
Towards Global South AI Sovereignty: A Federated Learning Framework for Collaborative LLM Development¹

Phemelo Maile Mueletshedzi Mahlomola-
University of Moses Mubvafhi Mohlomi
Cape Town Independent Independent

With
Apart Research

Abstract

Advanced AI systems are increasingly important for education, research, governance, and economic development, but their development remains concentrated in a small number of companies and countries. This creates dependency risks for Global South institutions, particularly where local languages, data, and infrastructure are underrepresented. We propose a federated fine-tuning framework for collaborative LLM development, allowing institutions to train locally and share only model updates rather than centralizing raw data. As a proof of concept, we simulate a federation of five African-language nodes using AfriSenti sentiment data and compare standard Federated Averaging with a fairness-aware q -FFL variant. We also evaluate how performance changes as more nodes join the federation. The results show that parameter-efficient updates reduce per-client communication from approximately 544 MB to 2.96 MB, a reduction of about $184\times$. Node scaling improves mean validation accuracy from approximately 0.36 with one node to 0.42 with three nodes before plateauing. However, worst-language accuracy remains low, and q -FFL does not meaningfully protect the lowest-resource node. These findings suggest that federated fine-tuning can support data-local collaboration and reduce participation costs, but stronger language-aware and privacy-preserving mechanisms are needed for equitable Global South AI development.

¹ Research conducted at the [The Global South AI Safety Hackathon](#), June 2026

1. Introduction

Large language models (LLMs) are becoming increasingly important infrastructure for education, research, governance, software development, and economic productivity. As these systems become more capable, access to advanced AI tools may increasingly determine how quickly individuals, institutions, and countries can learn, innovate, and compete. However, the development and deployment of state-of-the-art AI systems remain concentrated within a small number of companies and countries. For many Global South institutions, this creates a dependency risk: local users may benefit from AI systems, but they often do not control the data pipelines, model development processes, compute infrastructure, or access policies that shape those systems[8].

This dependency is especially important in regions where local languages and contexts are underrepresented in mainstream AI systems. If models are mainly trained and fine-tuned using data from dominant languages and well-resourced regions, they may perform unevenly across lower-resource communities. This creates both a capability gap and an AI safety concern: communities that are poorly represented in model development may receive less reliable, less useful, or less contextually appropriate AI outputs. At the same time, centralizing local data in external cloud platforms raises concerns about data sovereignty, privacy, institutional autonomy, and long-term technological dependence[8],[12].

We propose federated fine-tuning as a practical pathway toward Global South AI sovereignty. Instead of requiring all institutions to upload local data to a central server, federated learning allows each institution to train locally and share only model updates[1],[11]. This enables collaborative model improvement while keeping raw data within the originating institution. To reduce the cost of participation, we combine federated learning with parameter-efficient fine-tuning[3],[9], so that participating nodes transmit lightweight trainable updates rather than full model parameters.

Although the broader motivation applies to the Global South, this proof-of-concept focuses on African-language model adaptation as a concrete case study. We simulate a federation of five African-language nodes using real AfriSenti data and evaluate whether collaborative fine-tuning can improve a shared multilingual model under resource imbalance[4],[5]. The nodes represent different levels of data availability, including high-resource, medium-resource, and low-resource language settings. We compare standard Federated Averaging with a fairness-aware q -FFL variant and also evaluate how model performance changes as more nodes join the federation.

Our main contributions are:

- 1. We implement a proof-of-concept federated fine-tuning framework for collaborative African-language model adaptation using real multilingual sentiment data.*
- 2. We demonstrate that parameter-efficient federated updates substantially reduce communication cost compared with transmitting full model parameters, making participation more feasible for resource-constrained institutions.*
- 3. We empirically evaluate whether adding more participating nodes improves the shared model and whether fairness-aware aggregation protects low-resource language nodes.*

4. We show that node participation improves average shared-model performance, but low-resource language protection remains an unresolved challenge requiring stronger language-aware aggregation methods.

2. Related Work

2.1 Federated Learning and Federated Averaging

Federated learning was introduced to train shared models from decentralized data without requiring raw data to be collected in one location[1]. In the standard Federated Averaging approach, clients train local models on their own data and send model updates to a central coordinator, which aggregates the updates into a new global model[1]. This makes federated learning attractive for settings where data is distributed across institutions, devices, or jurisdictions. In the context of AI sovereignty, this architecture is relevant because it allows institutions to contribute to a shared model while retaining local control over their datasets[11].

However, standard federated learning is not automatically fair. When clients have unequal dataset sizes or highly heterogeneous data distributions, the global model may favour larger or easier clients. This is particularly important for multilingual AI development, where high-resource languages may dominate training while low-resource languages remain poorly served.

2.2 Fairness-Aware Federated Learning

To address uneven client performance, fairness-aware federated learning methods attempt to reduce disparities across participating nodes[2]. q-Fair Federated Learning (q-FFL) modifies the aggregation objective so that clients with higher losses can receive greater influence during training. This is useful in principle for multilingual settings, where weaker-performing language nodes may need additional weight to avoid being overwhelmed by high-resource nodes.

In this work, we test q-FFL as a simple fairness-aware baseline for low-resource language protection. Our goal is not to claim that q-FFL is the final solution, but to evaluate whether a straightforward fairness-aware aggregation rule is sufficient in an African-language federated setting. As shown in our results, q-FFL provides small average gains but does not fully protect the lowest-resource node, suggesting the need for stronger language-aware aggregation strategies.

2.3 Parameter-Efficient Fine-Tuning and LoRA

Fine-tuning large language models by updating all model parameters is expensive in both memory and communication cost. Parameter-efficient fine-tuning methods address this by freezing most of the model and training only a small number of additional parameters. Low-Rank Adaptation (LoRA) is one such method: it injects trainable low-rank matrices into a pretrained model while keeping the original model weights fixed[3].

LoRA is particularly suitable for federated learning in resource-constrained environments because each client only needs to train and transmit a small set of adapter parameters rather than the full model. This reduces both local compute requirements and communication overhead, making collaborative model adaptation more realistic for universities, civic organizations, and public institutions that may not have access to large AI compute clusters[3],[9].

2.4 African-Language NLP and AfriSenti

African languages remain underrepresented in many NLP benchmarks and large-scale language model training pipelines. AfriSenti helps address this gap by providing a multilingual sentiment analysis benchmark across several African languages, including Swahili, Hausa, Yoruba, Igbo, and Xitsonga[4],[5]. Because the dataset contains language-specific sentiment labels, it provides a useful testbed for simulating multilingual federated learning across language nodes.

In this project, we use AfriSenti as a proxy for institutionally held language data. Each language node represents a participating institution or community that keeps its local data private while contributing to a shared model through federated updates. This setup allows us to study both collaborative model improvement and the risk that low-resource languages remain under-served even in a federated system.

2.5 AI Sovereignty and Global South AI Safety

AI sovereignty refers to the ability of communities, institutions, and states to shape the development, deployment, and governance of AI systems according to their own priorities[8]. For Global South contexts, this includes reducing dependency on externally controlled infrastructure, improving representation of local languages and values, and ensuring that participation in AI development is not limited to actors with access to large centralized compute.

Federated learning does not solve all sovereignty problems by itself. It does not automatically guarantee privacy, fairness, or independence from foreign infrastructure. Model updates may still leak information, and low-resource nodes may still be underweighted during aggregation[9],[11]. However, federated fine-tuning provides a useful technical foundation for decentralized cooperation. Our work explores this foundation experimentally and identifies where further mechanisms, such as secure aggregation, differential privacy, and language-aware weighting, are needed.

3. Methods

3.1 Overview

We evaluate whether federated fine-tuning can support collaborative language model development under Global South resource constraints. The experiment simulates a federation of institutions, where each node holds data for one African language and trains locally without sharing raw text data. The central server only receives trainable model updates and aggregates them into a shared global model. This design is intended to test three practical goals: data-local collaboration, reduced communication cost, and improved shared-model performance as more nodes participate.

3.2 Dataset and Federated Node Setup

We use AfriSenti, a multilingual African sentiment analysis dataset, as the experimental testbed[4],[5]. Five language-specific nodes are created: Swahili, Hausa, Yoruba, Igbo, and Xitsonga. Each node is treated as a separate participating institution or language community. To simulate real-world resource imbalance, the nodes are assigned different training set sizes. Swahili and Hausa are treated as high-resource nodes with 500 training samples each, Yoruba and Igbo as medium-resource nodes with 125 samples each, and Xitsonga as a low-resource node with 25 samples. Each node is evaluated using a held-out validation set of 150 samples.

This setup creates a non-IID federated learning environment: each client has data from a different language and the amount of available training data differs significantly between clients. This reflects a common challenge in multilingual AI development, where some languages have substantially more labeled data than others.

3.3 Model and Parameter-Efficient Fine-Tuning

The base model is a multilingual transformer encoder used for three-class sentiment classification: negative, neutral, and positive[6],[7]. Instead of fine-tuning all model parameters, we use parameter-efficient fine-tuning based on low-rank trainable updates. The backbone model remains frozen, while only a small set of trainable adapter and classification-head parameters are updated locally.

This design reduces the cost of federated participation. In a standard full-model federated setup, each client would need to transmit the full model update after every local training stage. In our setup, each node transmits only the lightweight trainable update[3],[9]. This is important for resource-constrained environments where bandwidth, compute, and storage may be limited.

3.4 Federated Training Procedure

At the beginning of each communication round, the global model is copied to each participating node. Each node fine-tunes the trainable parameters using only its local dataset. After local training, the raw text data remains at the node, while the trainable model update is sent to the server for aggregation. The server combines the updates to form the next global model, which is then redistributed for the next communication round.

The standard Federated Averaging update is given by[1]:

$$w_{t+1} = \sum_{k=1}^K \frac{n_k}{N} w_{k,t}$$

where (w_{t+1}) is the updated global model, (K) is the number of participating nodes, ($w_{k,t}$) is the locally trained model update from node (k), (n_k) is the number of training samples at node (k), and $N = \sum_{k=1}^K n_k$ is the total number of samples across all participating nodes. Under this rule, nodes with more data contribute more strongly to the global update.

3.5 Fairness-Aware Aggregation

To test whether low-resource nodes can be better protected, we also evaluate a q -FFL-inspired aggregation method. Unlike standard FedAvg, which weights updates primarily by dataset size, q -FFL increases the influence of clients with higher training loss. In principle, this can help nodes that are poorly served by the current global model.

In our implementation, the aggregation weight for client (k) is proportional to[2]:

$$\alpha_k = \frac{n_k F_k(w_t)^q}{\sum_{j=1}^K n_j F_j(w_t)^q}$$

where $(F_{k(w)})$ is the local training loss of client (k) , (n_k) is the local dataset size, and (q) controls the strength of fairness weighting. We use $(q = 2.0)$ in the main experiment. This method is evaluated against standard FedAvg to test whether fairness-aware weighting improves performance for low-resource language nodes.

3.6 Experiments

We run three experiments.

First, we compare FedAvg and q -FFL across five African-language nodes. Each method is run for eight communication rounds across three random seeds. We report final validation accuracy per language, mean accuracy, worst-language accuracy, and macro-F1.

Second, we measure communication efficiency by comparing the size of the trainable federated update against the estimated size of a full model update. This evaluates whether parameter-efficient federated tuning can reduce the communication burden for participating institutions.

Third, we conduct a node-scaling experiment. The number of participating nodes is increased from one to five, and the resulting global model is evaluated across the same validation suite. This experiment tests whether adding more participating institutions improves the shared model while keeping raw data local.

3.7 Evaluation Metrics

We evaluate the model using validation accuracy and macro-F1 for each language node. To measure overall model quality, we compute the mean validation accuracy across all language nodes. To evaluate low-resource protection, we report the worst-language accuracy, defined as the minimum validation accuracy across all language nodes. We also report the fairness gap, defined as the difference between the best-performing and worst-performing language nodes.

4. Results

4.1 Multi-node federated fine-tuning

We first evaluated whether a shared model could be fine-tuned across five African-language nodes while keeping each node’s raw data local. The federation consisted of Swahili and Hausa as high-resource nodes, Yoruba and Igbo as medium-resource nodes, and Xitsonga as the low-resource node. Each method was run for eight communication rounds across three random seeds. We compared standard Federated Averaging (FedAvg) with a q -FFL aggregation variant designed to increase the influence of higher-loss clients.

Table 1 reports the final-round validation accuracy averaged across three seeds. q -FFL provides small gains for Swahili, Hausa, Yoruba, and Igbo, but it does not improve the lowest-resource node. Xitsonga remains the weakest-performing language under both aggregation methods.

Table 1. Final-round validation accuracy across three random seeds.

Language	Resource level	FedAvg	q -FFL	Difference
Swahili	High	0.500 ± 0.084	0.522 ± 0.079	+0.022
Hausa	High	0.551 ± 0.027	0.562 ± 0.025	+0.011
Yoruba	Medium	0.353 ± 0.022	0.369 ± 0.027	+0.016

<i>Igbo</i>	<i>Medium</i>	0.451 ± 0.008	0.462 ± 0.008	$+0.011$
<i>Xitsonga</i>	<i>Low</i>	0.178 ± 0.014	0.169 ± 0.014	-0.009

Across all languages, *q*-FFL improves the mean final accuracy from approximately 0.407 to 0.417. However, worst-language accuracy decreases slightly from approximately 0.178 under FedAvg to 0.169 under *q*-FFL. This suggests that *q*-FFL provides a small average performance gain, but does not sufficiently protect the lowest-resource node in this setting.

4.2 Communication efficiency

A key goal of the proposed framework is to reduce the cost of participation for resource-constrained institutions[9],[10]. The base model used in the experiment contains approximately 136 million parameters. A full 32-bit model update would require approximately 544 MB per client per communication round. In contrast, the parameter-efficient setup transmits only the trainable adapter and classification-head parameters, corresponding to approximately 740,000 trainable parameters or 2.96 MB per client update.

This represents an approximate communication reduction of:

$$\left[\frac{544.27\text{MB}}{2.96\text{MB}} = 183.9 \approx 184 \times \right]$$

This result supports the practical feasibility of parameter-efficient federated fine-tuning for Global South institutions where bandwidth, storage, and compute resources may be limited. Instead of requiring each institution to fine-tune and transmit a full model, each node can participate by training and sharing a much smaller update.

4.3 Node-scaling experiment

We next evaluated whether adding more participating nodes improves the shared model. The number of training nodes was increased from one to five, while each resulting global model was evaluated across the same validation suite. As shown in **Figure 1**, mean validation accuracy increases from approximately 0.36 with one participating node to approximately 0.42 with three participating nodes, after which performance begins to plateau. This suggests that adding institutions improves shared-model performance up to a point, but that the gains diminish as additional heterogeneous or lower-resource nodes are included.

To examine whether this improvement also benefits the weakest-performing language, we report worst-language accuracy in **Figure 2**. Worst-language accuracy improves modestly from approximately 0.16 to 0.18 as the number of participating nodes increases. However, the improvement remains small, indicating that broader participation alone is not sufficient to solve low-resource language underperformance.

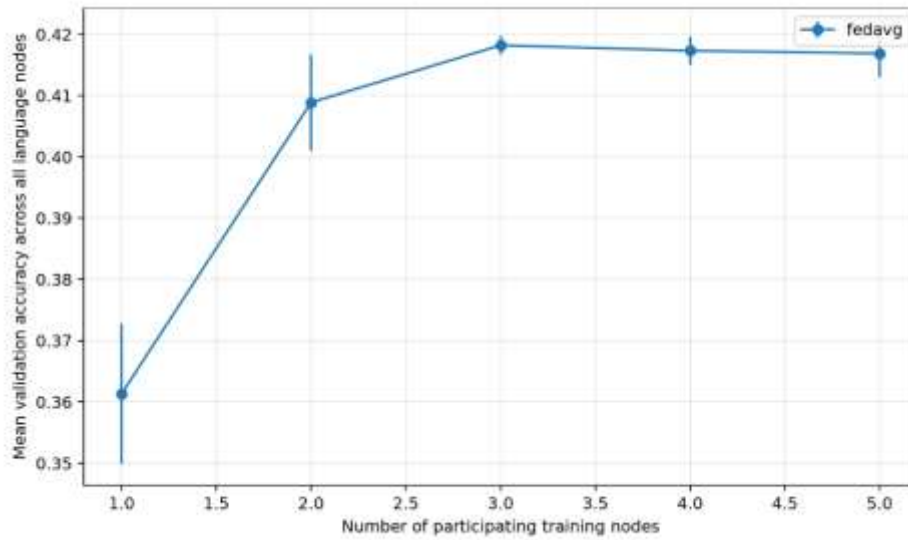


Figure 1: Mean validation accuracy as the number of participating federated nodes increases. Performance improves up to three nodes before plateauing.

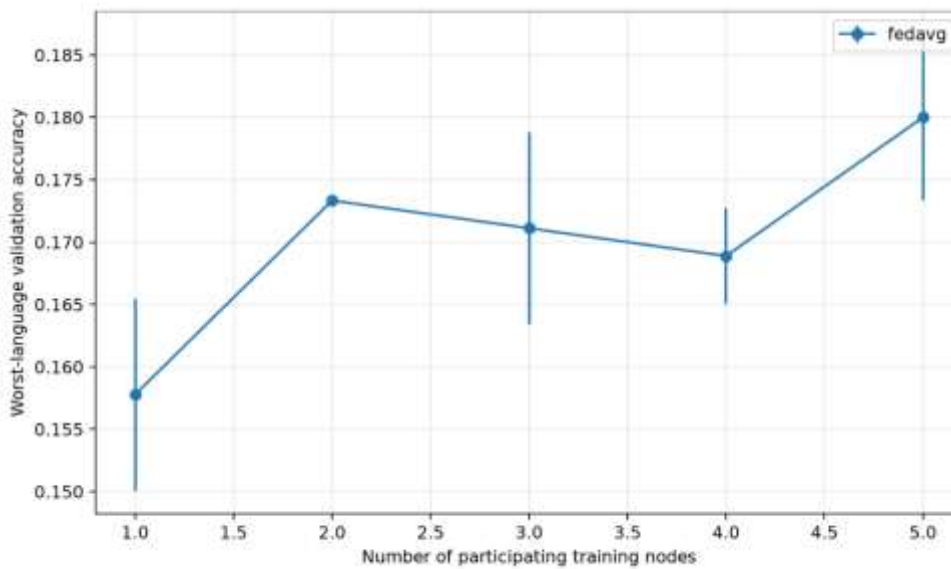


Figure 2: Worst-language validation accuracy as the number of participating federated nodes increases. Performance improves modestly but remains low.

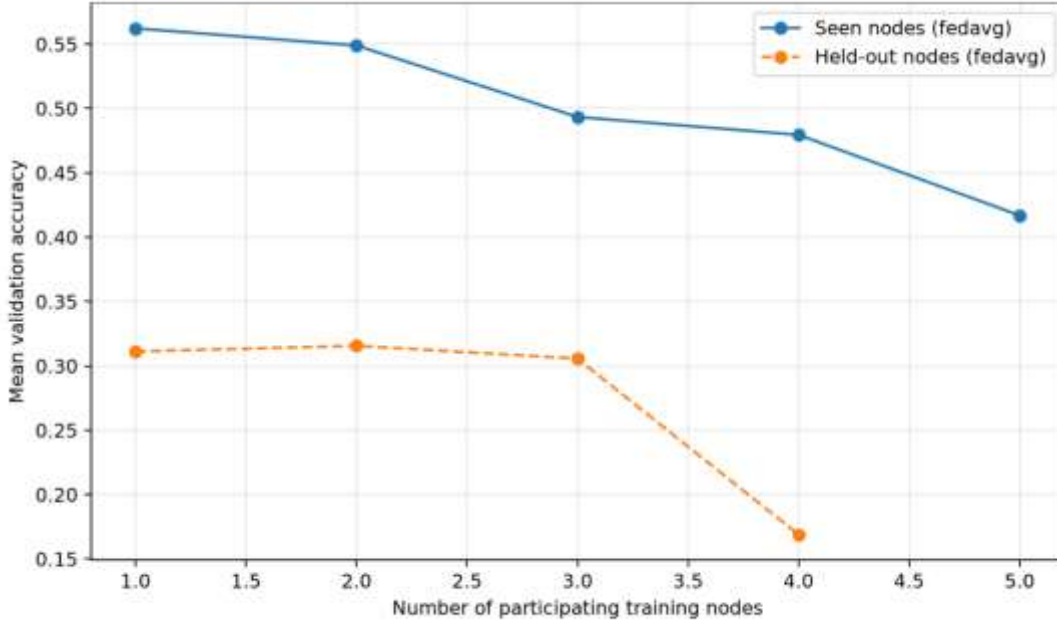


Figure 3: Mean validation accuracy for seen and held-out language nodes as federation size increases.

4.4 Seen versus held-out language performance

We also compare performance on language nodes included in federated training against those not yet included. As shown in **Figure 3**, seen-node accuracy is highest when the federation contains only the early high-resource nodes, but decreases as more medium- and low-resource languages enter the seen group. This should not be interpreted as the model becoming worse; rather, the seen-node evaluation group becomes more difficult as lower-resource languages are added. The held-out results further suggest that training on a subset of languages does not automatically generalize equally across all African-language nodes.

4.5 Summary of findings

The results support three main findings. First, federated fine-tuning can be used to collaboratively adapt a shared multilingual model across African-language nodes without centralizing raw text data. Second, parameter-efficient updates substantially reduce communication cost, making participation more feasible for resource-constrained institutions. Third, adding more participating nodes improves mean shared-model performance, but low-resource language protection remains unresolved. In particular, q-FFL provides small average gains but does not improve the weakest node, suggesting that future systems require stronger language-aware aggregation methods.

5. Discussion and Limitations

The results show that federated fine-tuning is a promising technical pathway for collaborative AI development in the Global South. By allowing each node to train locally and share only model updates, the framework supports a data-local model of cooperation. This is important for institutions

that may be unable or unwilling to centralize sensitive language data, institutional data, or community-specific datasets on external platforms.

The node-scaling experiment suggests that broader participation can improve shared-model performance. Mean validation accuracy improves as the number of participating nodes increases, indicating that collaboration across institutions can provide benefits beyond single-node training. This supports the central motivation of the work: instead of requiring one institution to collect all data and carry the full training burden, multiple institutions can contribute smaller local updates to improve a shared model.

The communication-efficiency result is also important. Since the framework transmits only trainable parameter-efficient updates rather than full model parameters, each client sends a much smaller update per communication round. This reduces the bandwidth and storage burden on participating nodes, making the approach more realistic for resource-constrained institutions.

However, the results also show that federated learning alone does not solve low-resource language underperformance. Xitsonga remains the weakest-performing node across experiments, and q-FFL does not improve its final accuracy compared with FedAvg. This suggests that simply adding more nodes or using a basic fairness-aware aggregation method is not enough to protect severely underrepresented languages. Stronger language-aware aggregation methods, targeted sampling, personalized adapters, or minimum-weight guarantees may be needed.

There are also important privacy limitations. Although raw text data is not centralized in our framework, this does not mean that privacy is fully guaranteed. Model updates can leak information under certain attacks, especially when clients have small datasets[9],[11]. Therefore, future versions of the system should incorporate secure aggregation, differential privacy, or other privacy-preserving mechanisms before being used with sensitive institutional or community data.

This work is also limited by scale. The experiment uses a small transformer model, five language nodes, and limited training samples. The setup is useful as a proof of concept, but it does not fully represent the complexity of deploying federated LLM development across real institutions with different hardware, network conditions, governance rules, and data-quality standards. The task is also limited to sentiment classification, so future work should test generative and instruction-following tasks.

Future Work

Future work should focus on four directions. First, stronger low-resource protection mechanisms should be developed, such as language-aware aggregation, client reweighting with minimum contribution guarantees, or personalized adapters for each language community. Second, privacy-preserving methods such as secure aggregation and differential privacy should be integrated to reduce the risk of information leakage from model updates. Third, the framework should be tested on larger models and more complex tasks, including instruction tuning and multilingual question answering. Finally, future experiments should involve more languages and more realistic institutional settings, where participating nodes differ not only in data size but also in compute capacity, bandwidth, and governance constraints.

6. Conclusion

This paper presented a proof-of-concept federated fine-tuning framework for collaborative LLM development in the Global South. Using African-language sentiment classification as a case study, we showed that multiple language-specific nodes can contribute to a shared model while keeping raw text data local. The framework also reduces communication cost by transmitting only parameter-efficient trainable updates instead of full model parameters.

Our results show that increasing the number of participating nodes improves mean shared-model performance up to a plateau, supporting the idea that institutions can benefit from collaborative training without centralizing their data. However, the experiments also show that low-resource language protection remains a major challenge. The lowest-resource node continued to perform poorly, and q-FFL did not substantially improve its outcome compared with FedAvg. These findings suggest that federated learning is a useful foundation for Global South AI sovereignty, but future systems must combine federation with stronger privacy-preserving and language-aware fairness mechanisms.

Code and Data

- **Code repository:** <https://github.com/Phemel01/global-south-federated-llm.git>
- **Data/Datasets:** We use the AfriSenti-Twitter dataset from Hugging Face: <https://huggingface.co/datasets/HausaNLP/AfriSenti-Twitter>. The experiments use five language subsets: Swahili, Hausa, Yoruba, Igbo, and Xitsonga.
- **Other artifacts:** The repository includes the experiment scripts, generated metric files, and figures used in the report, including the multi-seed FedAvg/q-FFL comparison and node-scaling plots

References

- [1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282. Available: <https://arxiv.org/abs/1602.05629>
- [2] T. Li, M. Sanjabi, A. Beirami, and V. Smith, “Fair Resource Allocation in Federated Learning,” in *International Conference on Learning Representations (ICLR)*, 2020. Available: <https://arxiv.org/abs/1905.10497>
- [3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-Rank Adaptation of Large Language Models,” in *International Conference on Learning Representations (ICLR)*, 2022. Available: <https://arxiv.org/abs/2106.09685>
- [4] S. H. Muhammad, I. Abdulmumin, A. A. Ayele, N. Ousidhoum, D. I. Adelani, S. Yimam, I. Ahmad, M. Beloucif, S. Ruder, O. O. Mayhew, K. Ogueji, and S. M. Mohammad, “AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages,” in *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023. Available: <https://aclanthology.org/2023.emnlp-main.862/>

[5] S. H. Muhammad, I. Abdulmumin, S. M. Yimam, D. I. Adelani, I. S. Ahmad, N. Ousidhoum, A. Ayele, S. M. Mohammad, M. Beloucif, and S. Ruder, “SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval),” in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2023. Available: <https://aclanthology.org/2023.semeval-1.315/>

[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter,” arXiv preprint arXiv:1910.01108, 2019. Available: <https://arxiv.org/abs/1910.01108>

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186. Available: <https://aclanthology.org/N19-1423/>

[8] V. S. S. A. Daliparthi, K. Tutschku, V. KEBANDE, and N. Momen, “Digital Sovereignty for Collaborative AI Engineering: A Survey,” *IEEE Access*, 2025. DOI: 10.1109/ACCESS.2025.3647085

[9] N. Yan, Y. Su, Y. Deng, and R. Schober, “Federated Fine-Tuning of LLMs: Framework Comparison and Research Directions,” *IEEE Communications Magazine*, 2025. DOI: 10.1109/MCOM.001.2400770

[10] Y.-W. Chu, D.-J. Han, and C. G. Brinton, “Only Send What You Need: Learning to Communicate Efficiently in Federated Multilingual Machine Translation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 33, pp. 1907–1921, 2025. DOI: 10.1109/TASLPRO.2025.3564854

[11] G. Schumann, J.-P. Awick, and J. M. Gómez, “Natural Language Processing using Federated Learning: A Structured Literature Review,” in *AIBThings*, 2023.

[12] J. Buys, “Large Language Models for African Languages,” SACAIR Tutorial, University of Cape Town, 2025.

LLM Usage Statement

We used ChatGPT and Claude to assist with brainstorming the project framing, developing and testing code, debugging implementation issues, interpreting experimental results, improving figure captions, and drafting sections of the report. All experiments were run locally by the authors, and all numerical claims, figures, and conclusions were checked against the generated result files before inclusion in the final submission.