Breaking Physical and Linguistic Borders: Privacy-Preserving Multilingual Prompt Tuning for Low-Resource Languages

Anonymous Author(s)
Affiliation
Address
email

Abstract

Pretrained 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 privacy 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 stronger cross-lingual generalization. Such findings underscore the potential of our approach to promote social equality, ensure user privacy, and champion linguistic diversity.

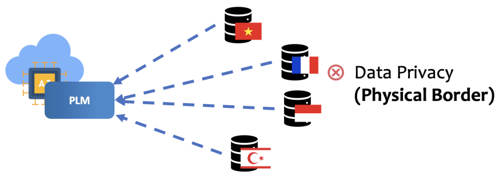
1 Introduction

Pretrained large language models (LLMs) have been driving the recent progress in natural language processing [11, 14, 3, 56, 57]. These large models, built on extensive corpora, offer valuable insights and impressive results across a range of applications. At the meantime, in order to provide universally accessible knowledge with LLMs, extending them to multiple languages has become a particularly relevant research target [17, 16, 5, 45].

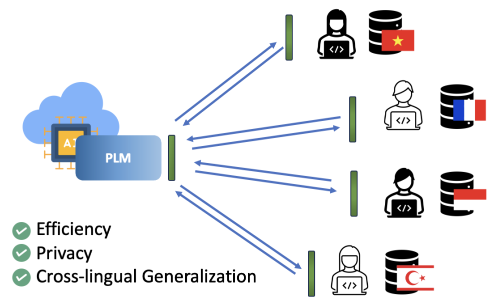
However, finetuning and deploying multilingual LLMs in practical downstream tasks are not as easy as its monolingual counterpart. First of all, sharing data across different regions can be difficult or even impossible. Regulations like the General Data Protection Regulation (GDPR) [32] limit cross-region data-sharing. Moreover, languages in various regions can be radically different, e.g. Sino-Tibetan and Indo-European, posing a Non-Independent and Identically Distributed (non-IID) challenge when learning a global multilingual model. This situation accentuates privacy concerns, and highlights the need for effective privacy-preserving techniques when using multilingual LLMs. To this end, recent works attempt to address privacy-constrained fine-tuning for multilingual tasks and explore how different languages impact the federated process [60]. However, they primarily target high-resources languages; research on low-resource languages remains largely unexplored.

Addressing low-resource languages is essential to promoting technological fairness and protecting the linguistic diversity. Unlike their high-resource counterparts, low-resources languages pose intriguing

(a) Monolingual Tuning.



(b) Centralized Tuning.



(c) Federated Prompt Tuning.

Figure 1: Comparison of three different fine-tuning paradigms for multilingual tasks.

research challenges: i) **Limited computational resources.** Regions of low-resources languages are often economically developing areas, with little access to huge computational resources required to either train language models from scratch or fully fine-tune pre-trained large language models [38, 1]. ii) **Limited data in the target language.** Due to a small speaking population or the spoken nature of the language, data is often scarce [2, 43, 20]. As depicted in Figure 4, the pretraining data for LLMs is predominantly in English, with little coverage of low-resource languages. Under such circumstances, the performance of low-resources languages is often unsatisfactory during fine-tuning because of their under-representation. iii) **Memorization risk.** Recent studies find that as pre-trained models scale up, their ability to memorize training data increases [55]. This implies that, when fine-tuning these models with limited data, the risk of overfitting and potential privacy issues arises.

To counteract the above challenges, we turn to federated learning (FL), where the model training is done across multiple decentralized devices or servers while the data is always kept localized [42, 29, 65]. In a multilingual setting, FL becomes particularly natural, as data from diverse linguistic backgrounds can be sourced without compromising user privacy, and due to the geographical spread and inherent linguistic diversity of devices, data on each node is likely to exhibit non-IID distribution.

In this paper, in order to mitigate the physical border and the linguistic border of multilinguality, 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, our global encoder can discern language patterns and cluster languages via federated prompt averaging, which allows each client to benefit from others' data without direct access. This strategy requires minimal computational resources and significantly improves performance, particularly for low-resource languages. We demonstrate the effectiveness of our method on standard NLP tasks including New Classification and XNLI. The performance of our paradigm achieves 6.9% accuracy improvement while protecting the privacy of mulitlingual source data. Compared with other Federated Learning approches, our paradigm reduces computational cost and communication cost by more than 99%. Our approach paves the way for fine-tuning multilingual large language models on resource-constraint devices in a privacy-preserving way, and holds the potential to promote social equality, privacy, and linguistic diversity in the research community.

2 A New Paradigm for Multilinguality: Federated Prompt Tuning

2.1 Notation and Preliminaries

In our federated learning setup, we have K clients. Each client k has a private dataset, either monolingual or multilingual, defined as:

$$\mathcal{D}_k = \{(x_{k,i}, y_{k,i}) \mid i = 1, \dots, n_k\}$$

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 h_g. Conversely, each client k has its version, h_k, adjusted based on its dataset. Each prompt encoder, whether global or local, is composed of a series of trainable prompts: $h_0, h_1, h_2, \dots$. These prompts are adjusted during training to better aid the model.

2.2 Virtual Prompt Encoder

Prompt Learning is a parameter-efficient alternative to fine-tuning pretrained language models (PLM). Instead of selecting discrete text prompts in a manual or automated fashion, prompt tuning utilize virtual prompt embeddings that can be optimized via gradient descent.

Given the utilization of a prompt encoder, for instance the version h_k for client k, a textual prompt p_k tailored for a specific task can be generated. This prompt is subsequently concatenated or combined in another manner with the original input x, resulting in a modified input x':

$$x'_k = p_k \oplus x \tag{1}$$

Here, p_k represents the prompt generated by h_k.

The modified input x'_k is then processed by the encoder E of the pre-trained language model:

$$h'_k = E(x'_k) \tag{2}$$

The primary objective of each prompt encoder is to generate an effective prompt, such as p_k for client k, to guide the pre-trained model in producing the desired outputs. During the fine-tuning phase, based on a task-specific loss, the parameters of the prompt encoder h_k are often adjusted:

$$\mathcal{L}(x, y; h_k) = \text{Loss}(D(E(p_k \oplus x)), y) \tag{3}$$

Where D is a decoder that maps the internal representation to task outputs, and Loss is an appropriate loss function, like cross-entropy loss. Throughout the fine-tuning, both the model's parameters and the prompt encoder h_k's parameters are updated in accordance with this loss function.

2.3 Federated Prompt Averaging

Derived from FedAvg [42] mentioned in Appendix B, we propose the following federated prompt averaging algorithm (also shown in Algorithm 1):

Initialization: The server initializes the global prompt encoder with its prompts $h_0, h_1, \ldots$. Each client sets its local version tailored to its dataset.

Client Selection: Every communication round selects a fraction C of the total K clients for training. This subset is $m = \max(C \times K, 1)$. A subset S of m clients is chosen.

Local Model Training and Tuning: Each client k in S goes through: The client fetches the current global prompt encoder. It assembles a model using its local prompt encoder with prompts $h_0, h_1, \ldots$ and PLM parameters. Training on $\mathcal{D}_k$ fine-tunes the local prompt encoder and its prompts, while most of the PLM remains unchanged. After training, each client computes model updates, especially the refined local prompt encoder and its prompts.

Aggregation: The server aggregates updates from all clients. The global prompt encoder and its prompts $h_0, h_1, \ldots$ are updated using:

$$h_g = \frac{1}{m} \sum_{k=1}^{m} h_k$$

3 Evaluation: How Federated Prompt Tuning helps Multilinguality

3.1 Experimental Setup

Tasks and Datasets We evaluate our model using the popular XGLUE benchmark [31], a cross-lingual evaluation benchmark for our multilingual evaluation. We conduct our experiments on two classification tasks: News Classification (NC) and XNLI [18], covering both high-resource languages and low-resource languages, with details provided in Appendix D. Accuracy (ACC) of the multi-class classification is used as the metric for both of the tasks. Our base model for both tasks is the XLM-RoBERTa base-sized model (270M parameters) [16].

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 (Non-IID) **(Ours)**	87.8	79.2	73.7	83.1	79.5	80.7

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

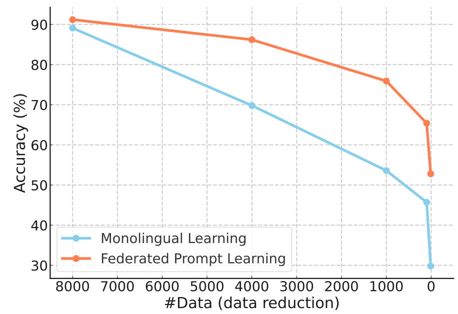


Figure 2: Performance comparison on NC with decreasing dataset size.

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
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
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**

Baselines 1) *Monolingual Fine-tuning (No FL):* Traditional local fine-tuning where a separate model is finetuned using the corresponding dataset for each single language; 2) *Centralized Fine-tuning (No FL):* Standard Fine-tuning using a combined dataset of all languages centralized in one location; 3) *Full Fine-tuning with FL:* Directly fine-tuning the whole pre-trained language model in a federated manner, with a full pre-trained model on each client; 4) *Prompt Fine-tuning with FL:* Only training the prompt encoder in a federated manner, with a prompt encoder on each client.

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.

3.2 Main Results

Table 1 presents the outcomes of experiments focused on news classification. When employing Prompt Tuning in comparison to Full Finetuning, 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 Prompt Tuning method as opposed to without it. For a comprehensive analysis of this, refer to the section 3.5.

Table 2 summarises the results of our FL experiments on the XNLI task. To accentuate the potency of our Federated Prompt Tuning approach, a juxtaposition was made with traditional monolingual training. As the data portrays, our Federated Prompt Tuning, particularly on Non-IID dataset, consistently outperformed the monolingual method across all languages. Remarkably, this superior performance was maintained even for languages with limited available data. The average score further substantiates the prowess of Federated Prompt Tuning, marking a noticeable improvement from 32.94% in the monolingual approach to 39.83% with Non-IID Federated Prompt Tuning.

From our results in section 3.2, 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. Additionally, the prompt encoder serves as a parameter-efficient alternative. By freezing the core language model parameters, we prevent the model from altering its foundational understanding of language. As a result, the model's propensity to overfit to a dataset is reduced, minimizing the risk of memorizing specific lexical cues and spurious correlations.

3.3 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 2, our Federated Prompt Tuning method consistently outperforms the traditional monolingual approach. As we reduce the dataset size from 8,000 to near-zero, the accuracy of 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.

3.4 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.

We define the *pretrained 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 serves as 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 cosine similarity to the pre-trained language. More details are shown in Appendix F. 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 3, a conspicuous trend is observed: As the language similarity to 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 finetuning becomes even more obvious for languages with less data, indicating that our Federated Prompt Tuning method offers significant advantages, particularly in low-resource scenarios.

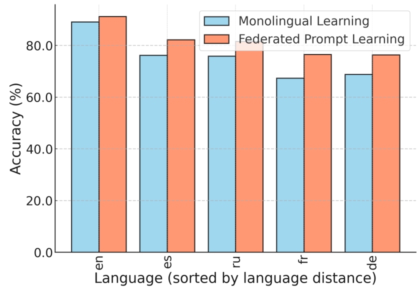
3.5 Ablation Study III: Parameter Efficiency

Computational Cost From the perspective of trainable parameters, this significant reduction in parameters demonstrates exceptional parameter efficiency. In both of the tasks, despite the total number of parameters exceeding 278 million, the trainable parameters are only around 1.2 million, accounting for less than 0.5% of the total. Such a design can substantially reduce training time and computational resources, while also mitigating the risk of overfitting. In the context of LLMs, this high parameter efficiency offers potential for model deployment in resource-constrained environments.

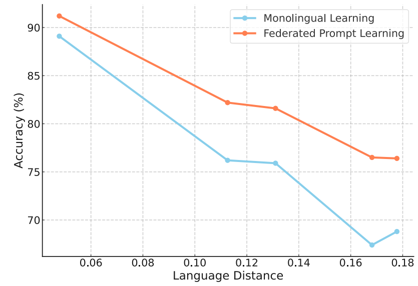
Table 3: Comparison of parameter efficiency and communication overhead in NC task.

	Federated Full Finetuning	**Federated Prompt Tuning (Ours)**	**Optimization Scale**
Trainable Params	278,655,764	1,202,708	**231.69**
Communication Cost	108GB	478.93MB	

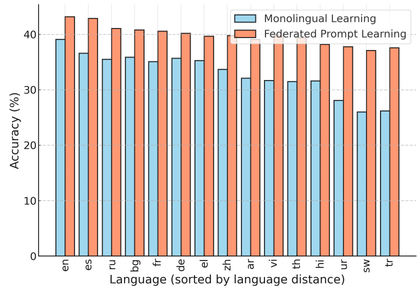
Communication Cost XLM-Roberta-Base's data transmission in Federated Learning with 5 clients and 10 communication rounds was 108 GB. After our optimization, using a prompt encoder with a 2x768 structure, the transmission size reduced to 478.93 MB, a 99% reduction shown in Table 3. This optimization enhances efficiency in federated prompt tuning and expands its applicability to bandwidth-constrained environments like edge devices and mobile networks.



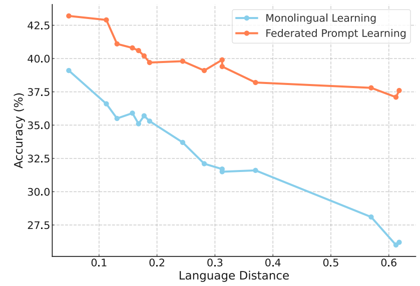
(a) Finetuning accuracy across different lanugages on the NC task.



(b) Finetuning accuracy across languages with varying similarity to the pre-trained language on the NC task.



(c) Finetuning accuracy across different lanugages on the XNLI task.



(d) Finetuning accuracy across languages with varying similarity to the pre-trained language on the XNLI task.

Figure 3: Comparative performance of traditional local finetuning and our Federated Prompt Tuning method across languages with varying similarity to the pre-trained language for XNLI and NC.

4 Conclusion

Future work Privacy attacks have been discussed in **(author?)** [21] on how gradient inversion can be used to attack language models and break the privacy protection that FL naturally adds. Future work may include privacy experiments and additional privacy protection with various secure aggregation (SA) [10, 7, 49] and differential privacy (DP) techniques [59, 50]. Future work will also explore the impact on the Multilingual Federated Prompt Tuning method as models scale up.

Social Impacts Addressing the physical and linguistic challenges 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. Its potential to democratize access to technology, preserve linguistic diversity, and ensure user privacy can have profound implications for the future of technology and society.

References

[1] 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)*, pages 3814–3841, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[2] 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.

[3] 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.

[4] 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)*, pages 1778–1796, 2022.

[5] 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*, pages 4623–4637, Online, July 2020. Association for Computational Linguistics.

[6] 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.

[7] James Henry Bell, Kallista A Bonawitz, Adrià Gascón, Tancrède Lepoint, and Mariana Raykova. Secure single-server aggregation with (poly) logarithmic overhead. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 1253–1269, 2020.

[8] 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)*, pages 1–9, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[9] 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.

[10] Keith 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. *arXiv preprint arXiv:1611.04482*, 2016.

[11] 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.

[12] 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*, pages 1324–1334, Online, November 2020. Association for Computational Linguistics.

[13] 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.

[14] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. 2022.

[15] 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)*, pages 4536–4546, Online, November 2020. Association for Computational Linguistics.

[16] 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*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.

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

[18] 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.

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

[20] 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)*, pages 6279–6299, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[21] Liam H. Fowl, Jonas Geiping, Steven Reich, Yuxin Wen, Wojtek Czaja, Micah Goldblum, and Tom Goldstein. Decepticons: Corrupted transformers breach privacy in federated learning for language models. *ArXiv*, abs/2201.12675, 2022.

[22] 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, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 09–15 Jun 2019.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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)*, pages 4483–4499, Online, November 2020. Association for Computational Linguistics.

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

[29] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

[30] 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)*, pages 4582–4597, Online, August 2021. Association for Computational Linguistics.

[31] 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.

[32] 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.

[33] 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*, pages 157–175, Seattle, United States, July 2022. Association for Computational Linguistics.

[34] 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, pages 8–14, 2017.

[35] 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, 2021.

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

[37] 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.

[38] 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*, pages 202–217, 2021.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[45] 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)*, pages 7654–7673, Online, November 2020. Association for Computational Linguistics.

[46] 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*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics.

[47] 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*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics.

[48] 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)*, pages 2362–2376, Online, November 2020. Association for Computational Linguistics.

[49] Xinchi Qiu, Heng Pan, Wanru Zhao, Chenyang Ma, Pedro PB Gusmao, and Nicholas D Lane. vfedsec: Efficient secure aggregation for vertical federated learning via secure layer. *arXiv preprint arXiv:2305.16794*, 2023.

[50] Thilina Ranbaduge and Ming Ding. Differentially private vertical federated learning, 2022.

[51] 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.

[52] Yu-An Chung Mariano Cora Meglioli David Dale Ning Dong Paul-Ambroise Duquenne Hady Elsahar Hongyu Gong Kevin Heffernan John Hoffman Christopher Klaiber Pengwei Li Daniel Licht Jean Maillard Alice Rakotoarison Kaushik Ram Sadagopan Guillaume Wenzek Ethan Ye Bapi Akula Peng-Jen Chen Naji El Hachem Brian Ellis Gabriel Mejia Gonzalez Justin Haaheim Prangthip Hansanti Russ Howes Bernie Huang Min-Jae Hwang Hirofumi Inaguma Somya Jain Elahe Kalbassi Amanda Kallet Ilia Kulikov Janice Lam Daniel Li Xutai Ma Ruslan Mavlyutov Benjamin Peloquin Mohamed Ramadan Abinesh Ramakrishnan Anna Sun Kevin Tran Tuan Tran Igor Tufanov Vish Vogeti Carleigh Wood Yilin Yang Bokai Yu Pierre Andrews Can Balioglu Marta R. Costa-jussà [3] Onur Celebi Maha Elbayad Cynthia Gao Francisco Guzmán Justine Kao Ann Lee Alexandre Mourachko Juan Pino Sravya Popuri Christophe Ropers Safiyyah Saleem Holger Schwenk Paden Tomasello Changhan Wang Jeff Wang Skyler Wang Seamless Communication, Loïc Barrault. Seamlessm4t—massively multilingual & multimodal machine translation. *ArXiv*, 2023.

[53] 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.

[54] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.

[55] 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.

[56] 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*, 2023.

[57] 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*, 2023.

[58] 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, page 1839–1850, New York, NY, USA, 2022. Association for Computing Machinery.

[59] 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.

[60] 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.

[61] 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*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

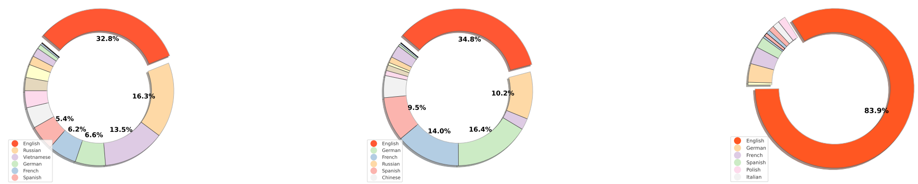
[62] Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844,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

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

[64] Hu Xu, Lei Shu, Philip Yu, and Bing Liu. Understanding pre-trained BERT for aspect-based sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 244–250, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[65] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

A Background



(a) XLM-R [16] (b) mBERT [46] (c) PaLM [14]

Figure 4: Linguistic coverage of different large language models.

B Related Work

Federated Learning. Federated Learning has garnered significant attention in the academic realm. A notable contribution from **(author?)** [26] underscores the potential of this methodology. One of its primary benefits is the execution of deep learning algorithms while maintaining an emphasis on user privacy, a premise originally posited by **(author?)** [54]. The overarching objective is to amalgamate insights from diverse data repositories without compromising sensitive particulars. In this spectrum, the FedAvg algorithm [42] stands out. This algorithm operates by independently training models on distinct client devices and subsequently aggregating their updates centrally for an averaged outcome. However, it's imperative to acknowledge that FedAvg, while powerful, is principally structured for identically and independently distributed (IID) data, while its application on Non-IID datasets may lead to potential discrepancies in results or even model instability [42, 26]. 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.

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 [6]. FedNLP introduces a benchmarking framework for evaluating various FL methods across NLP tasks, providing a universal interface between Transformer-based models and FL methods [33]. FedKC [58]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.

Multilingual Language Models. Multilingual Pretrained Language Models such as mBERT [46], XLM-R [16], and SeamlessM4T [52] have emerged as a viable option for bringing the power of pretraining to a large number of languages [19]. 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, translation dictionaries [47, 62, 5, 64].

However, these multilingual LMs are not without limitations. Particularly, **(author?)** [16] 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 [63, 24, 27] Prior work tried to address this issue by increasing the model capacity [5, 45, 12] or through additional training for particular language pairs [45, 48] or by clustering and merging the vocabularies of similar languages, before defining a joint vocabulary across all languages [15]. Despite these efforts, the multilingual LMs still struggle with balancing their capacity across many languages in an sample-efficient and parameter-efficient way [4, 41, 13].

Prompt Learning and Parameter-Efficient Fine-Tuning. The size of pre-trained language models has been increasing significantly [11], presenting challenges to traditional task transfer based on full-parameter finetuning. Recent research has shifted its attention to Parameter-Efficient Fine-Tuning techniques, such as prompt tuning [28, 30, 36], adapters [22], and combined approaches [23, 8]. These methods utilize a minimal number of tuning parameters, yet they offer transfer performance that is comparable with traditional finetuning.

Prompt learning is a burgeoning area in machine learning where models are steered towards desired outputs using prompts, typically without exposure to explicit label information [35]. This paradigm shows promise in effectively leveraging large pre-trained models in a data-efficient manner by reducing the need for extensive labeled datasets [51]. 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 [53].

C Federated Prompt Averaging Algorithm

Algorithm 1: Federated Prompt Averaging

Initialization:
Server initializes global prompt encoder h_g
Each client initializes local prompt encoder h_k

Server executes:
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:
  $h_g = \frac{1}{m} \sum_{k=1}^{m} h_k$
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

D Dataset

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.

E Implementation

We use Hugging Face's transformers library [61] and PEFT library [40] for loading pre-trained models and prompt tuning configurations. For our federated training and evaluation, we use the Flower framework [9] and PyTorch as the underlying auto-differentiation framework [44]. We use the AdamW optimizer [37, 25] for all experiments. All experiments are conducted using NVIDIA A40.

F Multilingual Distance Measurement

We leverage the database from **(author?)** [34, 39] 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 pretrained 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 $\phi(v_i) = -\log(\cos(v_i, V_{\text{p}}))$.