

Personalized Federated Fine-Tuning of Language Models

Gabriel Ukstin Talasso

Advisor: Leandro Aparecido Villas, Co-advisor: Allan Mariano de Souza

Resumo. Atualmente, modelos de linguagem vêm sendo utilizados em diversas aplicações, frequentemente adaptados com dados de maior qualidade das tarefas alvo. Porém, o difícil acesso aos dados, distribuídos em diversas fontes e com restrições de privacidade, inviabiliza o treinamento tradicional centralizado. O Aprendizado Federado surge como alternativa, possibilitando o treinamento distribuído sem o compartilhamento de dados brutos. No entanto, a adaptação federada apresenta dois principais desafios: altos custos de treinamento e comunicação e heterogeneidade dos dados, variando em domínios e distribuições. Dessa forma, essa dissertação desenvolve soluções baseadas na personalização dos modelos através do agrupamento de clientes combinadas a técnicas de ajuste eficiente, em quatro principais contribuições: (i) um método agrupamento de clientes baseado em similaridade de modelos para cenários clássicos de FL (ii) um pipeline de ajuste eficiente de modelos de linguagem federados e seus trade-offs (iii) um método que une o agrupamento de clientes com adaptação eficiente para construir modelos multilíngues especializados por idioma e (iv) um novo método de personalização centrado na tarefa, juntando mecanismos de agrupamento local e global com roteamento adaptativo na inferência a tempo de teste. Os resultados e desenvolvimentos destacam a melhora das métricas de performance e eficiência de treinamento quando comparados com diversas soluções da literatura para cenários restritos de recursos e sobre heterogeneidade de dados, demonstrando sua eficácia e importância.

Abstract. Language Models have recently been successfully applied to a wide range of applications, typically fine-tuned using high-quality task-specific data. However, the limited accessibility of such data, due to its distribution across multiple sources and privacy constraints, makes traditional centralized training impractical. Federated Learning emerges as an alternative by enabling distributed model training without sharing raw data. Nevertheless, federated adaptation introduces two main challenges: the high training and communication costs and the heterogeneity of data, which differ in domains and distributions. To address these challenges, this dissertation develops personalized federated adaptation methods that combine client clustering with parameter-efficient fine-tuning techniques, resulting in four main contributions: (i) a model-similarity-based client clustering method for classical FL scenarios; (ii) a parameter-efficient fine-tuning pipeline for federated language models and its trade-offs; (iii) a method that integrates client clustering with efficient adaptation to build multilingual language models specialized for each language; and (iv) a novel task-centric personalization approach that combines local and global clustering mechanisms with adaptive inference-time routing. The proposed methods improved both performance and training efficiency compared with several state-of-the-art approaches under resource-constrained and heterogeneous data settings, demonstrating their effectiveness and practical relevance.

Application Information:

Institution: Instituto de Computação, Universidade Estadual de Campinas (UNICAMP)

Advisor: Leandro Aparecido Villas, UNICAMP

Co-advisor: Allan Mariano de Souza, UNICAMP

Track: MSc

Dissertation Defense Date: 24/04/2026

Examination Committee Members:

1. Leandro Aparecido Villas [Advisor] (UNICAMP)
2. Roberto de Alencar Lotufo (UNICAMP)
3. Helena de Almeida Maia (UNICAMP)

Link to the complete dissertation: https://drive.google.com/drive/folders/1KRecTGpc3jjlhDPaIAopDDNw_VeFce3B?usp=sharing

Links to publications derived from the dissertation¹:

1. **Gabriel U. Talasso**, Allan M. de Souza, Meghdad Kurmanji, Nicholas D. Lane, Leandro A. Villas. *FedRouter: Task-Centric Personalized Federated Fine-Tuning of Multitask Language Models*. Accepted in **35th International Conference on Artificial Neural Networks**, 2026 DOI not available yet. [Talasso et al. 2026b].
2. **Gabriel U. Talasso**, Allan M. de Souza, Luis F. G. Gonzalez, Eduardo Cerqueira, Antonio A. F. Loureiro, Leandro A. Villas. *Leveraging Federated Learning for Multilingual and Private Language Models via Model Clustering*. In **3rd IEEE International Conference on Federated Learning Technologies and Applications**, 2025. DOI: 10.1109/FLTA67013.2025.11336764 [Talasso et al. 2025b].
3. **Gabriel U. Talasso**, Allan M. de Souza, Daniel Guidoni, Eduardo Cerqueira, Leandro A. Villas. *Fine-Tuning Eficiente de Modelos de Linguagem Para Detectar Anomalias em Logs Privados usando Aprendizado Federado*. **Honorable Mention** in **SBRC - Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos**, 2025. DOI: 10.5753/sbrc.2025.5859 [Talasso et al. 2025a]. 🏆²
4. **Gabriel U. Talasso**, Allan M. de Souza, Luiz F. Bittencourt, Eduardo Cerqueira, Antonio A. F. Loureiro, Leandro A. Villas. *FedSCCS: Hierarchical Clustering with Multiple Models for Federated Learning*. In **IEEE International Conference on Communications**, 2024. DOI: 10.1109/ICC51166.2024.10622346 [Talasso et al. 2024].
5. **Gabriel U. Talasso**, Allan M. de Souza, Leandro A. Villas. *FedDALE: Detecção de Anomalias Eficiente em Logs Distribuídos com Pequenos Modelos de Linguagem Federados*. Accepted in **Workshop SBRC - Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos - WIARC**, 2026. DOI: 10.5753/wiarc.2026.229369 [Talasso et al. 2026a].

Google Scholar profile: <https://scholar.google.com/citations?user=9Mh6btYAAAAJ>

Github profile: <https://github.com/GabrielTalasso>

¹Non-first-author contributions in the Appendix

²🏆 This paper received an Honorable Mention as one of the best papers in the conference's main track.

1. Introduction

In recent years, *Large Language Models (LLMs)* have become prominent due to their general capabilities across many contexts. However, both the self-supervised pre-training process and the post-training adaptation process remain highly expensive due to the requirement to update hundreds of billions of parameters over multiple training steps with massive datasets [Radford et al. 2019, Achiam et al. 2023, Touvron et al. 2023a, Touvron et al. 2023b, Dubey et al. 2024]. In response, *Small Language Models (SLMs)* have emerged, trading the flexibility of LLMs for specialization at a lower computational cost [Lu et al. 2024], as well as *Parameter-Efficient Fine-Tuning (PEFT)* techniques [Houlsby et al. 2019] have been proposed to reduce the training cost associated with updating all model parameters by adapting only an extra small subset of weights. Among them, *Low-Rank Adapters (LoRA)* [Hu et al. 2021] achieve state-of-the-art performance by optimizing models through low-rank matrices [Dettmers et al. 2024].

However, in both the pre-training and adaptation, the quantity and quality of the data used are important for resulting model performance [Kaplan et al. 2020]. Nevertheless, the scarcity of new data, the distribution of data across multiple sources, and recent regulations in the use of private data [Ye et al. 2024b, Kuang et al. 2024, Zhang et al. 2024] contribute to the difficulty of obtaining high-quality datasets for model adaptation [Villalobos et al. 2022]. In this scenario, *Federated Learning (FL)* has emerged as a solution by training models distributed and sharing knowledge model parameters, without requiring direct data sharing, thereby reducing privacy risks while enabling collaborative training [McMahan et al. 2017]. However, FL presents its own challenges when applied to the training and adaptation of Language Models (LMs), the main being (i) the training cost, often on resource-constrained devices [Bai et al. 2024], and (ii) the divergence of client data distributions, where clients may have distinct tasks that hinder global training and convergence, resulting in suboptimal performance [Li et al. 2022, Zhu et al. 2021, Arivazhagan et al. 2019, Talasso et al. 2024].

Specifically, ***personalization*** methods has consistently achieved state-of-the-art (SOTA) performance in many FL applications [Tan et al. 2022] by adapting the global model to each client’s data distribution, personalization mitigates the effects of heterogeneous datasets while leveraging collaboration to acquire general knowledge [Arivazhagan et al. 2019, T Dinh et al. 2020]. Approaches such as multi-task learning, meta-learning, local post-training adaptation, and clustered FL balance global collaboration with local specialization. However, most existing methods are not specifically designed to address the challenges of text data, NLP tasks, and the transformer-based architectures underlying LLMs [Ye et al. 2024b, Kuang et al. 2024, Ye et al. 2024a].

In light of these challenges, this dissertation aims to explore and develop new personalized FL methods, based on clustering of clients, capable not only of improving model performance on clients’ local data distributions under heterogeneous federation settings, but also of explicitly accounting for training and communication costs through the use of specialized PEFT techniques. Furthermore, we address limitations of traditional approaches in the literature, such as the loss of model generality, which increases sensitivity to distribution shifts, as well as the negative effects of multi-task training within individual clients, which can hinder local model learning. The key contributions are *sequential and progressive advancements* and are summarized as follows:

- **Contribution 1:** We propose FedSCCS, a clustered method for classical FL that groups clients based on model representation similarity to address client heterogeneity. We also analyze client selection, representation, and the tradeoff between cluster specialization and global generalization.
- **Contribution 2:** We propose a LoRA-based federated efficient anomaly detection method using SLMs for privacy-preserving distributed log analysis. We further evaluate the performance, computational, and communication tradeoffs of federated LM fine-tuning.
- **Contribution 3:** We introduce FedMLLM, a method that unifies clustered federated with an efficient fine-tuning method that groups multilingual clients according to linguistic similarity. The approach improves knowledge sharing while reducing negative transfer in highly heterogeneous settings.
- **Contribution 4:** We identify key limitations of existing personalized federated LM fine-tuning methods regarding task interference and generalization. We then propose FedRouter, a state-of-the-art two-stage clustering method that addresses these challenges, improving the performance in federated efficient fine-tuning.

2. Related Work

Recent advances in FL fine-tuning of LLMs have focused on improving both efficiency and scalability. Among PEFT methods, LoRA [Hu et al. 2021] has become the most adopted approach by introducing trainable adaptation matrices while keeping the pre-trained model frozen. Several works extend LoRA to FL, such as SLoRA [Babakniya et al. 2023] and FlexLoRA [Bai et al. 2024], which leverage Singular Value Decomposition (SVD) to improve performance with heterogeneous devices. In parallel, frameworks such as OpenFedLLM [Ye et al. 2024b], FederatedScopeLLM [Kuang et al. 2024], and FedLLM-Bench [Ye et al. 2024a] provide benchmarks, datasets, and evaluation protocols for FL LLM fine-tuning. Nevertheless, these studies also show that traditional FL methods, such as FedAvg, struggle when clients perform different tasks or distinct domains, motivating the development of personalized fine-tuning methods.

To address data heterogeneity, several personalized FL methods have recently been proposed. FedDPA [Long et al. 2024] introduces a dual-adaptor architecture composed of global and local adapters, enabling collaborative knowledge sharing while preserving client specialization. FFA-LoRA [Sun et al. 2024] exploits the observation that LoRA A matrices are highly similar across heterogeneous clients by freezing them and only adapting the B matrices, whereas FedSA [Guo et al. 2024] further extends this idea, based on the principles observed in HydraLoRA [Tian et al. 2024] by sharing only the A matrices while keeping the B matrices local, simultaneously reducing communication costs and improving performance. These methods consistently outperform classical FL approaches in client-specific tasks due to their capability of avoiding client heterogeneity.

Despite these advances, existing personalization methods are primarily evaluated under single-task client settings and do not address scenarios where clients simultaneously contain multiple tasks or experience distribution shifts over time. As a result, important challenges related to model generalization and internal task interference remain largely unexplored. These limitations motivate the investigation of new personalized FL fine-tuning approaches capable of improving generalization while mitigating conflicts caused by heterogeneous and multi-task client data.

3. Methodology

This section presents the four contributions of the dissertation, each described in a dedicated subsection through its methodology and main experimental results, together with the peer-reviewed publication of each. The contributions are progressively and sequentially integrated: we first study clustering-based personalization in *classical* FL (Section 3.2), we then investigate the efficiency of FL fine-tuning of LMs with PEFT in an anomaly-detection application (Section 3.2) next, we combine both directions to build language-specialized multilingual models through clustering of LoRA adapters (Section 3.3); and finally, we identify and address the main limitations of existing personalized methods of the recent literature by shifting the personalization from the client to the task (Section 3.4).

3.1. Contribution 1: Hierarchical Clustering with Model Similarity

Federated Learning enables collaborative training without sharing raw data, but data heterogeneity (non-IID data) degrades the global model when diverging client updates are aggregated. This first contribution tackles inter-client heterogeneity through clustering-based personalization in classical FL, establishing the principles for the techniques later extended to LMs. We propose **FedSCCS**, a hierarchical clustering method that groups similar clients from the internal representations of their models and trains one specialized model per group, exploring the trade-offs between cluster specialization and global generalization. This work was published as “*FedSCCS: Hierarchical Clustering with Multiple Models for Federated Learning*” at the IEEE International Conference on Communications (ICC 2024) [Talasso et al. 2024].

3.1.1. Methodology

After a few standard FedAvg rounds, aggregating all clients updates into a single model, at a designated round the server measures client’s similarities from the last-layer activations using Centered Kernel Alignment (CKA)[Kornblith et al. 2019] a representation-similarity metric computed over a small, unlabeled, server-side public dataset. After that, Ward’s hierarchical clustering is then applied to the resulting similarity matrix, and the server aggregates and maintains a separate model for each cluster, as shown in Figure 1.

The hierarchical formulation makes the level of personalization adaptable: varying the number of clusters from a single global model (FedAvg, maximum generalization) to one model per client (maximum specialization). We additionally apply intra-cluster client selection to reduce communication and training cost while preserving performance.

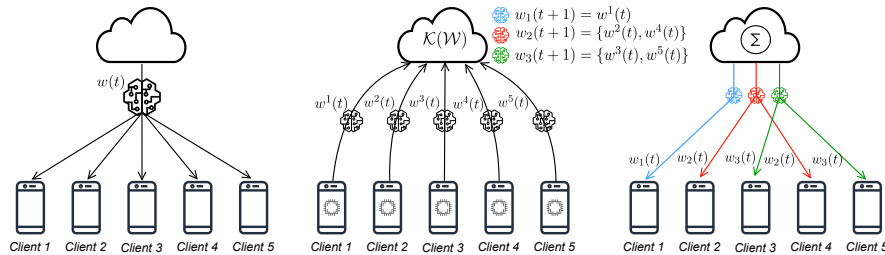


Figure 1. An overview of the FedSCCS workflow [Talasso et al. 2024]

3.1.2. Results

On the MotionSense Human Activity Recognition (HAR) dataset (24 heterogeneous clients) and with a simple Multi-Layer Perceptron, several baseline clustering strategies outperformed FedAvg by avoiding the negative interference of aggregating dissimilar clients. Combining CKA with hierarchical clustering achieved about 5% higher test accuracy than baselines using weight cosine similarity or K-Center clustering, and comparisons against random clustering confirmed the gains stem from meaningful similarity-based collaboration rather than specialization alone.

Additionally, for communication and computational efficiency, intra-cluster client selection preserved accuracy while reducing uplink communication: the Power-of-Choice [Cho et al. 2022] strategy required only 62.5% of the full-FedAvg cost and random per-cluster selection only 47.5%. The implementation was made publicly available³.

FedSCCS shows that similarity-based clustering is an effective criterion and that specialization and communication efficiency can be improved together. It also surfaced the limitations that drive the rest of the dissertation, such as the fixed number of clusters and the need to support transformer architectures in LMs and LoRA-based adaptation.

3.2. Contribution 2: Efficient Federated Fine-Tuning of Language Models

Federated fine-tuning of LMs is limited by the cost of training and transmitting large models [Ye et al. 2024b, Babakniya et al. 2023]. This contribution studies how PEFT mitigates these costs and the performance and efficiency trade-off to be further explored in the next contributions, using fully unsupervised log anomaly detection in distributed systems, which is a latency- and privacy-sensitive task. We propose a federated pipeline based on SLMs and LoRA that detects anomalies while transmitting only a small fraction of the model. This work was published as “*Efficient Fine-Tuning of Language Models to Detect Anomalies in Private Logs using Federated Learning*” at the Brazilian Symposium on Computer Networks and Distributed Systems (SBRC 2025), where it received an **Honorable Mention** [Talasso et al. 2025a].

3.2.1. Methodology

In the proposed pipeline for log anomaly detection, first, clients apply a common preprocessing procedure (windowing, message delimitation, and regular-expression sanitization that replaces some tokens, IPs, timestamps, and identifiers with placeholders) which protects privacy and prevents overfitting to specific numeric values. Each client then fine-tunes a locally stored, 4-bit-quantized SLM with LoRA, so that only the low-rank adapters (not the full model) are trained and communicated to the server for aggregation.

Detection is fully unsupervised and runs jointly with training: since the model learns to represent normal log patterns, anomalies are flagged when the next-token top- K prediction accuracy drops below a normality threshold β , removing any need for labeled anomalies, as only normal data is used, which is scarce in practice [Guo et al. 2021]. The full process are shown in Figure 2.

³<https://github.com/GabrielTalasso/FedSCCS>

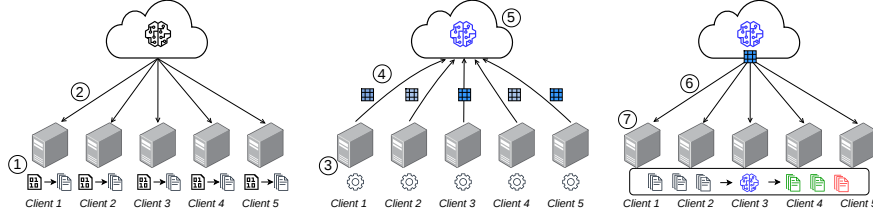


Figure 2. Illustration of the step-by-step operation of the language model tuning methodology for anomaly detection [Talasso et al. 2025a].

3.2.2. Results

Using the HDFS log dataset [Xu et al. 2009] with the SmolLM family (135M–1.7B parameters) [Allal et al. 2024] and LoRA ranks from 2 to 128, the pipeline reached F1 above 98% while communicating 10–100 $\times$ less data than full fine-tuning; using rank 2 on the 1.7B model reduced the per-round transmitted by more than 4300 $\times$ versus full fine-tuning, with only a marginal loss in performance. A top- K value of $K=1$ provided the most stable detection across configurations.

The study also focuses on the deployment trade-offs: large models with small adapters minimize communication with fewer rounds but demand substantial inference memory and compute, whereas smaller models require up to 7 $\times$ less memory and run inference up to 8 $\times$ faster, at the cost of more communication rounds. The experiments ran on a variety of hardware with different resources level to measure real values. The implementation was made publicly available⁴

This contribution establishes that PEFT makes federated fine-tuning of LMs viable under strict communication and computational budgets without sacrificing much task performance, providing the efficiency foundation on which the subsequent personalization methods build.

3.3. Contribution 3: Federated Clustering of Language Models Adapters

This contribution integrates the previous clustering and efficiency findings in statistical heterogeneity in federated fine-tuning of LMs, focusing on multilingual federations applications where aggregating updates across languages causes negative impact. We propose **FedMLLM**, which performs unsupervised clustering of LoRA adapters to group clients by model similarity and build language-specialized models, mitigating cross-domain interference while preserving collaborative knowledge sharing and privacy. This work was published as “*Leveraging Federated Learning for Multilingual and Private Language Models via Model Clustering*” at the IEEE International Conference on Federated Learning Technologies and Applications (FLTA 2025) [Talasso et al. 2025b].

3.3.1. Methodology

In the proposed method, FedMLLM, and based on the previous contributions’ observations, the base model is distributed once and kept frozen, clients train and transmit only their

⁴<https://github.com/GabrielTalasso/FL-LLM-AD>

LoRA adapters, enabling efficient on-device training and low communication overhead. On the server, adapters are compared by cosine similarity, grouped with Ward’s hierarchical clustering, and one specialized adapter is aggregated per cluster and returned to clients. The full process is shown in Figure 3.

A central choice is to compute similarity from the LoRA B matrices rather than the A matrices or their concatenation. Since B concentrates task- and language-specific information, as demonstrated in [Tian et al. 2024, Guo et al. 2024] and in the ablation experiments (whereas A is dominated by its Gaussian initialization), this yields better, more discriminative clusters at no additional communication cost. As before, the hierarchy spans the full range from joint aggregation (FedAvg) to fully local training.

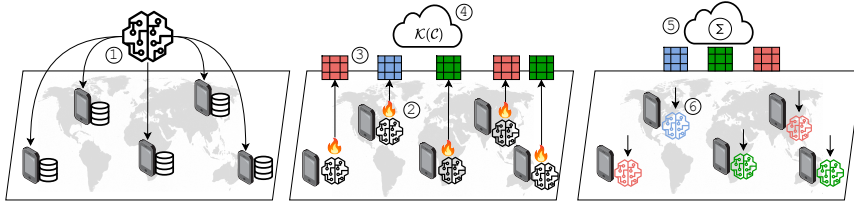


Figure 3. Illustration of the FedMLLM, training specialized adapters, creating better adapters for local usage with clustering [Talasso et al. 2025b].

3.3.2. Results

FedMLLM consistently outperformed both FedAvg and the Fully Distributed (FD) (no collaboration) baseline across languages and model sizes (LLaMA 3, 1B and 3B) on ROUGE-1. Notably, FedAvg often fell below FD, confirming that naively merging heterogeneous language updates is harmful, whereas similarity-aware clustering produced higher-quality specialized adapters. When scaling the experiment 60 clients and 15 languages, different settings (high-, mid-, and low-resource) and FedMLLM achieved the best average ROUGE-1 of 0.158, versus 0.149 (FD) and 0.098 (FedAvg), with the gap over FedAvg widening as heterogeneity increase.

Cluster-quality ablation analyses supported the method: B -matrix similarities showed clear diagonal structure grouping same-language clients, and a Separation Gap analysis confirmed stronger intra-cluster cohesion for B than for A or $A+B$. An ablation further showed that using the correct number of clusters is critical, especially for linguistically close languages such as Portuguese and Spanish. The implementation was made publicly available⁵.

FedMLLM demonstrates that unsupervised clustering of LoRA adapters is an effective and scalable strategy for federated multilingual instruction tuning, jointly increasing specialization, efficiency, and robustness to heterogeneity. It also exposed the two open challenges: (i) intra-client task interference and (ii) cross-task generalization, the target to be addressed by the final contribution.

⁵<https://github.com/GabrielTalasso/FedMLLM>

3.4. Contribution 4: Task-Centric Personalized Clustered Federated Fine-Tuning

The final contribution addresses two limitations of existing personalized federated fine-tuning methods: *intra-client task interference*, when a single client have multiple tasks with objectives that conflict during local training, and poor *generalization under test-time distribution shift*, when unseen tasks appear at inference. We propose **FedRouter**, that shifts the personalization from the client to the task by combining local and global clustering with adaptive routing at inference. Developed in collaboration with the **University of Cambridge and Flower Labs**, this work was accepted as “*FedRouter: Task-Centric Personalized Federated Fine-Tuning of Multitask Language Models*” at the 35th International Conference on Artificial Neural Networks (ICANN 2026) [Talasso et al. 2026b].

3.4.1. Methodology

The FedRouter method leverages two clustering levels: **Local clustering** split mixed tasks within each client: the frozen base model embeds the local data, a centroid-based algorithm (K-Means) clusterize it into latent tasks, and each partition trains a dedicated adapter, with a round-robin schedule bounding the per-round cost. **Global clustering** then aligns tasks across clients: the server clusters the union of local centroids and aggregates the adapters mapped to each global task, so that specialization is organized per task rather than per client. The full process is presented in Figure 4.

At inference, an **evaluation router** embeds each test sample and routes it to the nearest centroid (Euclidean distance) to select an adapter. A *local* mode uses client-specific centroids for personalized inference, while a *global* mode uses all global centroids to generalize to previously unseen tasks; samples are grouped by adapter to enable efficient batched inference.

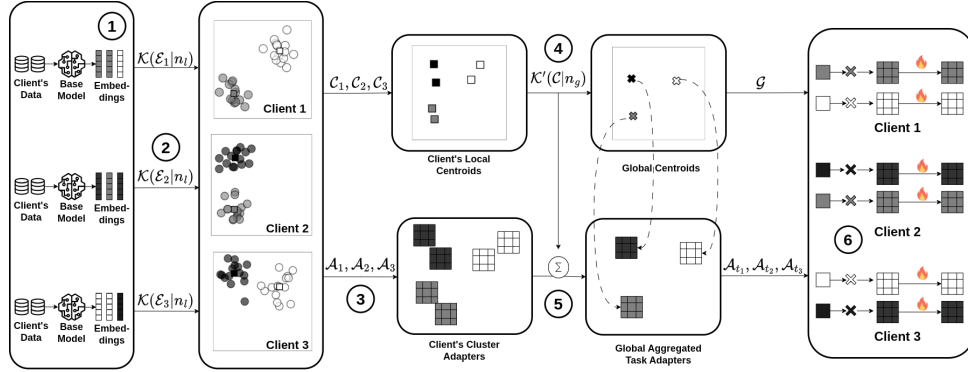


Figure 4. Proposed local and global clustering FedRouter Workflow Overview

3.4.2. Results

The evaluation setup was made across three scenarios of increasing intra-client interference: Single for one task per client, Dual for two tasks per client, and All for all tasks inside each client’s dataset. On four FLAN tasks with Llama 3.2 (1B), FedRouter matched state-of-the-art baselines (FedDPA [Long et al. 2024], FedSA [Guo et al. 2024],

FedCluster [Talasso et al. 2025b], FedIT [Zhang et al. 2024], Local) in the interference-free Single case and outperformed all of them as heterogeneity cases, reaching the best average ROUGE-1 (0.562; 0.567 for the relaxed FedRouter* variant, without resource limitation) and up to 3.5% absolute ($\approx 6.1\%$ relative) gain under interference, with local clustering accuracy of 100%/100%/95.4% in each of the three scenarios.

The advantage was largest under test-time distribution shift when training on a single task and evaluating on all tasks collapsed most personalization methods (e.g., 0.255 for Local), whereas FedRouter’s global routing reused adapters trained on other tasks to reach 0.583, up to 33.6% absolute ($\approx 136\%$ relative) improvement over client-centric personalization. Ablations further showed clean scaling with model size (1B–8B) and client count in federation, and that the Silhouette Score recovers the true number of underlying tasks both locally and globally. The implementation was made publicly available⁶.

By moving the personalization around tasks and adding adaptive inference routing, FedRouter simultaneously reduces negative interference and restores generalization, two goals usually unaddressed in personalized FL as the literature baselines, making it the main contribution of the dissertation and a robust solution for federated fine-tuning.

4. Conclusions

This dissertation tackled the two main challenges to fine-tuning LMs in federated settings: first, the statistical heterogeneity, which causes negative interference during aggregation, and the high cost of training and transmitting large models using clustering-based personalization combined with Parameter-Efficient Fine-Tuning (LoRA). Four sequential contributions were developed and integrated to solve specific gaps in the literature. Overall, the results show that integrating clustering-based personalization with parameter-efficient fine-tuning is an effective strategy for the federated fine-tuning of language models under heterogeneous data. Table 1 summarizes the contributions.

Tabela 1. Publication Summary and experimental scenarios.
Details in Appendix

Contr.	Venue	Qualis	Year	Ref.	Citations	Clusters	LM FT	Scenario
C1	IEEE ICC	A1	2024	[Talasso et al. 2024]	14	✓		HAR
C2	SBRC	A4	2025	[Talasso et al. 2025a]	4		✓	Log Anomaly Detection
C2	WIARC - SBRC	-	2026	[Talasso et al. 2026a]	0		✓	Log Anomaly Detection
C3	IEEE FLTA	-	2025	[Talasso et al. 2025b]	1	✓	✓	Multilingual Adaptation
C4	ICANN	A2	2026	[Talasso et al. 2026b]	2	✓	✓	Multitask Adaptation

⁶<https://github.com/GabrielTalasso/FedRouter>

Referências

- [Achiam et al. 2023] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- [Allal et al. 2024] Allal, L. B., Lozhkov, A., Bakouch, E., von Werra, L., and Wolf, T. (2024). Smollm - blazingly fast and remarkably powerful.
- [Arivazhagan et al. 2019] Arivazhagan, M. G., Aggarwal, V., Singh, A. K., and Choudhary, S. (2019). Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*.
- [Babakniya et al. 2023] Babakniya, S., Elkordy, A. R., Ezzeldin, Y. H., Liu, Q., Song, K.-B., El-Khamy, M., and Avestimehr, S. (2023). Slora: Federated parameter efficient fine-tuning of language models. *arXiv preprint arXiv:2308.06522*.
- [Bai et al. 2024] Bai, J., Chen, D., Qian, B., Yao, L., and Li, Y. (2024). Federated fine-tuning of large language models under heterogeneous language tasks and client resources. *arXiv e-prints*, pages arXiv–2402.
- [Braun et al. 2025] Braun, C., Jarczewski, R. O., Talasso, G. U., Villas, L. A., and de Souza, A. M. (2025). Beyond shortest path: Agentic vehicular routing with semantic context. *arXiv preprint arXiv:2511.04464*.
- [Cho et al. 2022] Cho, Y. J., Wang, J., and Joshi, G. (2022). Towards understanding biased client selection in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 10351–10375. PMLR.
- [de Souza et al.] de Souza, A. M., da Costa, J. B., Guidoni, D., Talasso, G., Maciel, F., Gonzalez, L. F., Cerqueira, E., Bittencourt, L. F., Loureiro, A. A., and Villas, L. A. Fine-tuning federado de modelos de linguagem na era da comunicação.
- [Dettmers et al. 2024] Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2024). Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.
- [Dubey et al. 2024] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- [Guo et al. 2021] Guo, H., Yuan, S., and Wu, X. (2021). Logbert: Log anomaly detection via bert. In *2021 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.
- [Guo et al. 2024] Guo, P., Zeng, S., Wang, Y., Fan, H., Wang, F., and Qu, L. (2024). Selective aggregation for low-rank adaptation in federated learning. *arXiv preprint arXiv:2410.01463*.
- [Houlsby et al. 2019] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhé, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

- [Hu et al. 2021] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- [Jarczewski et al. 2026] Jarczewski, R. O., Talasso, G. U., Villas, L., and de Souza, A. M. (2026). Agentic federated learning: The future of distributed training orchestration. *arXiv preprint arXiv:2604.04895*.
- [Kaplan et al. 2020] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- [Kornblith et al. 2019] Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR.
- [Kuang et al. 2024] Kuang, W., Qian, B., Li, Z., Chen, D., Gao, D., Pan, X., Xie, Y., Li, Y., Ding, B., and Zhou, J. (2024). Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5260–5271.
- [Li et al. 2022] Li, Q., Diao, Y., Chen, Q., and He, B. (2022). Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pages 965–978. IEEE.
- [Long et al. 2024] Long, G., Shen, T., Jiang, J., Blumenstein, M., et al. (2024). Dual-personalizing adapter for federated foundation models. *Advances in Neural Information Processing Systems*, 37:39409–39433.
- [Lu et al. 2024] Lu, Z., Li, X., Cai, D., Yi, R., Liu, F., Zhang, X., Lane, N. D., and Xu, M. (2024). Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*.
- [McMahan et al. 2017] McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.
- [Radford et al. 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- [Sun et al. 2024] Sun, Y., Li, Z., Li, Y., and Ding, B. (2024). Improving lora in privacy-preserving federated learning. *arXiv preprint arXiv:2403.12313*.
- [T Dinh et al. 2020] T Dinh, C., Tran, N., and Nguyen, J. (2020). Personalized federated learning with moreau envelopes. *Advances in neural information processing systems*, 33:21394–21405.
- [Talasso et al. 2025a] Talasso, G., de Souza, A., Guidoni, D., Cerqueira, E., and Villas, L. (2025a). Fine-tuning eficiente de modelos de linguagem para detectar anomalias em logs privados usando aprendizado federado. In *Anais do XLIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 126–139, Porto Alegre, RS, Brasil. SBC.

- [Talasso et al. 2026a] Talasso, G., Souza, A., and Villas, L. (2026a). Feddale: Detecção de anomalias eficiente em logs distribuídos com pequenos modelos de linguagem federados. In *Anais do I Workshop de Inteligência Artificial para Redes de Computadores*, pages 85–98, Porto Alegre, RS, Brasil. SBC.
- [Talasso et al. 2024] Talasso, G. U., de Souza, A. M., Bittencourt, L. F., Cerqueira, E., Loureiro, A. A. F., and Villas, L. A. (2024). Fedscs: Hierarchical clustering with multiple models for federated learning. In *ICC 2024 - IEEE International Conference on Communications*, pages 3280–3285.
- [Talasso et al. 2025b] Talasso, G. U., de Souza, A. M., Gonzalez, L. F. G., Cerqueira, E., Loureiro, A. A. F., and Villas, L. A. (2025b). Leveraging federated learning for multilingual and private language models via model clustering. In *2025 3rd International Conference on Federated Learning Technologies and Applications (FLTA)*, pages 25–32.
- [Talasso et al. 2026b] Talasso, G. U., Kurmanji, M., de Souza, A. M., Lane, N. D., and Villas, L. A. (2026b). Task-centric personalized federated fine-tuning of language models.
- [Tan et al. 2022] Tan, A. Z., Yu, H., Cui, L., and Yang, Q. (2022). Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12):9587–9603.
- [Tian et al. 2024] Tian, C., Shi, Z., Guo, Z., Li, L., and Xu, C.-Z. (2024). Hydralora: An asymmetric lora architecture for efficient fine-tuning. *Advances in Neural Information Processing Systems*, 37:9565–9584.
- [Touvron et al. 2023a] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023a). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- [Touvron et al. 2023b] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- [Villalobos et al. 2022] Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., and Hobbhahn, M. (2022). Will we run out of data? limits of llm scaling based on human-generated data. *arXiv preprint arXiv:2211.04325*.
- [Xu et al. 2009] Xu, W., Huang, L., Fox, A., Patterson, D., and Jordan, M. I. (2009). Detecting large-scale system problems by mining console logs. In *Proceedings of the ACM SIGOPS 22nd Symposium on Operating Systems Principles, SOSP '09*, page 117–132, New York, NY, USA. Association for Computing Machinery.
- [Ye et al. 2024a] Ye, R., Ge, R., Zhu, X., Chai, J., Du, Y., Liu, Y., Wang, Y., and Chen, S. (2024a). Fedllm-bench: Realistic benchmarks for federated learning of large language models. *arXiv preprint arXiv:2406.04845*.
- [Ye et al. 2024b] Ye, R., Wang, W., Chai, J., Li, D., Li, Z., Xu, Y., Du, Y., Wang, Y., and Chen, S. (2024b). Openfedllm: Training large language models on decentralized private data via federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6137–6147.

- [Zhang et al. 2024] Zhang, J., Vahidian, S., Kuo, M., Li, C., Zhang, R., Yu, T., Wang, G., and Chen, Y. (2024). Towards building the federatedgpt: Federated instruction tuning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6915–6919.
- [Zhu et al. 2021] Zhu, H., Xu, J., Liu, S., and Jin, Y. (2021). Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390.

Appendix

Non-First-Author Contributions

The following **book chapter** was also published as a result of this master’s research. The associate short-course was presented in the ”Brazilian Symposium of Computer Networks and Distributed Systems- SBRC 2025:

- Allan M. de Souza, Joahannes B. D. da Costa, Daniel Guidoni, **Gabriel Talasso**, Filipe Maciel, Luis F. G. Gonzalez, Eduardo Cerqueira, Luiz F. Bittencourt, Antonio A. F. Loureiro, Leandro A. Villas. *Fine-tuning Federado de Modelos de Linguagem na Era da Comunicação*. In **Mini-Cursos do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos**, 2025. DOI: <https://doi.org/10.5753/sbc.18862.9.2> [de Souza et al.]

Finally, this research also **contributes** directly to the following works:

- Braun, C., Jarczewski, R. O., **Talasso, G. U.**, Villas, L. A., de Souza, A. M. *Beyond Shortest Path: Agentic Vehicular Routing with Semantic Context*, 2025. In **Neurips 2025 - UrbanAI Workshop** [Braun et al. 2025].
- Jarczewski, R. O., **Talasso, G. U.**, Villas, L. A., de Souza, A. M. *Agentic Federated Learning: The Future of Distributed Training Orchestration*, 2026. In **ICLR 2026 - AIMS Workshop** [Jarczewski et al. 2026].
- Braun, Carnot; **Talasso, Gabriel U.**; Da Costa, Joahannes B. D.; Villas, Leandro A.; De Souza, Allan M. *A Hybrid Agent-based Navigation System with Semantic Reasoning*. Accepted in **8th Workshop on Urban Computing, International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)**, 2026.

Extended Version of Publication Summary Table

Tabela 2. Details of publication summary and experimental scenarios.

Contr.	Venue	Scope	Qualis	H5 Index	Year	Ref.	Citations	# Downloads	Code
C1	IEEE ICC	International	A1	76	2024	[Talasso et al. 2024]	14	724	Link
C2	SBRC	National	A4	8	2025	[Talasso et al. 2025a]	4	–	Link
C2	WIARC – SBRC	National	–	–	2026	[Talasso et al. 2026a]	0	–	Link
C3	IEEE FLTA	International	–	–	2025	[Talasso et al. 2025b]	1	78	Link
C4	ICANN	International	A2	28	2026	[Talasso et al. 2026b]	2	–	Link

Information marked as ” – ” are not available (recent conferences).

Highlights

1. **Honorable Mention Award.** The efficient federated fine-tuning method for private log anomaly detection (Contribution 2) received an *Honorable Mention* as one of the best papers in the main track of the Brazilian Symposium on Computer Networks and Distributed Systems (SBRC 2025) [Talasso et al. 2025a]
2. **International Collaboration.** FedRouter (Contribution 4), the main contribution of the dissertation, was developed in collaboration with the *University of Cambridge* and *Flower Labs*, and was accepted at the 35th International Conference on Artificial Neural Networks (ICANN 2026) [Talasso et al. 2026b].
3. **Journal Extended-Version Invitation.** FedDALE (Contribution 2), accepted at the WIARC workshop of SBRC 2026 [Talasso et al. 2026a], was invited to submit an extended version to the *Journal of Internet Services and Applications* (JISA), a Qualis A2 journal published by the Brazilian Computer Society (SBC).
4. **Fully Open-Source Implementations.** The source code of all five publication was publicly released as open-source repositories, ensuring full reproducibility: FedSCCS, FL-LLM-AD, FedMLLM, and FedRouter.