BREAKING PHYSICAL AND LINGUISTIC BORDERS: MULTILINGUAL FEDERATED PROMPT TUNING FOR LOW-RESOURCE LANGUAGES

Wanru Zhao[1]*, **Yihong Chen**[2], **Royson Lee**[1,3], **Xinchi Qiu**[1]
Yan Gao[1,4] **Hongxiang Fan**[1,3], **Nicholas D. Lane** [1,4]

[1] University of Cambridge, [2] University College London,
[3] Samsung AI Center, Cambridge, [4] Flower Labs

ABSTRACT

Pre-trained large language models (LLMs) have emerged as a cornerstone in modern natural language processing, with their utility expanding to various applications and languages. However, the fine-tuning of multilingual LLMs, particularly for low-resource languages, is fraught with challenges steming from data-sharing restrictions (the physical border) and from the inherent linguistic differences (the linguistic border). These barriers hinder users of various languages, especially those in low-resource regions, from fully benefiting from the advantages of LLMs.

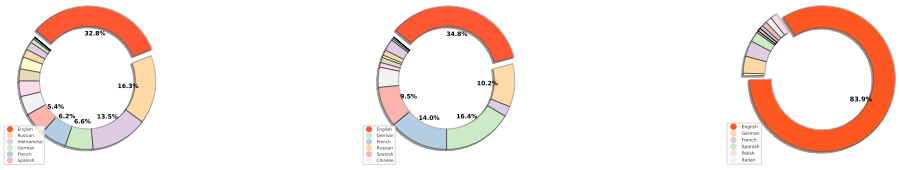
To address these challenges, we propose the Federated Prompt Tuning Paradigm for multilingual scenarios, which utilizes parameter-efficient fine-tuning while adhering to data sharing restrictions. We have designed a comprehensive set of experiments and analyzed them using a novel notion of language distance to underscore the strengths of this paradigm: Even under computational constraints, our method not only bolsters data efficiency but also facilitates mutual enhancements across languages, particularly benefiting low-resource ones. Compared to traditional local cross-lingual transfer tuning methods, our approach achieves 6.9% higher accuracy, reduces the training parameters by over 99%, and demonstrates better stability and generalization. Such findings underscore the potential of our approach to promote social equality and champion linguistic diversity, so that no language will be left behind. Our code is released at https://github.com/Ryan0v0/multilingual_borders.

1 INTRODUCTION

Large language models (LLMs) have been driving the recent progress in natural language processing (Brown et al., 2020; Chowdhery et al., 2022; Anil et al., 2023; Touvron et al., 2023a;b). These large models, built on extensive corpora, offer valuable insights and impressive results across a range of applications. In the meantime, in order to provide universally accessible knowledge with LLMs, extending the LLMs to multiple languages has become a particularly relevant research target (Conneau & Lample, 2019; Conneau et al., 2020; Artetxe et al., 2020)

Fine-tuning and deploying multilingual LLMs for practical downstream tasks present a unique set of challenges, distinct from those encountered with monolingual models. A primary concern is the geographical distribution of data across different languages, often stored in separate physical locations, making the sharing of data across regions difficult or, in some cases, prohibited due to legal constraints. For example, regulations such as the General Data Protection Regulation (GDPR) impose significant limitations on cross-region data sharing (Lim et al., 2020). Additionally, the linguistic diversity across regions, such as the differences between Sino-Tibetan and Indo-European languages, introduces a Non-Independent and Identically Distributed (non-IID) challenge in learning a unified multilingual model. This situation accentuates privacy concerns and highlights the need for effective privacy-preserving techniques when using multilingual LLMs. To this end, some recent works attempt to address privacy-constrained fine-tuning for multilingual tasks and explore how

*Corresponding to: Wanru Zhao (wz341@cam.ac.uk)



(a) XLM-R (Conneau et al., 2020) (b) mBERT (Pires et al., 2019a) (c) PaLM (Chowdhery et al., 2022)

Figure 1: Linguistic coverage of different large language models.

different languages impact the federated process (Weller et al., 2022). However, these works primarily target high-resources languages, leaving the low-resource languages under-explored.

Addressing low-resource languages is essential to promoting technological fairness and protecting linguistic diversity. Unlike their high-resource counterparts, low-resource languages pose intriguing research challenges: i) **Limited computational resources.** Regions of low-resource languages are often economically developing areas, with little access to enough computational resources required to either train language models from scratch or fully fine-tune pre-trained large language models (Mager et al., 2021; Adebara & Abdul-Mageed, 2022). ii) **Limited data in the target language.** Due to a small speaking population or the spoken nature of the language, data is often scarce (Adelani et al., 2021; Muhammad et al., 2022; Ebrahimi et al., 2022). As depicted in Figure 1, the pre-training data for LLMs is predominantly in English, with little coverage of low-resource languages. Under such circumstances, the performance of low-resource languages is often unsatisfactory during fine-tuning because of under-representation. iii) **Memorization risk.** Recent studies found that as pre-trained models scale up, their ability to memorize training data increases (Tirumala et al., 2022), which implies that, when fine-tuning these models with limited data, the risk of overfitting and potential privacy issues arises.

Recognizing the challenges posed by low-resource languages, Federated Learning (FL) has emerged as a promising training paradigm that addresses many of these concerns. In FL, model training occurs across multiple decentralized devices or servers, with the critical distinction that data remains localized McMahan et al. (2017); Beutel et al. (2020); Mathur et al. (2021). This approach is particularly well-suited to multilingual settings, where leveraging data from diverse linguistic backgrounds is essential without compromising data privacy. The geographical distribution and inherent linguistic diversity of devices in FL mean that the data on each node are likely to exhibit a non-IID (Non-Independent and Identically Distributed) distribution. This inherent characteristic of FL not only aligns with the privacy-preserving needs but also naturally accommodates the varied and complex linguistic features of low-resource languages, making it an ideal methodology for enhancing the inclusivity and effectiveness of language technologies. Furthermore, the efficiency of computational and communication processes Qiu et al. (2023; 2022) is paramount in such scenarios, underscoring the need for FL approaches that are not only privacy-centric but also resource-efficient, ensuring broad applicability and sustainability.

In this paper, in order to break the geographic border and the linguistic border between different language speaking countries, we propose a new paradigm grounded in FL, Multilingual Federated Prompt Tuning, focusing on parameter-efficient fine-tuning for multilingual tasks across various regions or devices. Specifically, by having each country fine-tune the model locally and then pass the updated parameters to a server for aggregation, we obtain a global model, which leverages collective knowledge and exposing models to a wider range of linguistic patterns. Considering the different linguistic patterns in various countries, our prompt encoders help generalize and adapt to the languages of different countries. We demonstrate the effectiveness of our paradigm on standard NLP tasks. The performance of our paradigm achieves 6.9% accuracy improvement while navigating privacy regulations that restrict cross-country data sharing. Compared with local monolingual fine-tuning, our paradigm reduces computational and communication cost by more than 99%. Our approach paves the way for fine-tuning multilingual large language models on resource-constraint devices across various regions, and holds the potential to promote social equality, privacy, and linguistic diversity in the research community. Our contributions are as follows:

- We demonstrate that federated prompt tuning can serve as a new paradigm for addressing the linguistic and physical challenges of multilingual fine-tuning across regions or devices.
- Compared to traditional local monolingual fine-tuning paradigm, federated prompt tuning is not only data and parameter efficient, suitable for situations with limited computational power, but also shows better generalization and stability, performing well in downstream tasks of low-resource languages with huge *language distance* from the *pre-trained language*.
- Federated prompt tuning also helps to alleviating data privacy leakage by reducing both the data transmission amount and data memorization, which opens up new avenues for exploration and has the potential to inform future research in this area.

2 Related Work

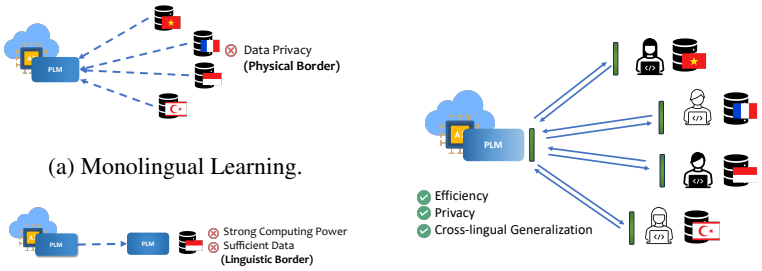
Multilingual Language Models. Multilingual Pre-trained Language Models (PLMs) such as mBERT (Pires et al., 2019a), XLM-R (Conneau et al., 2020), and SeamlessM4T (Loic Barrault, 2023) have emerged as a viable option for bringing the power of pre-training to a large number of languages (Doddapaneni et al., 2021). Many studies analyzed mBERT's and XLM-R's capabilities and limitations, finding that the multilingual models work surprisingly well for cross-lingual tasks, despite the fact that they do not rely on direct cross-lingual supervision (e.g., parallel or comparable data, and translation dictionaries (Pires et al., 2019b; Wu & Dredze, 2019; Artetxe et al., 2020)

However, these multilingual PLMs are not without limitations. Particularly, Conneau et al. (2020) observed the *curse of multilinguality* phenomenon: given a fixed model capacity, adding more languages does not necessarily improve the multilingual performance but can deteriorate the performance after a certain point, especially for underrepresented languages (Wu & Dredze, 2020; Hu et al., 2020; Lauscher et al., 2020) Prior work tried to address this issue by increasing the model capacity (Artetxe et al., 2020; Pfeiffer et al., 2020; Chau et al., 2020) or through additional training for particular language pairs (Pfeiffer et al., 2020; Ponti et al., 2020) or by clustering and merging the vocabularies of similar languages, before defining a joint vocabulary across all languages (Chung et al., 2020). Despite these efforts, the multilingual PLMs still struggle with balancing their capacity across many languages in an sample-efficient and parameter-efficient way (Ansell et al., 2022; Marchisio et al., 2022; Chen et al., 2023).

Prompt Learning and Parameter-Efficient fine-tuning. The size of pre-trained language models has been increasing significantly (Brown et al., 2020), presenting challenges to traditional task transfer based on full-parameter fine-tuning. Recent research has shifted its attention to Parameter-Efficient fine-tuning techniques, such as prompt tuning (Lester et al., 2021; Li & Liang, 2021; Liu et al., 2021b), adapters (Houlsby et al., 2019a), as well as combined approaches including LoRA (Hu et al., 2021) and BitFit (Ben Zaken et al., 2022). These methods utilize a minimal number of tuning parameters, yet they offer transfer performance that is comparable with traditional fine-tuning.

Prompt learning involves training a small set of prompt tokens while keeping the original pre-trained model parameters frozen, thus allowing for model personalization with minimal parameter updates. (Liu et al., 2021a). This paradigm shows promise in effectively leveraging large pre-trained models in a data-efficient manner by reducing the need for extensive labeled datasets (Schick & Schütze, 2022). Additionally, prompt learning has exhibited a remarkable ability to generalize across a variety of tasks, suggesting a step towards more flexible and adaptable machine learning systems (Shin et al., 2020). Another most widely used Parameter-Efficient fine-tuning technique is LoRA, or Low-Rank Adaptation, which involves freezing the pre-trained model weights and injecting trainable rank decomposition matrices into each layer of the Transformer architecture, thereby achieving fine-tuning without incurring any additional inference latency (Hu et al., 2021).

Federated Learning. Federated Learning has garnered significant attention in the academic realm. It bypasses the conventional model training process by sharing models instead of raw data. With Federated Averaging (FedAvg) (McMahan et al., 2017), participating clients train models using their respective private datasets locally, and the updated model parameters are aggregated. This preserves the privacy of the underlying data while collectively benefiting from the knowledge gained during the training process (Konečnỳ et al., 2016). Despite abundant research made on problems at hospitals, legal firms, and financial institutions, extending language models for multilingual usages effectively and efficiently, especially for low-resource languages remains under-explored.



(a) Monolingual Learning.

(b) Centralized Learning.

(c) Federated Prompt Learning.

Figure 2: Comparison of three different learning paradigms for multilingual tasks.

In the general NLP domain, FL has been instrumental in tasks such as language modeling, sentiment analysis, and machine translation, showcasing its potential to revolutionize the way models are trained and deployed (Banabilah et al., 2022). Lin et al. (2022) introduces a benchmarking framework for evaluating various FL methods across NLP tasks, providing a universal interface between Transformer-based models and FL methods. Wang et al. (2022)is a federated approach designed for multilingual Natural Language Understanding (NLU) that integrates knowledge from multiple data sources through federated learning techniques to enhance the efficacy and accuracy of multilingual text processing. However, considerations regarding computational and communication efficiency in resource-constrained environments have not been adequately addressed.

3 A New Paradigm for Multilinguality: Federated Prompt Tuning

In our federated learning setup, we have K clients. Each client k has a private dataset, either monolingual or multilingual, defined as: $\mathcal{D}_k = \{(x_{k,i}, y_{k,i}) \mid i = 1, \ldots, n_k\}$, where $x_{k,i}$ denotes the textual content, and $y_{k,i}$ is its corresponding label. The server sets up and maintains a global prompt encoder. We denote the parameters of the global prompt encoder as h_g, and each client k has its own prompt encoder, denoted its parameters by h_k, tuned based on its dataset.

3.1 Virtual Prompt Encoder

Instead of selecting discrete text prompts in a manual or automated fashion, in our Multilingual Federated Prompt Tuning paradigm, we utilize virtual prompt embeddings that can be optimized via gradient descent. Specifically, each prompt encoder, whether global or local, takes a series of virtual tokens, which are updated during tuning to better aid the model.

Figure 3 shows how our prompt tuning works on both clients and server. Specifically, on each client k, a textual prompt tailored for a specific task and input text x are passed to the model. Then task specific virtual tokens are retrieved based on the textual prompt.

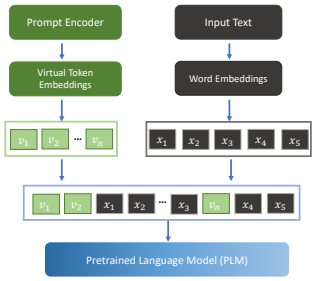


Figure 3: The pipeline of prompt tuning.

The primary objective of each prompt encoder is to generate an effective prompt embedding $v_0, v_1, v_2, \ldots$ for each client based on task specific virtual tokens, to guide the PLM in producing the desired outputs. With the input text tokenized, the discrete word token embeddings $x_1, x_2, x_3, \ldots$ are retrieved. Then virtual token embeddings are inserted among discrete token embeddings and passed together into the PLM. During the fine-tuning phase, based on a task-specific loss, the parameters of the prompt encoder, h_k, are often tuned: $\mathcal{L}(x, y; h_k) = \mathit{Loss}(D(v \oplus x), y)$, where D can be a decoder that maps the internal representation to task outputs, and *Loss* is an appropriate loss function,

such as Cross-Entropy Loss. During the fine-tuning, the PLM's parameters keep frozen, whereas the prompt encoder's parameters h_k are updated in accordance with the loss.

3.2 FEDERATED PROMPT AVERAGING

In every communication round t, Federated Prompt Averaging includes the following steps.

Initialization: The server initializes the global prompt encoder h_g^t. Each client initializes its local prompt encoder $h_0^t, h_1^t, \dots h_k^t$.

Client Selection: We select a fraction C of the total K clients for training. This subset size is $m = \max(C \times K, 1)$. The subset we choose is denoted as S.

Local Encoder Tuning: Each client k in S fetches the current global prompt encoder h_g^t, and assembles it with the PLM. During the local training on the local data $\mathcal{D}_k$, The PLM's parameters stay fixed while local prompt encoder parameters h_k^t are tuned.

Aggregation: The server aggregates updates from all clients using weighted average. The global prompt encoder h_g^{t+1} is updated based on the received parameters h_k^t from clients for the next round of federated prompt tuning: $h_g^{t+1} = \sum_{k=1}^{K} \frac{|\mathcal{D}_k|}{\sum_{k=1}^{K} |\mathcal{D}_k|} h_k^t$.

4 EXPERIMENTAL SETUP

4.1 TASKS AND DATASETS

We evaluate our model using the popular XGLUE benchmark (Liang et al., 2020), a cross-lingual evaluation benchmark for our multilingual evaluation. We conduct our experiments on classification tasks of News Classification (NC), XNLI (Conneau et al., 2018) and MasakhaNEWS (Adelani et al., 2023). Accuracy (ACC) of the multi-class classification is used as the metric for both of the tasks. The details regarding each dataset can be found in Appendix D. Our base model for both tasks is the XLM-RoBERTa base-sized model (270M parameters), shown to perform well across many languages (Conneau et al., 2020).

4.2 MULTILINGUAL FINE-TUNING PARADIGMS

1) **Local Monolingual fine-tuning:** Traditional fine-tuning where a separate model is fine-tuned using the corresponding dataset for each single language locally. 2) **Centralized fine-tuning:** Standard fine-tuning using a combined dataset of all languages centralized in the cloud. 3) **Federated Full fine-tuning:** Directly fine-tuning the whole pre-trained language model in a federated manner, with the full pre-trained model on the server or each client. 4) **Federated Prompt Tuning:** Only training the prompt encoder in a federated manner, with the prompt encoder on the Server or each client. 5) **Federated LoRA (Low-Rank Adaptation):** Only training over-parameterized models with a low intrinsic rank in a federated manner, with the trainable rank decomposition matrices on the server or each client. In our tables, we use *PE* to denote the methods with parameter-efficient techniques.

5 EVALUATION AND ANALYSIS

5.1 MAIN RESULTS

Table 3 presents the experimental results on news classification. When employing parameter-efficient fine-tuning in comparison to full parameter fine-tuning, there is an acceptable decline in accuracy. Despite this decrease, the overall performance remains consistent and stable. A significant gain in accuracy is observed when adopting the FL approach. It is worth noting that the fine-tuning time is considerably reduced when employing the parameter-efficient method as opposed to without it. For a comprehensive analysis of this, refer to the section 5.4.

Table 2 summarises the results of our FL experiments on the XNLI task. To demonstrate the advantage of our Federated Prompt Tuning approach, a comparison was made with traditional monolingual training. As the result shows, our Federated Prompt Tuning, particularly on Non-IID

Table 1: Results for FL experiments on the NC task. Bold scores indicate the best in the column.

Method	en	es	fr	de	ru	Avg
Monolingual	92.4	84.7	79.5	88.3	89.0	86.8
Centralized	93.9	86.7	**82.9**	**89.5**	88.6	88.3
FL (IID)	**94.1**	**86.9**	82.7	89.4	**88.8**	**88.4**
FL (Non-IID)	92.4	86.3	81.2	88.9	84.7	86.7
PE_Monolingual	82.9	59.7	47.3	71.4	60.0	64.3
PE_Centralized	89.1	76.2	67.4	78.8	75.9	77.5
PE_FL (IID) **(Ours)**	**91.2**	**82.2**	**76.5**	**86.4**	**81.6**	**83.6**
PE_FL (Prompt Tuning) (Non-IID) **(Ours)**	87.8	79.2	73.7	83.1	79.5	80.7
PE_FL (LoRA) (Non-IID) **(Ours)**	89.3	76.0	75.4	75.8	83.2	79.9

Table 2: Results for FL experiments on the XNLI task. Bold scores indicate the best in the column. The PE_FL is evaluated under the Non-IID setting.

Method	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
PE_Monolingual	39.1	35.1	36.6	35.7	35.3	35.9	35.5	26.2	32.1	31.7	31.5	33.7	31.6	26.0	28.1	32.94
PE_Centralized	35.3	36.9	33.3	35.3	30.5	36.5	33.7	35.7	33.3	**40.1**	36.1	30.5	37.3	**38.6**	29.3	34.86
PE_FL **(Ours)**	**43.2**	**40.6**	**42.9**	**40.2**	**39.7**	**40.8**	**41.1**	**37.6**	**39.1**	39.9	**39.4**	**39.8**	**38.2**	37.1	**37.8**	**39.83**

Table 3: Results for FL experiments on MasakhaNEWS. Bold scores indicate the best in the column.

Method	eng	fra	hau	swa	yor	Avg
PE_Monolingual	79.08	84.91	75.18	76.64	52.8	73.7
PE_Centralized	79.81	87.10	**80.78**	84.18	64.48	79.3
PE_FL (Prompt Tuning) **(Ours)**	82.99	**89.81**	65.96	**86.16**	57.20	76.4
PE_FL (LoRA) **(Ours)**	**87.10**	85.64	76.64	83.21	**72.50**	**81.0**

setting, consistently outperformed the monolingual method across all languages. Remarkably, this superior performance was maintained even for languages with limited available data. The average accuracy further shows the advantage of Federated Prompt Tuning, marking a noticeable improvement from 32.94% in the monolingual approach to 39.83% with Non-IID Federated Prompt Tuning.

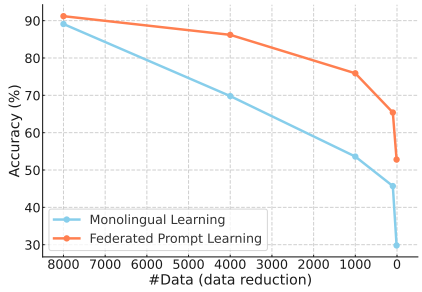


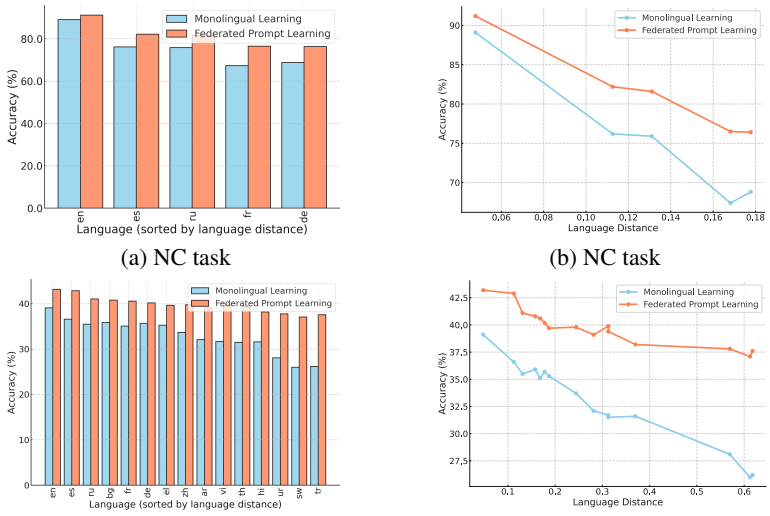
Figure 4: Performance comparison between traditional local fine-tuning and our federated prompt tuning method.

From our results in section 5.1, we observe that some languages demonstrate superior accuracy with the FL method compared to the centralized approach. This enhanced performance might be attributed to the Federated Prompt Averaging in FL, which could introduce similar implicit regularization effects (Izmailov et al., 2018; Rehman et al., 2022). Additionally, the prompt encoder, by freezing the core language model parameters, prevents the model from altering its foundational understanding of language. As a result, the model's tendency to overfiting is reduced, minimizing the risk of memorizing specific lexical cues and spurious correlations.

5.2 Ablation Study I: Data Efficiency

As previous sections mentioned, one characteristic of low-resource languages is their limited available data. Hence, enhancing data sample efficiency is crucial when fine-tuning pre-trained models for downstream tasks. To better validate and simulate the advantages of our approach in real-world scenarios, we reduced the data volume for one language and observed the performance under traditional local fine-tuning as well as our Federated prompt fine-tuning method. We conducted experiments on German News Classification. German was chosen because it represents the language with the fewest resources among the five languages included in this task.

As shown in the Figure 4, our Federated Prompt Tuning method consistently outperforms the traditional monolingual approach. As we reduce the dataset size from 8,000 to 30, the accuracy of



(a) NC task (b) NC task

(c) XNLI task (d) XNLI task

Figure 5: Comparative performance for both XNLI and NC tasks. (a)(c) reports the fine-tuning accuracy across different languages; (b)(d) reports the fine-tuning accuracy across languages with varying similarity to the pre-trained languages.

the traditional method drops significantly. On the other hand, the Federated Prompt Tuning method retains its performance, demonstrating its robustness even with limited data. This clearly indicates that our Federated Prompt Tuning approach is better suited for scenarios with limited data availability.

5.3 Ablation Study II: Language Distance

As previously mentioned, another characteristic of low-resource languages is that their linguistic features differ from those of high-resource languages, particularly in aspects including syntax, phonology, and inventory. Consequently, direct fine-tuning on models pre-trained with highly dissimilar languages often yields unsatisfying results. Therefore, we conducted an ablation study to examine the impact of language similarity on performance, comparing our Federated Prompt fine-tuning method to the traditional local fine-tuning approach.

5.3.1 Multilingual Distance Measurement

We define the *pre-trained language* as a representative composite language formed by blending each language in the multilingual corpus used for pre-training, in proportion to their amount. This is a formal representation for the mixed dataset composition. We define distance for a specific language in the downstream tasks, in terms of the negative logarithm of its similarity to the pre-trained language.

We leverage the database from Littell et al. (2017); Malaviya et al. (2017) to extract feature vectors for each language. These vectors are then weighted according to the token count of each language in the pre-trained corpus to calculate the feature vector of the pre-trained language. Given the feature vector V_i for the i-th language, token count T_i, and total tokens T_{total}, the weight w_i is given by $w_i = \frac{T_i}{T_{\text{total}}}$ and the feature vector V_p for the pre-trained model is computed as $V_p = \sum_{i=1}^{n} w_i \cdot V_i$.

We define distance for a specific language in the downstream tasks, in terms of the negative logarithm of its cosine similarity to the pre-trained language. Let v represent the feature vector of a specific language in the downstream task. The diversity measure ϕ between this language and the average language of the pre-trained model is defined as $\phi(v_i) = -\log(\cos(v_i, V_{\text{p}}))$.

5.3.2 Fine-tuning Languages Distance from Pre-trained Language

Leveraging the distance metric, we compared model performance of languages with varying degrees of distance to the pre-trained language. We present our results from two key experiments on the NC and XNLI tasks. From Figure 5, a conspicuous trend is observed: As the language similarity to

	# Trainable Params	Communication Cost
Full fine-tuning	278,655,764	110,592 MB
Prompt Tuning (Ours)	1,202,708	479 MB
LoRA (Ours)	1,491,476	594 MB

Table 4: Comparison of parameter efficiency and communication overhead in NC task.

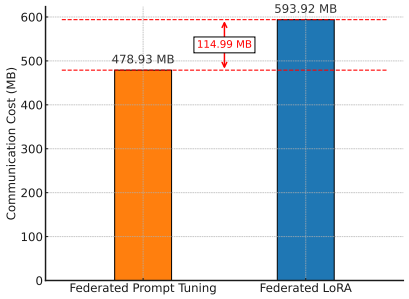


Figure 6: Communication Cost Comparison between Federated Prompt Tuning and Federated LoRA.

the pre-trained language decreases, the model's accuracy tends to drop. However, when we apply our Federated Prompt method, this decline is notably less steep. This means that even when we are dealing with languages that are quite different from the pre-trained one, our method manages to retain a decent level of accuracy. The difference between our method and the traditional local fine-tuning becomes even more obvious for languages with less data, indicating that our Federated Prompt Tuning method offers significant advantages, particularly in low-resource scenarios.

5.4 Ablation Study III: Parameter Efficiency

We evaluate the efficiency both in terms of computation and communication. From the perspective of trainable parameters, our method demonstrates exceptional parameter efficiency. In both tasks, despite the total number of parameters exceeding 278 million, the trainable parameters are only around 1.2 million, accounting forless than 0.5% of the total. Such design can substantially reduce training time and computational resources, while mitigating the risk of overfitting. Also, high parameter efficiency offers the potential for model deployment in resource-constrained environments.

Regarding communication, XLM-Roberta-Base's data transmission in FL with 5 clients and 10 communication rounds was 108 GB. After our optimization, using a prompt encoder with a 2×768 structure, the transmission size reduced to 478.93 MB, a 99% reduction shown in Table 4. This optimization enhances efficiency in federated prompt tuning and expands its applicability to bandwidth-constrained environments including edge devices and mobile networks.

6 Conclusion

Addressing the complexities of multilingual LLMs, especially for low-resource languages, requires innovative approaches that can balance efficiency, privacy concerns, and performance. Our Multilingual Federated Prompt Tuning paradigm provides a solution to these challenges. By aggregating lightweight multilingual prompts, this approach offers enhanced fine-tuning capabilities with minimal computational demand. The robustness of our method is especially pronounced for low-resource languages with sparse data and rare linguistic features. Overall, this approach promises to advance privacy and linguistic diversity in the realm of NLP. Future work will focus on exploring the impact on the Multilingual Federated Tuning method based on prompt learning as the model scale increases.

Limitation In our current paradigm, we have not added extra privacy protection techniques to defend against potential privacy attacks, such as gradient inversion (Geiping et al., 2020; Huang et al., 2021). We need to introduce differential privacy (Wei et al., 2020), secure aggregation (Bonawitz et al., 2016), and/or other methods to protect the training data. While FL by itself offers some privacy benefits, we understand it still needs more methods and attention to be really privacy-preserving, and we're interested in these limitations being addressed in follow-up work.

ETHICS STATEMENTS

This paper presents work whose goal is to advance the field of decentralized language model training and multilingual NLP. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

ACKNOWLEDGEMENT

This work was supported by the European Research Council via the REDIAL project (805194). We would also like to thank David Ifeoluwa Adelani, Shaozuo Yu and the anonymous reviewers for the insightful discussions and useful suggestions for the camera-ready version.

REFERENCES

Ife Adebara and Muhammad Abdul-Mageed. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3814–3841, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.265.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-Azzawi, Blessing K. Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Oluwaseyi Ajayi, Tatiana Moteu Ngoli, Brian Odhiambo, Abraham Toluwase Owodunni, Nnaemeka C. Obiefuna, Shamsuddeen Hassan Muhammad, Saheed Salahudeen Abdullahi, Mesay Gemeda Yigezu, Tajuddeen Rabiu Gwadabe, Idris Abdulmumin, Mahlet Taye Bame, Oluwabusayo Olufunke Awoyomi, Iyanuoluwa Shode, Tolulope Anu Adelani, Habiba Abdulganiy Kailani, Abdul-Hakeem Omotayo, Adetola Adeeko, Afolabi Abeeb, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Raphael Ogbu, Chinedu E. Mbonu, Chiamaka I. Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola F. Awosan, Tadesse Kebede Guge, Sakayo Toadoum Sari, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Ussen Abre Kimanuka, Kanda Patrick Tshinu, Thina Diko, Siyanda Nxakama, Abdulmejid Tuni Johar, Sinodos Gebre, Muhidin Mohamed, S. A. Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, and Pontus Stenetorp. Masakhanews: News topic classification for african languages. 2023.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1778–1796, 2022.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421.

Syreen Banabilah, Moayad Aloqaily, Eitaa Alsayed, Nida Malik, and Yaser Jararweh. Federated learning review: Fundamentals, enabling technologies, and future applications. *Information Processing & Management*, 59(6):103061, 2022. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2022.103061.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of*

the Association for Computational Linguistics (Volume 2: Short Papers), pp. 1–9, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.1.

Daniel J. Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, and Nicholas D. Lane. Flower: A friendly federated learning research framework, 2020.

K. A. Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

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

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1324–1334, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.118.

Yihong Chen, Kelly Marchisio, Roberta Raileanu, David Ifeoluwa Adelani, Pontus Stenetor, Sebastian Riedel, and Mikel Artetx. Improving language plasticity via pretraining with active forgetting. *NeurIPS 2023*, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

Hyung Won Chung, Dan Garrette, Kiat Chuan Tan, and Jason Riesa. Improving multilingual models with language-clustered vocabularies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4536–4546, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.367.

Alexis Conneau and Guillaume Lample. *Cross-Lingual Language Model Pretraining*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *EMNLP*, 2018.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747.

Sumanth Doddapaneni, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. A primer on pretrained multilingual language models. *CoRR*, abs/2107.00676, 2021.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained

multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6279–6299, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.435.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in neural information processing systems*, 33:16937–16947, 2020.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 09–15 Jun 2019a.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019b.

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

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080, 2020.

Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in Neural Information Processing Systems*, 34:7232–7241, 2021.

Pavel Izmailov, Dmitrii Podoprikhin, T. Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence*, 2018.

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

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 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)*, pp. 4483–4499, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.363.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *CoRR*, abs/2104.08691, 2021.

Xiang Lisa Li and Percy Liang. 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)*, pp. 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353.

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. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv*, abs/2004.01401, 2020.

Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22(3):2031–2063, 2020.

Bill Yuchen Lin, Chaoyang He, Zihang Ze, Hulin Wang, Yufen Hua, Christophe Dupuy, Rahul Gupta, Mahdi Soltanolkotabi, Xiang Ren, and Salman Avestimehr. FedNLP: Benchmarking federated learning methods for natural language processing tasks. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 157–175, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.13.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 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*, volume 2, pp. 8–14, 2017.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586, 2021a.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *arXiv:2103.10385*, 2021b.

Mariano Cora Meglioli Loic Barrault, Yu-An Chung. Seamlessm4t—massively multilingual & multimodal machine translation. *ArXiv*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios Gonzales, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, et al. Findings of the americasnlp 2021 shared task on open machine translation for indigenous languages of the americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pp. 202–217, 2021.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. Learning language representations for typology prediction. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark, September 2017.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`, 2022.

Kelly Marchisio, Patrick Lewis, Yihong Chen, and Mikel Artetxe. Mini-model adaptation: Efficiently extending pretrained models to new languages via aligned shallow training. *ACL 2023, Findings of the Association for Computational Linguistics*, 2022.

Akhil Mathur, Daniel J Beutel, Pedro Porto Buarque de Gusmao, Javier Fernandez-Marques, Taner Topal, Xinchi Qiu, Titouan Parcollet, Yan Gao, and Nicholas D Lane. On-device federated learning with flower. *arXiv preprint arXiv:2104.03042*, 2021.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 2017.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, et al. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. *arXiv preprint arXiv:2201.08277*, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7654–7673, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, Florence, Italy, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2362–2376, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.185.

Xinchi Qiu, Javier Fernandez-Marques, Pedro PB Gusmao, Yan Gao, Titouan Parcollet, and Nicholas Donald Lane. Zerofl: Efficient on-device training for federated learning with local sparsity. *arXiv preprint arXiv:2208.02507*, 2022.

Xinchi Qiu, Titouan Parcollet, Javier Fernandez-Marques, Pedro PB Gusmao, Yan Gao, Daniel J Beutel, Taner Topal, Akhil Mathur, and Nicholas D Lane. A first look into the carbon footprint of federated learning. *Journal of Machine Learning Research*, 24(129):1–23, 2023.

Yasar Abbas Ur Rehman, Yan Gao, Jiajun Shen, Pedro Porto Buarque de Gusmão, and Nicholas Lane. Federated self-supervised learning for video understanding. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 506–522, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19821-2.

Timo Schick and Hinrich Schütze. True few-shot learning with Prompts—A real-world perspective. *Transactions of the Association for Computational Linguistics*, 10:716–731, 2022. doi: 10.1162/tacl_a_00485.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Haoyu Wang, Handong Zhao, Yaqing Wang, Tong Yu, Jiuxiang Gu, and Jing Gao. Fedkc: Federated knowledge composition for multilingual natural language understanding. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pp. 1839–1850, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3511988.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE transactions on information forensics and security*, 15:3454–3469, 2020.

Orion Weller, Marc Marone, Vladimir Braverman, Dawn J Lawrie, and Benjamin Van Durme. Pretrained models for multilingual federated learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2022.

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. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6.

Shijie Wu and Mark Dredze. 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)*, pp. 833–844, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1077.

Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pp. 120–130, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.repl4nlp-1.16.

Wanru Zhao, Xinchi Qiu, Javier Fernandez-Marques, Pedro PB de Gusmão, and Nicholas D Lane. Protea: Client profiling within federated systems using flower. In *Proceedings of the 1st ACM Workshop on Data Privacy and Federated Learning Technologies for Mobile Edge Network*, pp. 1–6, 2022.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The United Nations parallel corpus v1.0. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3530–3534, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

APPENDIX

A CHALLENGES AND OPPORTUNITIES OF MULTILINGUAL NLP

As natural language processing technologies advance, not all languages have been treated equally by developers and researchers. There are around 7,000 languages spoken in the world, and approximately 400 languages have more than 1 million speakers. However, there is scarce coverage of multilingual datasets. This is especially true for low-resource languages, where data scarcity is a major bottleneck. Furthermore, the under-indexing of certain languages is also driven by access to compute resources. Mobile data, compute, and other computational resources may often be expensive or unavailable in regions that are home to under-represented languages. Unless we address this disproportionate representation head-on, we risk perpetuating this divide and further widening the gap in language access to new technologies. One pressing example is biomedical data. Due to its global scale, this digital content is accessible in a variety of languages, yet most existing NLP tools remain English-centric. This situation highlights the need for effective strategies: how can we exploit abundant labeled data from resource-rich languages to make predictions in resource-lean languages?

Also, the problem is very timely compared to other application scenarios. It was not even considered a year ago. Previously, due to the smaller size of language models, the demand for data was not as high, and different kinds and sources of data were treated equally. Currently, the progress of LLMs, their usability, the amount of attention they receive, and the increased regulation on data, compound and lead to the urgency of this problem, where we are among the first batch to attempt to break both lingual and physical barriers.

B PROMPT CONSTRUCTION AND INITIALIZATION

When we use Prompt Tuning to optimize the parameter efficiency, the prompt tuning initialization text is *Predict the category given the following news article* for all the News Classification tasks. By providing the string of words, we initialize virtual token embeddings from existing embedding weights. This string is tokenized and tiled or truncated to match the number of virtual tokens.

C HYPERPARAMETERS AND IMPLEMENTATION

For all of the experiments, we report results using the 1e-3 learning rate and early stopping (5 epochs of no improvement). For a fair comparison with the setup in Houlsby et al. (2019b), we restrict the model sequence length to 512 and use a fixed batch size for all tasks. For FL experiments, we adjust the parameter α that controls the mixture of languages in the dataset. An α value of 1.0 signifies a uniform mixture of all languages, while values closer to 0 indicate a dominant representation of individual languages or a more separated mixture. When we use Prompt Tuning to optimize the parameter efficiency, the number of virtual tokens is 1, and the prompt tuning init text is *Predict the category given the following news article* for all the News Classification tasks. When we use LoRA to optimize the parameter efficiency, We use two different ranks (1 and 8), LoRA α is 16 and LoRA dropout is 0.1.

We use Hugging Face's transformers library (Wolf et al., 2020) and PEFT library (Mangrulkar et al., 2022) for loading pre-trained models and prompt tuning configurations. For our federated training and evaluation, we use the Flower framework (Beutel et al., 2020; Zhao et al., 2022) and PyTorch as the underlying auto-differentiation framework (Paszke et al., 2019). We use the AdamW optimizer (Loshchilov & Hutter, 2019; Kingma & Ba, 2015) for all experiments. All experiments are conducted using NVIDIA A40.

D DATASETS FOR GENERATIVE TASKS

UN Corpus (Ziemski et al., 2016) is a Machine Translation dataset of official records from the UN proceedings over the years 1990 to 2014, covering six languages: English, French, Spanish, Russian, Chinese, and Arabic. we sample 10k in each direction for training and 5k each for evaluation sets.

We cover three machine translation directions: En → Fr, Ar → Es, Ru → Zh, and sample 10k in each direction for training and 5k each for evaluation sets.

News Classification (NC) is a classification problem with 10 classes across 5 languages: English, Spanish, French, German, and Russian. This task aims to predict the category given a news article. Since only 10k annotated examples are available for each language (excluding the official test set), we sample 8k instances for training and 1k for evaluation sets.

Cross-lingual Natural Language Inference (XNLI) is a cross-lingual sentence understanding problem which covers 15 languages, including high-resource languages (English, French, Spanish, German, Russian and Chinese), medium-resource languages (Arabic, Turkish, Vietnamese and Bulgarian), and low-resource languages (Greek, Thai, Hindi, Swahili and Urdu). The task involves determining the relationship between a premise and a hypothesis sentence, and this relationship can be categorized into one of three classes: entailment, contradiction, or neutral. We sample 2k instances for training and 250 for evaluation sets for each language. NLI serves as an effective benchmark for assessing cross-lingual sentence representations, and better approaches for XNLI will lead to better general Cross-Lingual Understanding (XLU) techniques.

MasakhaNEWS is a benchmark dataset for news topic classification covering 16 languages widely spoken in Africa, where African languages are severely under-represented in NLP research due to lack of datasets covering several NLP tasks. The task involves categorizing news articles into different categories like sports, business, entertainment, and politics.We choose English, Hausa, Kiswahili, French and Yorùbá in our experiments. We sample 1433 instances for training and 411 for evaluation sets for each language.

E Federated Prompt Averaging Algorithm

Algorithm 1 Federated Prompt Averaging

Server executes:
Initialize h_{g}
for each round t **do**
  Select subset S of m clients
  for each client k in S **do**
    Send h_{g} to client k
  end for
  Aggregate client updates:
  $h_g^{t+1} = \sum_{k=1}^{K} \frac{|\mathcal{D}_k|}{\sum_{k=1}^{K}|\mathcal{D}_k|} h_k^t$
end for

Client k **executes:**
Retrieve current h_{g}
Assemble full model using h_k and PLM parameters
Train model on $\mathcal{D}_k$
Update local prompt encoder h_k
Send updated h_k to server