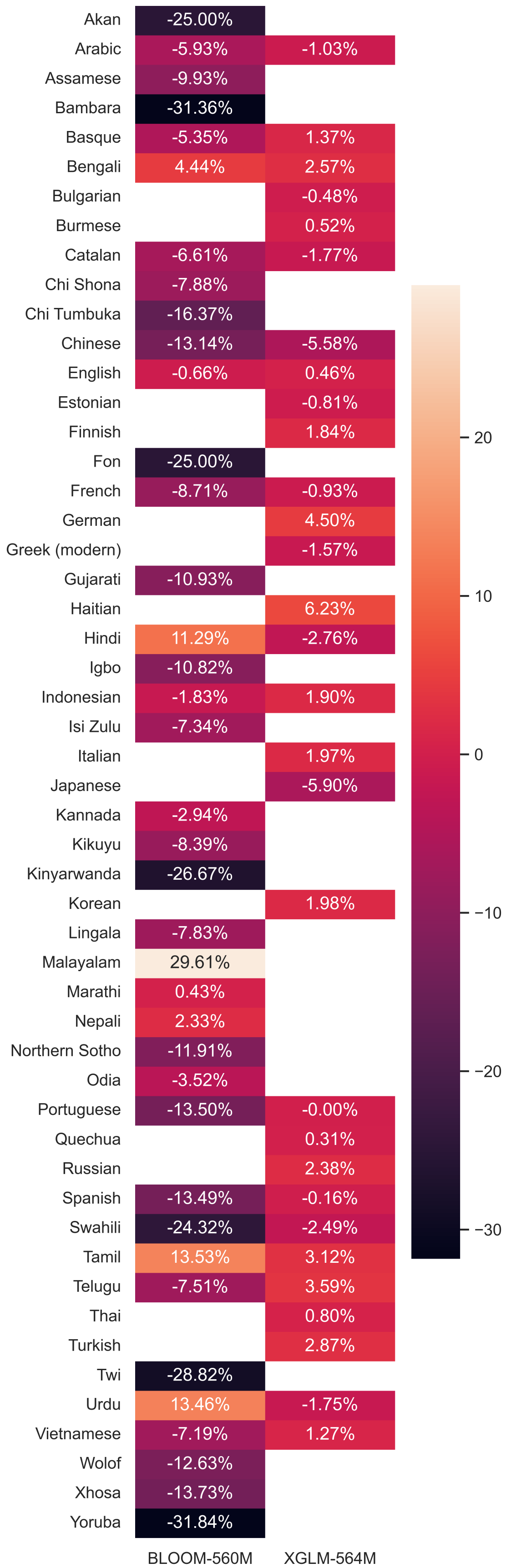

Figure 5: Impact of reparameterization (expressed in %) on the cross-lingual transfer performance of BLOOM-560M and XGLM-564M for different target languages.

Language	BLOOM-560M w/o MF	BLOOM-560M w/ MF	BLOOM-1.1B w/o MF	BLOOM-1.1B w/ MF	XGLM-564M w/o MF	XGLM-564M w/ MF	XGLM-1.7B w/o MF	XGLM-1.7B w/ MF
Akan	22,18$_{9,67}$	**34,80$_{6,33}$**	19,36$_{6,52}$	**35,05$_{0,85}$**	-	-	-	-
Arabic	55,51$_{3,56}$	**70,22$_{1,62}$**	42,03$_{11,9}$	**63,48$_{3,50}$**	57,60$_{3,34}$	**71,69$_{1,89}$**	74,75$_{1,67}$	**78,68$_{5,49}$**
Assamese	27,45$_{7,99}$	**37,01$_{4,85}$**	29,41$_{9,66}$	**53,06$_{4,92}$**	-	-	-	-
Bambara	16,67$_{3,63}$	**26,96$_{8,74}$**	17,03$_{5,94}$	**29,17$_{3,68}$**	-	-	-	-
Basque	43,50$_{12,5}$	**61,89$_{1,90}$**	38,73$_{7,69}$	**63,97$_{13,0}$**	67,40$_{1,30}$	**71,32$_{2,70}$**	71,08$_{3,07}$	**72,43$_{6,64}$**
Bengali	56,62$_{4,48}$	**60,78$_{2,41}$**	46,69$_{12,2}$	**71,81$_{2,90}$**	68,14$_{3,51}$	**71,45$_{4,24}$**	71,57$_{3,05}$	**76,23$_{5,22}$**
Bulgarian	-	-	-	-	72,79$_{5,13}$	**77,33$_{2,45}$**	78,92$_{2,40}$	**81,37$_{4,33}$**
Burmese	-	-	-	-	63,60$_{6,06}$	**71,20$_{3,38}$**	72,67$_{3,03}$	**73,41$_{7,79}$**
Catalan	63,48$_{13,1}$	**72,30$_{2,28}$**	48,77$_{7,93}$	**73,77$_{4,03}$**	68,50$_{7,52}$	**76,35$_{3,03}$**	77,33$_{4,01}$	**79,04$_{4,13}$**
Chi Shona	19,98$_{4,79}$	**24,88$_{2,67}$**	17,89$_{5,69}$	**31,00$_{3,95}$**	-	-	-	-
Chi Tumbuka	20,34$_{4,54}$	**27,70$_{2,55}$**	18,14$_{4,95}$	**33,70$_{4,62}$**	-	-	-	-
Chinese	60,54$_{11,1}$	**73,65$_{6,47}$**	47,30$_{13,9}$	**72,43$_{3,36}$**	59,93$_{8,33}$	**79,04$_{1,85}$**	77,94$_{5,08}$	**81,74$_{4,28}$**
English	**75,00$_{5,87}$**	74,63$_{2,09}$	69,36$_{2,67}$	**75,12$_{2,90}$**	78,68$_{1,67}$	**79,90$_{2,62}$**	80,88$_{2,94}$	**82,84$_{5,41}$**
Estonian	-	-	-	-	72,30$_{3,24}$	**75,86$_{3,13}$**	76,35$_{1,76}$	**81,13$_{5,78}$**
Finnish	-	-	-	-	76,72$_{1,81}$	**79,90$_{1,44}$**	79,78$_{1,76}$	**82,35$_{5,92}$**
Fon	19,36$_{10,0}$	**25,49$_{7,88}$**	13,97$_{3,98}$	**26,84$_{5,51}$**	-	-	-	-
French	69,61$_{6,52}$	**73,16$_{1,89}$**	57,23$_{6,29}$	**72,92$_{5,51}$**	71,94$_{4,26}$	**79,29$_{2,98}$**	79,04$_{5,48}$	**79,90$_{2,80}$**
German	-	-	-	-	71,57$_{7,19}$	**76,23$_{4,67}$**	81,62$_{5,04}$	81,62$_{5,79}$
Greek (modern)	-	-	-	-	73,90$_{3,47}$	**78,19$_{2,93}$**	80,27$_{2,70}$	**82,97$_{5,11}$**
Gujarati	**41,79$_{7,92}$**	37,01$_{9,35}$	27,08$_{7,85}$	**54,29$_{10,3}$**	-	-	-	-
Haitian	-	-	-	-	65,44$_{1,30}$	**68,87$_{2,55}$**	74,39$_{1,72}$	**74,75$_{6,70}$**
Hindi	42,52$_{4,28}$	**45,59$_{4,47}$**	50,12$_{10,0}$	**64,95$_{2,85}$**	74,14$_{3,28}$	**75,37$_{2,41}$**	75,74$_{2,95}$	**78,19$_{4,88}$**
Igbo	18,50$_{1,57}$	**23,77$_{6,42}$**	15,20$_{4,85}$	**27,70$_{4,57}$**	-	-	-	-
Indonesian	49,26$_{2,55}$	**66,91$_{1,86}$**	49,14$_{11,9}$	**68,75$_{3,38}$**	73,90$_{1,29}$	**77,57$_{2,45}$**	77,21$_{2,48}$	**79,90$_{5,34}$**
Isi Zulu	19,24$_{6,01}$	**21,69$_{2,72}$**	15,69$_{5,98}$	**29,66$_{2,48}$**	-	-	-	-
Italian	-	-	-	-	73,41$_{4,82}$	**74,75$_{1,52}$**	78,43$_{4,95}$	**80,02$_{5,52}$**
Japanese	-	-	-	-	54,29$_{5,98}$	**76,84$_{3,89}$**	**80,64$_{1,47}$**	77,94$_{4,65}$
Kannada	22,30$_{8,24}$	**25,00$_{8,46}$**	22,92$_{3,85}$	**55,76$_{7,93}$**	-	-	-	-
Kikuyu	28,19$_{8,36}$	**35,05$_{2,42}$**	19,49$_{4,44}$	**33,70$_{3,81}$**	-	-	-	-
Kinyarwanda	19,00$_{3,21}$	**25,74$_{6,26}$**	15,69$_{3,80}$	**30,39$_{4,33}$**	-	-	-	-
Korean	-	-	-	-	73,77$_{1,67}$	**74,26$_{2,28}$**	74,75$_{4,46}$	**77,45$_{5,41}$**
Lingala	23,90$_{3,85}$	**28,19$_{4,74}$**	21,69$_{8,43}$	**36,15$_{3,29}$**	-	-	-	-
Malayalam	**23,53$_{11,1}$**	21,94$_{7,56}$	30,39$_{9,95}$	**59,93$_{4,17}$**	-	-	-	-
Marathi	**34,68$_{11,1}$**	28,31$_{5,83}$	29,78$_{6,21}$	**60,05$_{4,41}$**	-	-	-	-
Nepali	30,15$_{6,99}$	**42,03$_{6,95}$**	36,76$_{13,3}$	**67,03$_{6,25}$**	-	-	-	-
Northern Sotho	20,59$_{6,62}$	**28,80$_{0,47}$**	18,38$_{4,09}$	**33,82$_{2,40}$**	-	-	-	-
Odia	**34,80$_{7,64}$**	31,37$_{6,62}$	25,00$_{5,25}$	**47,06$_{9,22}$**	-	-	-	-
Portuguese	66,67$_{5,02}$	**75,37$_{3,19}$**	53,19$_{5,69}$	**73,77$_{2,17}$**	74,26$_{1,90}$	**79,53$_{1,09}$**	80,15$_{1,98}$	**82,48$_{3,95}$**
Quechua	-	-	-	-	35,66$_{8,69}$	**39,71$_{2,23}$**	49,88$_{4,84}$	**51,59$_{6,37}$**
Russian	-	-	-	-	76,96$_{3,23}$	**77,21$_{1,98}$**	78,19$_{3,43}$	**80,27$_{4,30}$**
Spanish	63,36$_{8,94}$	**72,67$_{0,47}$**	46,69$_{9,79}$	**73,65$_{5,11}$**	71,45$_{0,74}$	**76,47$_{2,30}$**	77,33$_{3,63}$	**79,78$_{4,84}$**
Swahili	35,05$_{7,95}$	**49,88$_{6,02}$**	25,12$_{6,22}$	**49,75$_{7,40}$**	61,40$_{8,61}$	**69,00$_{2,84}$**	**73,77$_{2,45}$**	72,79$_{7,91}$
Tamil	44,85$_{9,58}$	**50,74$_{4,09}$**	34,44$_{13,1}$	**67,40$_{4,71}$**	68,75$_{5,68}$	**70,59$_{2,12}$**	73,90$_{1,01}$	**75,86$_{7,91}$**
Telugu	24,51$_{3,94}$	**31,00$_{6,71}$**	26,96$_{1,20}$	**66,05$_{7,13}$**	62,75$_{3,33}$	**68,26$_{5,15}$**	74,14$_{3,76}$	**76,23$_{6,46}$**
Thai	-	-	-	-	67,77$_{6,42}$	**76,35$_{1,16}$**	**79,53$_{1,72}$**	77,33$_{5,02}$
Turkish	-	-	-	-	73,16$_{2,84}$	**76,96$_{3,18}$**	74,63$_{4,30}$	**79,17$_{5,89}$**
Twi	23,41$_{9,5}$	**35,29$_{6,64}$**	18,75$_{6,83}$	**36,52$_{3,32}$**	-	-	-	-
Urdu	42,28$_{6,67}$	**44,61$_{8,95}$**	31,74$_{8,12}$	**48,41$_{9,35}$**	54,90$_{8,37}$	**70,10$_{2,86}$**	70,10$_{3,12}$	**75,25$_{5,69}$**
Vietnamese	46,08$_{19,4}$	**68,14$_{7,21}$**	43,87$_{7,49}$	**64,58$_{3,76}$**	70,71$_{3,06}$	**76,96$_{3,18}$**	78,31$_{3,63}$	**79,90$_{7,30}$**
Wolof	25,49$_{7,88}$	**34,93$_{4,17}$**	21,81$_{9,77}$	**41,42$_{4,64}$**	-	-	-	-
Xhosa	21,94$_{7,35}$	**28,55$_{1,23}$**	15,32$_{5,51}$	**32,23$_{6,14}$**	-	-	-	-
Yoruba	13,36$_{1,62}$	**21,94$_{9,49}$**	16,30$_{2,28}$	**33,21$_{4,76}$**	-	-	-	-

Table 4: Cross-lingual transfer results, reported as accuracy, along with standard deviation across 4 runs, after 8-shot soft prompt tuning (SPT) in English, with and without model freezing (MF).