FedReFT+: Federated Representation Fine-Tuning with ALL-BUT-ME Aggregation

Anonymous ACL submission

Abstract

Parameter-efficient fine-tuning (PEFT) has attracted significant attention for adapting large pre-trained models by modifying a small subset of parameters. Recently, Representation Fine-tuning (ReFT) has emerged as an effective alternative. ReFT shifts the fine-tuning paradigm from updating model weights to directly manipulating hidden representations that capture rich semantic information, and perform better than state-of-the-art PEFTs in standalone settings. However, its application in Federated Learning (FL) remains challenging due to heterogeneity in clients' data distributions, model capacities, and computational resources. To address these challenges, we introduce **Fed**erated **Re**presentation **F**ine-**T**uning (FedReFT+), a novel approach to fine-tune the client's hidden representation. FedReFT+ applies sparse intervention layers to steer hidden representations directly, offering a lightweight and semantically rich fine-tuning alternative ideal for edge devices. However, representation-level updates are especially vulnerable to aggregation mismatch under different task heterogeneity, where naive averaging can corrupt semantic alignment. To mitigate this issue, we propose **All-But-Me (ABM)** aggregation, where each client receives the aggregated updates of others and partially incorporates them, enabling stable and personalized learning by balancing local focus with global knowledge. We evaluate FedReFT+ on commonsense reasoning, arithmetic reasoning, instruction-tuning, and GLUE, where it consistently outperforms state-of-the-art PEFT methods in FL, achieving 7×–15× higher parameter efficiency compared to leading LoRA-based approaches. The paper code is available at Anonymous Repository

Fine-tuning has emerged as a core strategy for adapting large language models (LLMs) to various downstream tasks, allowing for a broad generalization from minimal task-specific data (Ding et al., 2023; Ziegler et al., 2019). However, traditional fine-tuning is computationally expensive and memory-intensive, which poses scalability challenges. This is further amplified in resource-constrained environments, such as smartphones, where full model updates are often infeasible due to limited resources. To address these challenges, parameter-efficient fine-tuning (PEFT) methods such as Adapter Tuning (Houlsby et al., 2019), BitFit (Zaken et al., 2022), Prefix Tuning (Li and Liang, 2021), Prompt Tuning (Lester et al., 2021), and Low-Rank Adaptation (LoRA) (Hu et al., 2021a), have been proposed, significantly reducing the cost of adaptation by updating only a small subset of model weights.

PEFT has emerged as the preferred method for efficiently adapting large language models (LLMs) without sacrificing performance. However, most PEFT approaches assume centralized data access, which is unrealistic in many real-world scenarios where data is distributed across users or devices with varying tasks and privacy concerns. Federated Learning (FL) offers a solution by enabling collaborative model training without centralizing data, but prior FL work often emphasizes task-specific tuning rather than learning generalizable representations. In practice, clients frequently work on diverse or specialized tasks, making global representation learning both more difficult and more essential.

While PEFT typically modifies model weights, recent interpretability research highlights the potential of hidden representations, which encode rich semantic information. Representation Fine-Tuning (ReFT) (Wu et al., 2024b) leverages this by directly intervening in hidden layers, achieving stronger performance than methods like LoRA. Despite ReFT's success in centralized settings, it has yet to be adapted for FL, where challenges such as data heterogeneity, varying model capacities, and limited computational resources complicate aggregation and reduce effectiveness. In order to study the challenges of representation-level fine-tuning under heterogeneous federated settings, and to evaluate the effectiveness of our proposed aggregation strategy, we put forward the following research questions:

1) How can we aggregate representation-level updates in Federated Learning without compromising semantic alignment across task-heterogeneous clients?

2) Is weighted averaging sufficient for aligning semantically rich hidden representations, or is a more robust and personalized strategy needed to preserve local semantics while leveraging global knowledge?

To address these questions, we introduce **Fed**erated **Re**presentation **F**ine-**T**uning (FedReFT+), a novel framework for personalized and parameter-efficient federated representation fine-tuning. FedReFT+ builds on the core idea of ReFT by enabling clients to inject lightweight intervention components (i.e., sparse low-rank matrices W, R, b) directly into hidden representations, making it particularly useful for edge devices with limited resources. To mitigate the degradation in semantic alignment caused by naive aggregation such as vanilla weighted average aggregation (FedAvg) (McMahan et al., 2017), we propose the *All-But-Me (ABM)* aggregation strategy. Instead of averaging all client updates uniformly, ABM constructs a personalized global intervention for each client by computing the geometric median over updates from all other clients.

The key contributions of our work are as follows: **Contribution 1**: We address a critical gap in the utilization of ReFT in FL setting by introducing a novel aggregation strategy, *All-But-Me (ABM)*, specifically designed for low-rank, representation-level interventions. ABM addresses the challenge of semantic misalignment caused by naive aggregation strategies, preserving client-specific semantics while enabling stable collaboration.

Contribution 2: We propose **FedReFT+**, a novel framework for personalized and parameter-efficient federated fine-tuning based on representation-level interventions. **Contribution 3**: We evaluate the framework by simulating task heterogeneity, i.e., assigning different tasks to clients, all derived from a common dataset. This setup mimics real-world scenarios where clients pursue distinct objectives over structurally similar data, allowing us to evaluate the effectiveness of FedReFT+ and ABM under realistic conditions.

Paper outline: The remainder of the paper is organized as follows. Section 1 formally defines the problem setting, the motivation behind our work and the challenges posed by heterogeneous FL setting. Section 2 details our methodology, including the FedReFT+ mechanism and the All-But-Me (ABM) aggregation strategy. Section 3 presents our experimental results and evaluates the effectiveness of our approach on multiple benchmarks. Section 4 concludes the paper with key insights and future directions. We defer additional details to the appendices.

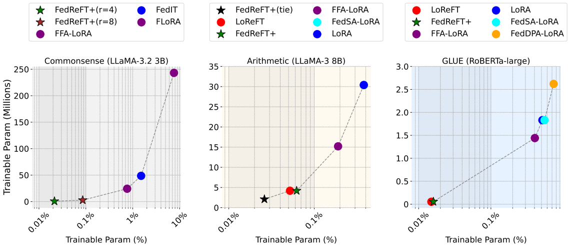

Figure 1: Illustration of the relationship between the number of trainable parameters (in millions and %) for various federated PEFT methods on Commonsense, Arithmetic, and GLUE benchmarks using LLaMA-3.2B, LLaMA-3 8B, and RoBERTa-large models, respectively. FedReFT+ achieves competitive or state-of-the-art performance while training significantly fewer parameters, resulting in improved communication efficiency and reduced transmission cost in federated learning settings.

1 Problem Formulation and Motivation

In this section, we present the motivation to apply ReFT in FL and formalize the problem of adapting personalized representation. Although ReFT offers parameter-efficient updates in the representation space, its application in FL faces key challenges, including task heterogeneity, semantic misalignment, and unstable aggregation among heterogeneous clients. We highlight these challenges and formulate the objective of enabling parameter-efficient and semantically aligned adaptation in FL using ReFT. **Challenge 1: LoReFT in FL Settings**: ReFT (Wu et al., 2024b) offers an attractive alternative by modifying hidden activations instead of model weights. By intervening directly in structured semantic subspaces, ReFT supports interpretable, modular, and task-aligned adaptation, particularly advantageous in task-heterogeneous FL settings. However, full ReFT incurs considerable communication costs and poses integration challenges when clients use different model capacities or architectures.

To bridge this gap, we adopt **Low-Rank Linear Subspace ReFT (LoReFT)**(Wu et al., 2024b), a lightweight ReFT variant that constrains interventions to a learnable low-rank subspace. This design significantly reduces overhead while maintaining semantic control, making it a promising candidate for FL. We follow the LoReFT intervention formulation from (Wu et al., 2024b) on hidden representations $h \in \mathbb{R}^d$ which is defined as:

$$\Phi_{\text{LoReFT}}(h) = h + R^\top(Wh + b - Rh), \quad (1)$$

where, $W \in \mathbb{R}^{r\times d}$ is a low-rank projection matrix with d as the representation dimension and r as the subspace intervention dimension, $R \in \mathbb{R}^{r\times d}$ is a low-rank projection matrix with orthonormal rows, and $b \in \mathbb{R}^r$, with $r \ll d$. This structure, inspired by Distributed Interchange Intervention (DII) (Geiger et al., 2024), enables semantically grounded, low-rank adaptation suitable for scalable and privacy-preserving FL. Despite its efficiency, applying LoReFT in FL raises several non-trivial challenges: LoReFT modifies internal representations that are sensitive to client-specific data distributions. Aggregating these interventions naïvely using FedAvg can cause semantic interference or collapse. Without global synchronization, low-rank updates may evolve in divergent directions, especially when tasks are dissimilar. Applying shared LoReFT interventions across clients risks overfitting to shared patterns while ignoring local semantics. Considering all these challenges, the major research question is:

Can representation-level adaptation via LoReFT achieve personalization and stability in federated environments without collapsing under task and data heterogeneity?

FedReFT+ uses All-But-Me(ABM) aggregation to robustly combine intervention parameters while preserving personalization in heterogeneous FL.

Challenge 2: Federated Fine-Tuning under Task Heterogeneity: A central motivation of our work is to address task heterogeneity in real-world FL, where clients perform fundamentally different tasks rather than optimizing a shared objective. For example, clients may work on distinct reasoning tasks within natural language QA that demand different semantic skills. While centralized fine-tuning has proven effective for such tasks, it assumes access to all data, which is unrealistic in decentralized settings. In FL, each client sees only a local, task-specific subset of the broader reasoning space, leading to highly heterogeneous training distributions, a common challenge in multi-department or cross-domain deployments. This raises the question:

How can we learn a global representation that generalizes across tasks when each client trains only on a fragment of the broader task distribution?

Standard methods like FedAvg (McMahan et al., 2017) struggle in this regime, as they average semantically misaligned updates, often resulting in degraded performance or collapsed representations. Formally, let each client i have a dataset $\mathcal{T}_i = \{X_i, Y_i\}$ and optimize a personalized model $\boldsymbol{\theta}_i$ by solving:

$$\min_{\boldsymbol{\Theta}} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(X_i, Y_i, \boldsymbol{\theta}_i), \quad (2)$$

where $\mathcal{L}$ is the task-specific loss and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_i\}_{i=1}^N$ is the set of client-specific models. Our proposed method, FedReFT+, can successfully address this research challenge. FedReFT+ enables scalable, personalized representation learning across heterogeneous tasks, allowing global reasoning capabilities to emerge from decentralized, task-specific updates.

Table 1: Comparison of different aggregation methods of GLUE task on ROBERTa, Commonsense and Arithmetic reasoning task on LLaMA-2 7B.

Task	**FedAvg (%)**	**FedReFT+ (%)**
Commonsense	70.16	**70.77**
Arithmetic	15.36	**17.21**
GLUE	88.17	**89.77**

Challenge 3: Learnable Parameter Sharing with the Server: When applying ReFT (Wu et al., 2024b) in a FL setting from the perspective of learnable parameter sharing, a fundamental question is: *Which of these parameters should be communicated to the server for collaborative aggregation?* In FedReFT+, each client fine-tunes hidden representations by introducing learnable low-rank intervention parameters W, R, and a bias b into a frozen backbone model. Sharing only some part of the intervention parameters leads to incomplete information transfer and breaks the low-rank structure critical for generalization. W projects representations into a low-dimensional space, and R reconstructs them; omitting either disrupts compositionality and limits alignment across heterogeneous clients. In Table 2, empirical results show that partial sharing

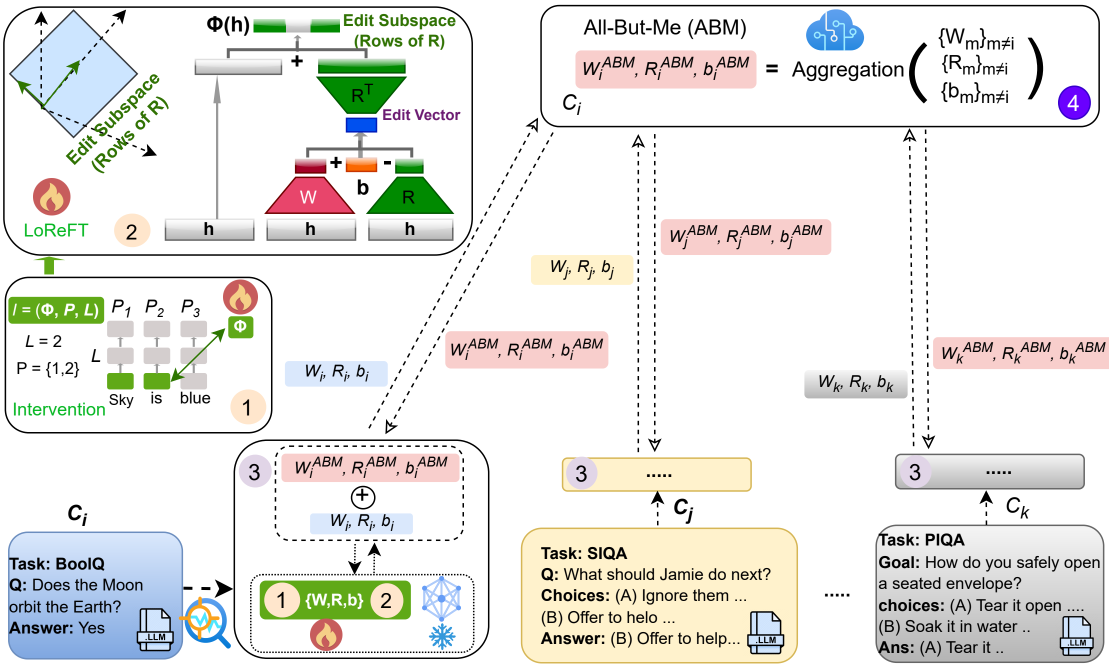
Φ(h)
Edit Subspace (Rows of R)
+
R^T
Edit Vector
+
-
W
b
R
h
h
h
Edit Subspace (Rows of R)
LoReFT
2
I = (Φ, P, L)
P1 P2 P3
L = 2
P = {1,2}
L
Φ
Sky is blue
Intervention
1
All-But-Me (ABM)
W_i^ABM, R_i^ABM, b_i^ABM = Aggregation ({W_m}m≠i {R_m}m≠i {b_m}m≠i)
C_i
4
W_j, R_j, b_j
W_j^ABM, R_j^ABM, b_j^ABM
W_i^ABM, R_i^ABM, b_i^ABM
W_i, R_i, b_i
W_k, R_k, b_k
W_k^ABM, R_k^ABM, b_k^ABM
3
W_i^ABM, R_i^ABM, b_i^ABM
⊕
W_i, R_i, b_i
1 {W,R,b} 2
3 .....
3 .....
C_i
C_j
C_k
Task: BoolQ
Q: Does the Moon orbit the Earth?
Answer: Yes
LLM
Task: SIQA
Q: What should Jamie do next?
Choices: (A) Ignore them ...
(B) Offer to helo ...
Answer: (B) Offer to help...
LLM
.....
Task: PIQA
Goal: How do you safely open a seated envelope?
choices: (A) Tear it open ....
(B) Soak it in water ..
Ans: (A) Tear it ..
LLM

Figure 2: **FedReFT+ with ABM Aggregation.** Clients cross-task demonstrate personalization while maintaining alignment with the global representation. (1)-(2): Each client applies LoReFT(Wu et al., 2024b) interventions to train learnable parameter $\{W, R, b\}$ to modify hidden representations h in a low-rank edit subspace. (3): Clients fine-tune $\{W, R, b\}$ locally and partially fuse received *All-But-Me* aggregated updates with their own. (4): The server performs ABM aggregation using the geometric median over other clients' intervention parameters to generate $W_k^{\text{ABM}}, R_k^{\text{ABM}}, b_k^{\text{ABM}}$.

significantly degrades performance and representation alignment under task heterogeneity. Therefore, FedReFT+ shares the full set of learnable intervention parameters (W, R, B) from each client with the server.

2 Methodology

In this section, we introduce FedReF+, designed to address the challenges we discussed in the previous section. An illustrative overview of FedReFT+ is shown in Figure 2.

2.1 Intervention Parameter Sharing Strategies

To reduce communication overhead while maintaining personalization, we explore three strategies for sharing local intervention parameters with the server. These strategies are summarized in Table 2 and represent different trade-offs between expressiveness and communication efficiency:

1) Full Intervention Sharing: $\{\mathbf{W} \in \mathbb{R}^{r\times d}, \mathbf{R} \in \mathbb{R}^{r\times d}, \mathbf{b} \in \mathbb{R}^{r}\}$ This strategy shares the complete set of intervention parameters, capturing client-specific compression ($\mathbf{W}$), transformation ($\mathbf{R}$), and translation ($\mathbf{b}$). It enables the most accurate reconstruction of local updates and yields the best global performance, especially under high heterogeneity.

2) No Bias Sharing: $\{\mathbf{W} \in \mathbb{R}^{r\times d}, \mathbf{R} \in \mathbb{R}^{r\times d}\}$ This variant omits the bias term $\mathbf{b}$ but retains the directional transformation via $\mathbf{W}$ and $\mathbf{R}$. While it allows the server to align low-rank subspace transformations across clients, it lacks the ability to model per-dimension translation shifts, which can hinder fine-grained personalization. **3) No W Sharing:** $\{\mathbf{R} \in \mathbb{R}^{r\times d}, \mathbf{b} \in \mathbb{R}^{r}\}$ This configuration excludes $\mathbf{W}$, giving the server access only to the reconstruction and shift parameters. Without knowledge of how the local signals were encoded, the server's ability to interpret or align updates is severely limited. The $\{\mathbf{W}, \mathbf{R}, \mathbf{b}\}$ strategy provides the highest fidelity for aggregation, $\{\mathbf{W}, \mathbf{R}\}$ offers a balanced compromise, and $\{\mathbf{R}, \mathbf{b}\}$ prioritizes communication efficiency at the cost of semantic alignment and global performance.

Table 2: Performance vs. parameter efficiency for different LoReFT sharing strategies (Uplink) for C clients on commonsense reasoning task following the second experiment design. GLUE task on ROBERTa, Arithmetic and Commonsense on LLaMa-2 7B model.

Task	Strategy	TP(% ↓)	Score
GLUE	`W,R,b`	0.01384	94.31
	`W,R`	0.01383	64.03
	`R,b`	0.00693	74.12
Arithmetic	`W,R,b`	0.03114	29.01
	`W,R`	0.03114	26.13
	`R,b`	0.01557	25.77
Commonsense	`W,R,b`	0.03114	73.82
	`W,R`	0.03114	70.63
	`R,b`	0.01557	68.02

2.2 Intervention Design for Federated Classification Tasks

Following the formulation in ReFT (Wu et al., 2024b), for a given client, we define the classification head H_ψ with parameters $\psi = \{W_o, b_o, W_d, b_d\}$ operates on the CLS token representation $z \in \mathbb{R}^d$ from the final layer:

$$H_\psi(z) = \text{softmax}\left(W_o \cdot \tanh(W_d z + b_d) + b_o\right). \tag{3}$$

We jointly optimize the intervention parameters ϕ and the classifier ψ using cross-entropy loss over input x and label y:

$$\min_{\phi,\psi} \left\{-\log H_\psi\left(y \mid z_\phi(x)\right)\right\}. \tag{4}$$

2.3 All-But-Me (ABM) Aggregation

In heterogeneous FL, the integration of shared knowledge without compromising local task-specific adaptation remains a core challenge. Standard aggregation methods such as FedAvg (McMahan et al., 2017), which averages client models into a single global model, are often suboptimal in non i.i.d. scenarios. They risk overwriting valuable client-specific representations and rely on fixed mixing weights that may further reduce personalization. Table 1 depicts the comparison. To overcome these limitations, we propose the *All-But-Me (ABM)* aggregation strategy. Instead of initializing clients with a global model, each client continues to update its local parameters while partially incorporating knowledge aggregated from other clients. Specifically, each client k receives a robustly aggregated set of intervention parameters $\{\mathbf{W}_k^{\text{ABM}}, \mathbf{R}_k^{\text{ABM}}, \mathbf{B}_k^{\text{ABM}}\}$, calculated from the updates of all other clients using a geometric median:

$$\begin{aligned} \mathbf{W}_k^{\text{ABM}} &= \text{ABM}\left(\left\{\mathbf{W}_m^{\text{local}}\right\}_{m\neq k}\right), \\ \mathbf{R}_k^{\text{ABM}} &= \text{ABM}\left(\left\{\mathbf{R}_m^{\text{local}}\right\}_{m\neq k}\right), \\ \mathbf{B}_k^{\text{ABM}} &= \text{ABM}\left(\left\{\mathbf{B}_m^{\text{local}}\right\}_{m\neq k}\right). \end{aligned} \tag{5}$$

The client then performs a personalized update by interpolating between its local parameters and the ABM-aggregated ones using a mixing factor $\alpha \in [0, 1]$. We have discussed in detail how the evaluation loss-based α tuning works in the appendix D.

$$\begin{aligned} \mathbf{X}_k^{\text{new}} &= (1-\alpha)\cdot \mathbf{X}_k^{\text{local}} + \alpha \cdot \mathbf{X}_k^{\text{ABM}}, \\ \mathbf{X} &\in \{\mathbf{W}, \mathbf{R}, \mathbf{B}\}. \end{aligned} \tag{6}$$

Before the next local training using the $\mathbf{R}_k^{\text{new}}$, we do the orthogonal transformation of $\mathbf{R}_k^{\text{new}}$ to keep the original property of R.

ABM via Geometric Median. The geometric median (also known as the spatial or L_1 median) offers a robust alternative to the arithmetic mean, particularly under client heterogeneity and adversarial conditions (Maronna and Martin, 2006; Weiszfeld, 1937). Given a set of vectors $\mathcal{S} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, it is defined as:

$$x^* = \arg\min_{x\in\mathbb{R}^d} \sum_{i=1}^{n} \|x - x_i\|_2, \tag{7}$$

which minimizes the sum of Euclidean distances to all elements in the set. This estimator is robust to outliers and misaligned updates, making it well-suited for federated settings. We instantiate the ABM function using the geometric median, where each client k receives an aggregated intervention vector computed from $\mathcal{S}_k = \{x_m\}_{m\neq k}$:

$$\text{ABM}(\mathcal{S}_k) = \arg\min_{x\in\mathbb{R}^d} \sum_{x_m\in\mathcal{S}_k} \|x - x_m\|_2. \tag{8}$$

To solve this optimization efficiently, we employ Weiszfeld's algorithm (Weiszfeld, 1937), an iterative method known to converge under mild conditions. Details of the algorithm are provided in Appendix E. By avoiding direct averaging and incorporating semantically meaningful low-rank intervention updates through robust aggregation, ABM enables each client to benefit from the knowledge of others without sacrificing local personalization. This approach enhances stability and generalization across non-i.i.d. and task-heterogeneous FL environments.

Table 3: Federated fine-tuning performance of LlaMa-3.2 3B across five commonsense reasoning tasks with MT experimental setup where clients train on heterogeneous task mixtures to promote generalizable representations. *Performance results of all baseline methods and the experimental setup are taken from (Singhal et al., 2025).

Method	R	# TP(M) ↓	TP(%)	BoolQ	PIQA	SIQA	HellaS.	WinoG	Avg ↑
FLoRA*	32	243.15	7.58	65.05	82.81	74.67	81.84	76.01	78.83
FedIT*	32	48.63	1.51	62.99	81.50	73.13	76.83	71.51	75.74
FFA-LoRA*	32	24.31	0.76	62.87	80.03	68.53	70.02	65.56	71.11
Fed-SB *	120	2.83	0.0884	64.86	81.66	74.87	81.67	75.22	75.66
FedReFT+ (Ours)	4	1.38	0.0428	63.09	82.10	72.36	90.27	69.22	75.41
	8	2.75	0.0857	64.01	81.18	72.11	90.71	71.01	75.66
	16	6.194	0.1927	63.42	81.61	73.64	91.23	71.35	76.05
	32	11.01	0.3427	64.53	81.34	73.39	91.51	71.32	76.22
FedReFT+ (tie ϕ, Ours)	4	0.688	0.0214	49.94	81.23	72.72	89.84	68.43	72.43
	8	1.38	0.0428	57.15	81.22	72.77	90.56	68.50	74.04

Table 4: Performance comparison across arithmetic reasoning tasks with the Distict Task and Mixed Task setup.

FedReFT+	Distinct Task (DT)				Mixed Task (MT)			
Models	AQuA	GSM8K	SVAMP	Avg ↑	AQuA	GSM8K	SVAMP	Avg ↑
LLaMa 7B	25.59	25.47	49.80	33.62	22.83	14.33	27.10	21.42
LLaMa-2 7B	29.53	32.45	57.3	39.76	21.65	20.39	31.5	24.51
LLaMa-3 8B	34.64	48.98	73.60	52.41	31.89	48.90	70.04	50.48

3 Experimental Validation

To evaluate FedReFT+, we conduct extensive experiments on three different NLP benchmarks covering over 12 datasets. Our objective is to present a comprehensive assessment of how this approach performs in various NLP tasks. We experiment with both masked and autoregressive language models, including RoBERTa-large (Liu et al., 2019), TinyLlama-1B (Community, 2023), LLaMA 7B (Touvron et al., 2023a), LLaMA-2 7B and 13B (Touvron et al., 2023b), LLaMA-3.2B and LLaMA-3 8B (AI, 2024), across multiple settings and scales. Our comparisons include state-of-the-art baselines, such as LoRA(Hu et al., 2021b), FedIT (Zhang et al., 2024), FFA-LoRA(Sun et al., 2024), FedDPA-LoRA(Long et al., 2024), FedSA-LoRA(Guo et al., 2024), Fed-SB (Singhal et al., 2025) and FLoRA(Wang et al., 2024) focusing on both parameter efficiency and performance tradeoffs. We align the experimental setup configurations with the baseline papers to ensure fair comparisons. To optimize memory usage, we load all base language models with `torch.bfloat16` precision. All experiments are executed on a single NVIDIA A100-SXM4-80GB GPU, except for LLaMA-2 13B, which is run on a GPUH200x8 141GB system to accommodate the computational demands of large-scale federated fine-tuning. The results are averaged over two runs to report the mean performance.

Hyperparameter Configuration. In the experiments, we determine how many interventions to learn, as well as which input positions and layers to apply them to. We apply interventions at a fixed number of layers L, and at prefix (p) and suffix (s) positions in the input prompt. We narrow the hyperparameter search space for Federated Learning by adopting the configuration used in the centralized ReFT(Wu et al., 2024b) paper. The appendix B provides a brief overview of the hyperparameter search space. We experiment with whether to share (tie) the intervention parameters ϕ across different input positions within the same layer. Given the positions $P = \{1, \ldots, p\} \cup \{n - s + 1, \ldots, n\}$, we define the untied and tied variants (Wu et al., 2024b):

$$\begin{aligned} \mathbf{I}_{\text{untied}} &= \{\langle \Phi, \{p\}, l \rangle \mid p \in P,\, l \in L\}\,, \\ \mathbf{I}_{\text{tied}} &= \{\langle \Phi, P, l \rangle \mid l \in L\}\,. \end{aligned} \tag{9}$$

Task Distribution Rationale. We design two experimental setups to study how global representations converge under diverse task distributions. In the Mixed-Task (MT) setup, each client trains on a subset of a combined reasoning dataset but is evaluated on a single task, encouraging generalized, transferable representations through ABM aggregation. This reflects collaborative learning across varied yet related tasks. In the Distinct Task (DT) setup, each client trains on a unique reasoning task, enabling personalized fine-tuning while still leveraging global updates. Despite higher task heterogeneity, this setup maintains stable performance as model capacity increases. Both setups show that FedReFT+ supports effective generalization in MT and robustness in DT.

3.1 Commonsense Reasoning

We evaluate global representation generation on eight commonsense reasoning tasks using the Commonsense170K dataset inspired by (Singhal et al., 2025; Wu et al., 2024b). We use the same hyperparameter of (Singhal et al., 2025) and tune the intervention parameter in the Appendix B.1. This helps us tune important hyperparameters efficiently and also test their robustness across multiple commonsense reasoning tasks.

Datasets. For the first setup (MT design), we split the combined COMMONSENSE170K (Hu et al., 2023) commonsense reasoning tasks among clients and use them for fine-tuning. Each client evaluates one of the commonsense reasoning tasks. BoolQ (Clark et al., 2019), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), HellaSwag (Zellers et al., 2019), and WinoGrande (Sakaguchi et al., 2021). For the second setup (DT design), each client fine-tunes on only one of these five commonsense reasoning tasks and is evaluated using the same task. All examples are formatted as multiple-choice questions, requiring the model to directly generate the correct answer without providing rationales. We adopt the prompt template from Hu et al. (Hu et al., 2023) with minor modifications, including additional string normalization by removing leading and trailing whitespace.

Results. In Table 3, our proposed FedReFT+ method demonstrates strong parameter efficiency while maintaining competitive accuracy across five commonsense reasoning tasks. Notably, FedReFT+ with rank 8 uses only 2.75M(0.0857%) trainable parameters, achieving accuracy close to or better than several baselines. Compared to existing methods our approach reduces the trainable parameter count by factors of 9× to 89×, with minimal to no compromise in performance. Figure 1 also depicts so. The experiments results on MT setup in Commonsense reasoning are shown in Appendix Table 14.

3.2 Arithmetic Reasoning

For the arithmetic reasoning tasks, we design three experimental settings to fine-tune models on various arithmetic reasoning tasks. We follow the same hyperparameter tuning strategy as used in COMMONSENSE170K in Appendix B.1, which uses a development set to select the best-performing configuration. Evaluation is based solely on the final numeric or multiple-choice answer, disregarding intermediate reasoning steps.

Datasets. In the first setting following MT, we split a combined arithmetic reasoning dataset, MATH10K (Hu et al., 2023) which includes four arithmetic reasoning tasks with chain-of-thought solutions generated by a language model. Each client reports performance using test set one of three tasks: AQuA (Ling et al., 2017), GSM8K (Cobbe et al., 2021), and SVAMP (Patel et al., 2021). In the second setting following DT, each client is assigned one arithmetic reasoning task for both fine-tuning and evaluation. Both MT and DT setup results are reported in Table 4. In the third setting following (Guo et al., 2024; Kuang et al., 2024), we split the dataset GSM8K into three clients under an IID distribution, and the results are shown in Table 6. All optimization hyperparameters remain consistent with that setup.

Results. In Table 6, FedReFT+ demonstrates strong performance while using significantly fewer trainable parameters than existing baselines. Notably, FedReFT+ achieves the highest accuracy among all methods, while FedReFT+ (tie ϕ) offers a compelling trade-off between performance and efficiency. These results highlight the scalability and efficiency of our representation-tuning approach. The Distinct Task (DT) setup represents task heterogeneity. Hence, fine-tuning allows clients to learn highly personalized, task-specific representations while benefiting from global aggregation. As a result, the DT setup yields higher performance, as seen in Table 4. In contrast, the MT setup, where clients train on heterogeneous task mixtures to promote global generalizable representations. This blending of tasks during fine-tuning leads to general representation learning but can de-

Table 5: Performance comparison across GLUE Tasks on RoBERTa model for $C = 3$, FedReFT+ use rank rank 1. *Performance results of all baseline methods are taken from (Guo et al., 2024) and use LoRA rank 8.

Setup	Method	# TP(M) ↓	TP(%)	MNLI-m	SST-2	QNLI	QQP	Avg ↑
Standalone	FT	355	100	88.8	96.0	93.8	91.5	91.87
	LoRA*	1.83	0.515	88.71	95.16	91.16	85.33	89.33
	LoReFT	0.053	0.015	89.2	96.2	94.1	88.5	92.0
FL	FFA-LoRA*	1.44	0.405	88.83	94.95	91.52	86.71	89.39
	FedDPA-LoRA*	2.62	0.737	88.99	95.50	90.74	85.73	89.47
	FedSA-LoRA*	1.83	0.551	90.18	96.00	92.13	87.48	90.43
	FedReFT+ (ours)	**0.053**	**0.015**	88.86	95.17	94.52	86.57	**90.93**

Table 6: Performance comparison on arithmetic reasoning tasks for GSM8K on LLaMa-3 8B model with LoRA rank 8, where clients enable consistent evaluation of representation generalization. *Performance results of all baseline methods are taken from (Guo et al., 2024).

Method	# TP(M) ↓	TP(%)	GSM8K
LoReFT	4.19	0.052	48.33
LoRA*	30.40	0.38	46.23
FedSA-LoRA*	30.40	0.38	46.63
FFA-LoRA*	15.2	0.19	46.32
FedReFT+ (tie ϕ)	**2.09**	**0.0261**	**49.35**
FedReFT+	**4.19**	**0.0622**	**49.68**

grade performance on specific evaluation tasks due to misaligned representation and conflicting task objectives.

3.3 Natural Language Understanding

We evaluate the effectiveness of FedReFT+ in learning generalizable representations for Natural Language Understanding (NLU) using the GLUE benchmark (Wang et al., 2018). The objective is to fine-tune NLU to learn global representations that capture task-level semantics. By aligning intermediate representations for downstream classification performance. This setup allows us to test whether lightweight intervention tuning can align representations across clients within a single NLU task.

Hyperparameter Tuning. We tune hyperparameters separately for each task, following common practice for PEFT methods (Hu et al., 2023) in FL. To reduce the impact of random seed variability, we perform hyperparameter tuning using a fixed seed and report the average performance across that seed and two additional unseen seeds. Details are provided in the Appendix B.2.

Results. Table 5 depicts that our approach performs strongly across GLUE tasks while using very few trainable parameters. It performs competitive or outperforms other methods, showing that it can learn good representations even in a federated setting. Despite using over 30× fewer parameters than some baselines, it still achieves competitive results, making it both efficient and effective.

3.4 Ablation Studies

We conduct ablation studies to better understand the effectiveness of FedReFT+, specifically examining the role of geometric median-based All-But-Me aggregation. Details of this analysis are provided in Appendix F.

4 Conclusion

In this work, we addressed a critical gap in the deployment of Representation Fine-Tuning within Federated Learning settings by proposing a novel aggregation strategy tailored to its low-rank, representation-level interventions. While ReFT improves upon traditional PEFT methods like LoRA by operating on semantically rich hidden representations, its application in FL is limited by data heterogeneity, model diversity, and the shortcomings of standard aggregation methods. To address these, we propose All-But-Me aggregation, enabling clients to adapt their local ReFT parameters using a robust average of others' interventions. **FedReFT+** ensures parameter efficiency and semantic alignment. Extensive experiments under task heterogeneity and different heterogeneous settings show that **ABM** consistently enhances convergence, generalization, and robustness, making it a practical and effective solution for personalized representation learning in federated systems.

Limitations

Due to computational constraints, our current study focuses primarily on LoReFT-based interventions within language models under a fixed set of hyperparameters. In future work, we aim to automate the parameter search space using a multi-agent coordination framework to better explore optimal low-rank configurations for each client. Although our current set-up does not explicitly address privacy, we are actively investigating how to integrate differential privacy mechanisms, such as DP-SGD, into the FedReFT framework without sacrificing personalization. Initial experiments in this direction are ongoing. Additionally, we are exploring the theoretical properties of ABM aggregation under adversarial or noisy clients, and whether it can be extended to other modalities beyond language, such as vision-language models in federated systems.

Data and Model Usage

We use publicly available models including LLaMA-1.1B, LLaMA-2 (7B, 13B), LLaMA-3 8B, LLaMA-3.2 3B and RoBERTa-large. LLaMA-2 and LLaMA-3 models are licensed under Meta's community license permitting commercial use. RoBERTa-large is under the MIT License, and TinyLLaMA use Apache 2.0, while the original LLaMA-1 7B is for non-commercial research only. We will release code and configurations under an open-source license with usage documentation to support reproducibility and responsible use.

We employ publicly available datasets across commonsense and arithmetic reasoning tasks, each released under open-source licenses. For commonsense reasoning, BoolQ is under CC BY-SA 3.0, PIQA under Apache 2.0, SIQA and WinoGrande under CC BY 4.0, HellaSwag under MIT, ARC under CC BY-SA 4.0, and OBQA under CC BY 4.0. For arithmetic reasoning, AddSub, AQuA, MAWPS, and MultiArith are under Apache 2.0, GSM8K and SVAMP under MIT, and SingleEq under CC BY 4.0. For natural language understanding, GLUE consists of multiple datasets, each with its own license, allowing for research use and redistribution.

Environmental Impact

Our approach FedReFT+ achieves 7×–15× higher parameter efficiency than existing PEFT methods, using fewer trainable parameters. This reduces energy consumption and training time, making our method more resource-efficient and environmentally friendly.

Societal Impacts

Our method FedReFT+ adapts ReFT for Federated Learning, enabling efficient model personalization with minimal computational overhead. This promotes broader accessibility of large language models on edge devices, including in low-resource or privacy-sensitive environments. While improving inclusivity and deployment scalability, care must be taken to mitigate potential misuse or bias propagation across decentralized systems.

Bias and Fairness

Our approach FedReFT+ considers the potential for bias introduced by non-IID client data in Federated Learning. While we do not explicitly optimize for fairness, we acknowledge that imbalanced participation or data diversity may lead to uneven model performance. Future work should explore fairness-aware objectives to mitigate such disparities across clients and demographic groups.

Responsible Deployment

To support responsible use, we include clear documentation outlining the intended use cases of our framework and advise against applying it in safety-critical settings without thorough validation. We encourage users to follow ethical standards, such as the ACL Code of Ethics, when deploying our method. Our released code comes with usage instructions to promote safe adoption and reduce the risk of misuse. This work is licensed under CC BY 4.0, allowing reuse and adaptation, even commercially, with proper attribution.

AI Assistants in Research Writing

We used AI assistants to support writing and code refinement during the preparation of this paper. All AI-generated content was reviewed and verified by the authors.

References

Meta AI. 2024. Llama 3: Meta's next-generation large language model. https://ai.meta.com/llama/.

Matan Avitan, Ryan Cotterell, Yoav Goldberg, and Shauli Ravfogel. 2024. What changed? converting representational interventions to natural language. *arXiv preprint arXiv:2402.11355*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Dongqi Cai, Yaozong Wu, Shangguang Wang, Felix Xiaozhu Lin, and Mengwei Xu. 2023. Efficient federated learning for modern nlp. In *Proceedings of the Annual International Conference on Mobile Computing and Networking*, pages 1–16.

Jinyu Chen, Wenchao Xu, Song Guo, Junxiao Wang, Jie Zhang, and Haozhao Wang. 2022. Fedtune: A deep dive into efficient federated fine-tuning with pre-trained transformers. *arXiv preprint arXiv:2211.08025*.

Lili Chen, Lijun Su, and Jinhui Xu. 2017. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *arXiv preprint arXiv:1705.10301*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2924–2936.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

TinyLlama Community. 2023. Tinyllama: An open reproduction of llama-1b with 1.1b parameters. `https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0`.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, and 1 others. 2023. Parameter-efficient fine-tuning of large-scale pretrained language models. *Nature Machine Intelligence*, 5(3):220–235.

Shangqian Gao, Ting Hua, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. 2024. Adaptive rank selections for low-rank approximation of language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 227–241.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pages 160–187. PMLR.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. 2024. Selective aggregation for low-rank adaptation in federated learning. *arXiv preprint arXiv:2410.01463*.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *ICML*.

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

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

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.

Weirui Kuang, Bingchen Qian, Zitao Li, Daoyuan Chen, Dawei Gao, Xuchen Pan, Yuexiang Xie, Yaliang Li, Bolin Ding, and Jingren Zhou. 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.

Kenneth Lange. 2016. *MM optimization algorithms*. SIAM.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *EMNLP*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *KR*.

Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. 2024a. Mixlora: Enhancing large language models fine-tuning with lora based mixture of experts. *arXiv preprint arXiv:2404.15159*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024b. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*.

Vladislav Lialin, Sherin Muckatira, Namrata Shivagunde, and Anna Rumshisky. 2023. Relora: High-rank training through low-rank updates. In *The Twelfth International Conference on Learning Representations*.

Wang Ling, Dani Yogatama, Chris Dyer, and Philip Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 158–167.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2023. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zefang Liu and Jiahua Luo. 2024. Adamole: Fine-tuning large language models with adaptive mixture of low-rank adaptation experts. *arXiv preprint arXiv:2405.00361*.

Guodong Long, Tao Shen, Jing Jiang, Michael Blumenstein, and 1 others. 2024. Dual-personalizing adapter for federated foundation models. *Advances in Neural Information Processing Systems*, 37:39409–39433.

Qikai Lu, Di Niu, Mohammadamin Samadi Khoshkho, and Baochun Li. 2024. Hyperflora: Federated learning with instantaneous personalization. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pages 824–832.

Ricardo A Maronna and Douglas Martin. 2006. Yohai. robust statistics. *Wiley Series in Probability and Statistics. John Wiley and Sons*, 2:3.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

J. Pablo Muñoz, Jinjie Yuan, and Nilesh Jain. 2025. Low-rank adapters meet neural architecture search for llm compression. In *AAAI'25 workshop on CoLoRAI - Connecting Low-Rank Representations in AI*.

Yahao Pang, Xingyuan Wu, Xiaojin Zhang, Wei Chen, and Hai Jin. 2025. Fedeat: A robustness optimization framework for federated llms. *arXiv preprint arXiv:2502.11863*.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*.

Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. 2022. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154.

Hossein Rajabzadeh, Mojtaba Valipour, Tianshu Zhu, Marzieh Tahaei, Hyock Ju Kwon, Ali Ghodsi, Boxing Chen, and Mehdi Rezagholizadeh. 2024. Qdylora: Quantized dynamic low-rank adaptation for efficient large language model tuning. *arXiv preprint arXiv:2402.10462*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. 2024. Mimic: Minimally modified counterfactuals in the representation space. *arXiv preprint arXiv:2402.09631*.

Raghav Singhal, Kaustubh Ponkshe, Rohit Vartak, Lav R Varshney, and Praneeth Vepakomma. 2025. Fed-sb: A silver bullet for extreme communication efficiency and performance in (private) federated lora fine-tuning. *arXiv preprint arXiv:2502.15436*.

Guangyu Sun, Matias Mendieta, Taojiannan Yang, and Chen Chen. 2022. Exploring parameter-efficient fine-tuning for improving communication efficiency in federated learning. In *International Conference on Learning Representations (ICLR)*.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. 2024. Improving lora in privacy-preserving federated learning. *arXiv preprint arXiv:2403.12313*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Somya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Van-Tuan Tran, Quoc-Viet Pham, and 1 others. 2025. Revisiting sparse mixture of experts for resource-adaptive federated fine-tuning foundation models. In

ICLR 2025 Workshop on Modularity for Collaborative, Decentralized, and Continual Deep Learning.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. 2022. Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. *arXiv preprint arXiv:2210.07558*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. 2022. Adamix: Mixture-of-adaptations for parameter-efficient model tuning. *arXiv preprint arXiv:2205.12410*.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. 2024. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *arXiv preprint arXiv:2409.05976*.

Endre Weiszfeld. 1937. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal, First Series*, 43:355–386.

Xun Wu, Shaohan Huang, and Furu Wei. 2024a. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2024b. Reft: Representation fine-tuning for language models. *arXiv preprint*. ArXiv:2404.03592.

Yifan Yang, Kai Zhen, Ershad Banijamal, Athanasios Mouchtaris, and Zheng Zhang. 2024. Adazeta: Adaptive zeroth-order tensor-train adaption for memory-efficient large language models fine-tuning. *arXiv preprint arXiv:2406.18060*.

Liping Yi, Han Yu, Gang Wang, and Xiaoguang Liu. 2023. Fedlora: Model-heterogeneous personalized federated learning with lora tuning. *arXiv preprint arXiv:2310.13283*.

Dong Yin, Yudong Chen, Ravi Kannan, Peter L. Bartlett, and Kannan Ramchandran. 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

Ben Zaken, Yoav Goldberg, and Amir Globerson. 2022. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *ACL Findings*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. 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. IEEE.

Zhuo Zhang, Yuanhang Yang, Yong Dai, Qifan Wang, Yue Yu, Lizhen Qu, and Zenglin Xu. 2023. Fedpetuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models. In *Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 9963–9977.

Changhai Zhou, Shijie Han, Shiyang Zhang, Shichao Weng, Zekai Liu, and Cheng Jin. 2024. Rankadaptor: Hierarchical dynamic low-rank adaptation for structural pruned llms. *arXiv preprint arXiv:2406.15734*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Related Works

A.1 Parameter-Efficient Fine-Tuning (PEFT)

Fine-tuning LLMs is resource-intensive due to their large parameter counts. Parameter-efficient fine-tuning (PEFT) methods mitigate this by updating only a small subset of parameters while keeping pre-trained weights frozen (Li and Liang, 2021; He et al., 2021; Wang et al., 2022). Several PEFT approaches have been proposed, Adapter Tuning (Houlsby et al., 2019), BitFit (Zaken et al., 2022), Prefix Tuning (Li and Liang, 2021), Prompt Tuning (Lester et al., 2021), and Low-Rank Adaptation (LoRA) (Hu et al., 2021a). Among them, LoRA is widely adopted for its efficiency in approximating weight updates via low-rank matrices. Extensions such as ReLoRA (Lialin et al., 2023) and RankAdapter (Zhou et al., 2024) improve memory use and adapt ranks dynamically, though they lack theoretical guarantees. AdaZeta (Yang et al., 2024) introduces zeroth-order optimization with convergence guarantees, while others (Gao et al., 2024; Rajabzadeh et al., 2024; Valipour et al., 2022) explore adaptive ranks without formal proofs. LoRA has been integrated with Mixture-of-Experts models (Li et al., 2024a; Wu et al., 2024a), as in AdaMoLE (Liu and Luo, 2024), to enable dynamic expert selection, and also with Neural Architecture Search for LLM compression (Muñoz et al., 2025). These approaches primarily target weight updates, overlooking direct interventions in hidden representations, which are discussed next.

A.2 Representation Fine-Tuning (ReFT)

ReFT shifts fine-tuning from model weights to hidden representations, leveraging their semantic structure for efficient adaptation (Wu et al., 2024b). Inspired by activation steering and representation engineering (Avitan et al., 2024; Li et al., 2024b; Liu et al., 2023; Singh et al., 2024), ReFT enables task-specific control through fixed or learned interventions without updating the full model. Notably, Inference-Time Intervention (ITI) (Li et al., 2024b) improves LLM truthfulness by modifying activations, while representation engineering (Zou et al., 2023) combines representation reading and control for interpretable model behavior. Minimally Modified Counterfactuals (MMC) (Singh et al., 2024) unify erasure and steering to reduce bias, and can be mapped to natural language edits (Avitan et al., 2024), enhancing interpretability. These findings support direct representation manipulation as a lightweight and effective alternative to weight-based PEFT methods like LoRA.

A.3 Federated Fine-Tuning

Federated Learning (FL) (McMahan et al., 2017) poses challenges for fine-tuning LLMs, including data heterogeneity, communication constraints, and model diversity. PEFT methods have emerged to address these issues efficiently (Sun et al., 2022; Chen et al., 2022; Zhang et al., 2023). LoRA-based approaches such as FedLoRA (Yi et al., 2023), Hyper-FloRA (Lu et al., 2024), and Efficient FL Adapter (Cai et al., 2023) offer modular and personalized adaptation across clients. Recent advances further incorporate privacy (FFA-LoRA (Sun et al., 2024)), heterogeneous adaptation (FloRA (Wang et al., 2024)), instruction tuning (FedIT (Zhang et al., 2024)), and expert routing (DualFed (Long et al., 2024), Sparse-FedMoE (Tran et al., 2025)). In contrast, our proposed FEDREFT+ shifts from weight updates to direct representation-level tuning via sparse intervention layers and introduces an All-But-Me (ABM) aggregation strategy to preserve semantic alignment while enabling robust knowledge sharing across non-IID clients.

A.4 Aggregation Methods in FL

To address the inherent heterogeneity and robustness challenges in federated learning, median-based aggregation strategies have been extensively studied as alternatives to simple averaging. Unlike the arithmetic mean, the geometric and coordinate-wise medians are significantly more resilient to outliers and adversarial updates, making them suitable for secure and personalized FL scenarios. For instance, coordinate-wise median aggregation has been proposed to defend against Byzantine clients in distributed optimization (Blanchard et al., 2017). This was extended with geometric median-based gradient descent to improve statistical guarantees across diverse loss landscapes (Yin et al., 2018). Further work demonstrated that coordinate-wise median and trimmed-mean-based methods achieve order-optimal convergence not only for strongly convex losses but also under non-strongly convex and even non-convex population losses (Chen et al., 2017). Additionally, a one-round median-based algorithm was shown to maintain statistical optimality under quadratic convexity, offering a communication-efficient solution (Chen et al., 2017). RFA (Pillutla et al., 2022) maintains privacy and demonstrates improved robustness over standard averaging techniques, particularly in environments with high levels of data corruption. FedEAT (Pang et al., 2025) integrates adversarial training in the embedding space with geometric median-based aggregation to enhance robustness while preserving performance. This work demonstrates that LoRA-based FL systems can effectively leverage geometric median aggregation. Inspired by these findings, we adopt geometric median aggregation in our FL framework to aggregate the All-But-Me (ABM) intervention parameter, weight $\mathbf{W}$, rotation $\mathbf{R}$, and bias $\mathbf{b}$. This provides stability across diverse client behaviors and loss geometries, improving personalization performance under data and objective heterogeneity.

B Hyperparameter Search Space

B.1 Hyperparameter Search Space for Commonsense and Arithmetic Reasoning

Following the ReFT framework (Wu et al., 2024b), we construct a development set using the GSM8K dataset and consider only the last 300 samples. We trained the clients using LLaMa 7B model with the remaining training data and determined the best-performing hyperparameters based on the model's performance on the development set. We further use this hyperparameter in another model directly. We set the maximum input sequence length to 512 tokens during training and tuning, and limit inference to 32 generated tokens. We use the same setup for commonsense reasoning with COMMONSENSE170k dataset. The hyperparameter search space is summarized in Tables 7 and 8.

During inference, we use greedy decoding (without sampling) for the commonsense reasoning benchmark, as it is a multi-token classification task. For arithmetic reasoning, we follow the decoding setup from (Hu et al., 2023), using a higher temperature of 0.3. This change helps avoid errors in HuggingFace's decoding caused by unstable probabilities

B.2 Hyperparameter Search Space for GLUE Benchmark

We perform hyperparameter (HP) tuning on RoBERTa-large separately for each GLUE task, selecting the optimal settings based on validation performance using a fixed random seed of 42. Final evaluations are conducted using two additional unseen seeds, {43, 44}, to ensure robustness. Table 9 depicts this.

Table 7: Narrow down the hyperparameter(HP) search space of LLaMA 7B models with FedReFT+ on the GSM8K development set, inspired from (Wu et al., 2024b). The best-performing settings are underlined. We apply greedy decoding without sampling during hyperparameter tuning.

HP	**FedReFT+**
prefix+suffix, $p+s$	{p5+s5, p7+s7, p9+s9}
Tied weight ϕ	{True, False}
Rank r	{8, 16, 32, 64}
Layer L	{all}
Dropout	{0.00, 0.05}
Optimizer	AdamW
LR	$\{6, 9\}\times10^{-4}$
Weight decay	$\{0, 1\times10^{-3}, 2\times10^{-3}\}$
LR scheduler	Linear
Batch size	{16, 32}
Warmup ratio	{0.06, 0.10}
Clients	{3, 5}
Epochs	{3, 4, 5, 6}
Rounds	10

C Dataset Description

C.1 Commonsense Reasoning

We train and evaluate our models on eight commonsense reasoning datasets spanning different types of open-ended QA tasks, following (Hu et al., 2021a), we construct all examples. Table 10 shows the dataset samples.

- **BoolQ** (Clark et al., 2019): A yes/no question answering dataset consisting of naturally occurring questions. We remove the associated passages to ensure a fair comparison.
- **PIQA** (Bisk et al., 2020): A dataset for physical commonsense reasoning. The model must select the more plausible solution to everyday physical tasks.
- **SIQA** (Sap et al., 2019): Focuses on social interaction reasoning by asking the model to choose responses based on human intent and consequences.
- **HellaSwag** (Zellers et al., 2019): Requires choosing the most coherent sentence completion given a context, often involving physical or temporal common sense.

Table 8: Narrow down the hyperparameter (HP) search space of LLaMA 7B models with FedReFT+ on the COMMONSENSE170k development set, following the Appendix B.1. The best-performing settings are underlined. We apply greedy decoding without sampling during hyperparameter tuning.

HP	**FedReFT+**
prefix+suffix, $p+s$	{p5+s5, p7+s7}
Tied weight p, s	{True, False}
Rank r	{8, 16, 32, 64}
Layer L	{all}
Dropout	{0.00, 0.05}
Optimizer	AdamW
LR	$\{4, 6, 9\}\times10^{-4}$
Weight decay	{0}
LR scheduler	Linear
Batch size	{16, 32}
Warmup ratio	{0.1}
Clients	{3, 5}
Epochs	{2, 3, 4}
Rounds	10

Table 9: Hyperparameter(HP) settings of RoBERTa-large models on selected GLUE tasks for FedReFT+, inspired from (Wu et al., 2024b)

HP	**MNLI**	**SST-2**	**QNLI**	**QQP**
position p	$p1$	$p3$	$p11$	$p11$
Tied weight	False			
Rank r	1			
Layer L	all			
Dropout	0.05			
Optimizer	AdamW			
LR	2×10^{-2}			
Weight decay	0.00			
LR scheduler	Linear			
Batch size	32			
Warmup ratio	0.00	0.10	0.10	0.06
Epochs	10			
Rounds	50			

- **WinoGrande** (Sakaguchi et al., 2021): Inspired by the Winograd Schema Challenge (Levesque et al., 2012), this dataset contains fill-in-the-blank problems with binary choices requiring commonsense coreference reasoning.

We follow the experimental setup in (Hu et al., 2021a) by fine-tuning our models on a combined training corpus referred to as **COMMONSENSE170K**, which merges all of the above datasets. Evaluation is conducted individually on each dataset's test split.

Table 10: Examples from commonsense reasoning tasks: BoolQ(Clark et al., 2019), PIQA(Bisk et al., 2020), HellaSwag(Zellers et al., 2019), and SIQA(Sap et al., 2019). Each instruction is followed by the answer selected during evaluation.

Dataset	**Instruction / Question**	**Answer**
BoolQ	*Please answer the following question with true or false:* Question: Do Iran and Afghanistan speak the same language?	True
PIQA	*Please choose the correct solution to the question:* Question: When boiling butter, when it's ready, you can Solution1: Pour it onto a plate Solution2: Pour it into a jar	Solution2
HellaSwag	*Please choose the correct ending to complete the given sentence:* Removing ice from car: Then, the man writes over the snow covering the window of a car, and a woman wearing winter clothes smiles. then Ending1: , the man adds wax to the windshield and cuts it. Ending2: , a person boards a ski lift... Ending3: , the man puts on a christmas coat... Ending4: , the man continues removing the snow on his car.	Ending4
SIQA	*Please choose the correct answer to the question:* Cameron decided to have a barbecue and gathered her friends together. How would others feel as a result? Answer1: like attending Answer2: like staying home Answer3: a good friend to have	Answer1

Table 11: Examples from math reasoning tasks: AQUA (Ling et al., 2017), GSM8K (Cobbe et al., 2021), and SVAMP (Patel et al., 2021). Each instruction is followed by the correct answer derived through step-by-step reasoning.

Dataset	Instruction / Question	Answer
AQUA	*Solve the following word problem:* A car is driven in a straight line toward the base of a vertical tower. It takes 10 minutes for the angle of elevation to change from 45° to 60°. After how much more time will the car reach the base of the tower? Answer Choices: (A) $5(\sqrt{3}+1)$, (B) $6(\sqrt{3}+\sqrt{2})$, (C) $7(\sqrt{3}-1)$, (D) $8(\sqrt{3}-2)$, (E) None of these.	(A)
GSM8K	*Solve the following question:* Janet's ducks lay 16 eggs per day. She eats 3 eggs and uses 4 for baking. She sells the rest at $2 per egg. How much money does she make daily?	$18
SVAMP	*Solve the following arithmetic question:* Each pack of DVDs costs $76. A discount of $25 is applied. What is the final price per pack?	$51

C.2 Arithmetic Reasoning

We evaluate arithmetic reasoning using seven benchmark datasets that cover a range of math word problem types. As in (Hu et al., 2021a), we construct all examples without using golden or retrieved passages. Data samples are shows in Table 11.

- **AQuA** (Ling et al., 2017): Presents algebraic word problems in a multiple-choice format.
- **GSM8K** (Cobbe et al., 2021): A widely used benchmark of grade-school math problems requiring multi-step reasoning.
- **SVAMP** (Patel et al., 2021): A more challenging dataset that tests robustness to paraphrased and structurally altered word problems.

Following (Hu et al., 2021a), we train our models on a combined training set named **MATH10K**.

C.3 Natural Language Understanding

For NLU, we evaluate on the GLUE benchmark following the evaluation protocol in (Wu et al., 2024b). Data samples for shown in Table 12.

- The validation set is split into two subsets one for in-training evaluation and the other for final testing.
- For large datasets (QQP, MNLI, QNLI), 1,000 samples are used for in-training validation.
- For smaller datasets, half of the validation set is used during training.

Table 12: Examples from GLUE benchmark (Wang et al., 2018) tasks: MNLI, SST-2, QNLI, and QQP.

Dataset	Instruction / Question	Answer
MNLI	Premise: The dog is running through the field. Hypothesis: An animal is moving. Label: entailment	Entailment
SST-2	Sentence: A touching and thought-provoking piece of cinema. Label: positive	Positive
QNLI	Question: What is the capital of France? Sentence: Paris is the capital and most populous city of France. Label: entailment	Entailment
QQP	Question1: How do I learn to play guitar? Question2: What is the best way to learn guitar? Label: duplicate	Duplicate

D Evaluation Loss-Based α Tuning

To personalize the blending of local and aggregated intervention parameters in FedReFT+, we employ an *evaluation loss-based* tuning strategy for the mixing coefficient $\alpha \in [0, 1]$. This coefficient governs how much each client integrates the aggregated *All-But-Me (ABM)* intervention parameters with its own local updates:

$$\theta_k^{\text{new}} = (1 - \alpha) \cdot \theta_k^{\text{local}} + \alpha \cdot \theta_k^{\text{ABM}}, \qquad (10)$$

where θ_k can represent the LoReFT intervention components W, R, and b for client k.

Tuning Procedure:

1. For a set of candidate α values ($\alpha \in \{0.0, 0.1, \ldots, 1.0\}$), the client computes interpolated parameters.

2. For each α, the client evaluates its performance using a local validation set and records the evaluation loss $\mathcal{L}(\alpha)$.

3. The optimal mixing coefficient is selected by:

$$\alpha^* = \arg\min_{\alpha} \mathcal{L}(\alpha). \qquad (11)$$

E Theoretical Foundation: Geometric Median via Weiszfeld's Algorithm

The geometric median offers a robust alternative to the arithmetic mean, particularly suitable for federated settings with heterogeneous or noisy client updates. For a given set of vectors $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \subset \mathbb{R}^d$, the geometric median $\mathbf{y}^*$ is defined as:

$$\mathbf{y}^* = \arg\min_{\mathbf{y}\in\mathbb{R}^d} \sum_{i=1}^{n} \|\mathbf{y} - \mathbf{x}_i\|_2. \qquad (12)$$

This optimization is non-smooth and convex, and generally lacks a closed-form solution. However, Weiszfeld's algorithm (Weiszfeld, 1937) provides an efficient iterative method to approximate $\mathbf{y}^*$. We now derive and justify this algorithm via the Majorization-Minimization (MM) framework.

We define the cost function to be minimized:

$$f(\mathbf{y}) = \sum_{i=1}^{n} \|\mathbf{y} - \mathbf{x}_i\|_2. \qquad (13)$$

This function is convex but non-differentiable at points where $\mathbf{y} = \mathbf{x}_i$. Weiszfeld's algorithm avoids such points during updates by construction.

The MM algorithm minimizes a difficult objective $f(\mathbf{y})$ by iteratively minimizing a surrogate function $Q(\mathbf{y}|\mathbf{y}^{(k)})$ that: Majorizes f: $Q(\mathbf{y}|\mathbf{y}^{(k)}) \geq f(\mathbf{y})$ for all $\mathbf{y}$, Touches f at the current iterate: $Q(\mathbf{y}^{(k)}|\mathbf{y}^{(k)}) = f(\mathbf{y}^{(k)})$.

We define the surrogate using Jensen's inequality and the convexity of the norm:

$$Q(\mathbf{y}|\mathbf{y}^{(k)}) = \sum_{i=1}^{n} \frac{\|\mathbf{y} - \mathbf{x}_i\|_2^2}{2\|\mathbf{y}^{(k)} - \mathbf{x}_i\|_2} + C(\mathbf{y}^{(k)}), \qquad (14)$$

where $C(\mathbf{y}^{(k)})$ is a constant that does not depend on $\mathbf{y}$. This function is differentiable and strictly convex in $\mathbf{y}$.

To find the minimizer of $Q(\mathbf{y}|\mathbf{y}^{(k)})$, we take the gradient and set it to zero:

$$\nabla Q(\mathbf{y}) = \sum_{i=1}^{n} \frac{\mathbf{y} - \mathbf{x}_i}{\|\mathbf{y}^{(k)} - \mathbf{x}_i\|_2} = 0. \qquad (15)$$

Solving the above yields the Weiszfeld update rule:

$$\mathbf{y}^{(k+1)} = \frac{\sum_{i=1}^{n} \frac{\mathbf{x}_i}{\|\mathbf{y}^{(k)} - \mathbf{x}_i\|_2}}{\sum_{i=1}^{n} \frac{1}{\|\mathbf{y}^{(k)} - \mathbf{x}_i\|_2}}. \qquad (16)$$

The update is only valid when $\mathbf{y}^{(k)} \neq \mathbf{x}_i$ for all i, a condition that can be enforced by initialization and step-size dampening if needed.

From MM theory (Lange, 2016), each iteration satisfies:

$$\begin{aligned} f(\mathbf{y}^{(k+1)}) &\leq Q(\mathbf{y}^{(k+1)}|\mathbf{y}^{(k)}) \\ &\leq Q(\mathbf{y}^{(k)}|\mathbf{y}^{(k)}) = f(\mathbf{y}^{(k)}), \end{aligned} \qquad (17)$$

ensuring that $f(\mathbf{y}^{(k)})$ is non-increasing. Under mild conditions (excluding cases where $\mathbf{y}^{(k)} = \mathbf{x}_i$), Weiszfeld's algorithm converges to the geometric median $\mathbf{y}^*$.

E.1 Application in FedReFT+: ABM Aggregation

In our FedReFT+ framework, each client receives an All-But-Me (ABM)aggregated update for intervention parameters computed as the geometric median of the corresponding parameters from all other clients. For client k, the ABM aggregated parameter is:

$$\mathbf{W}_k^{\text{ABM}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \sum_{m\neq k} \|\mathbf{w} - \mathbf{W}_m^{\text{local}}\|_2. \qquad (18)$$

We compute this using Weiszfeld's algorithm for each parameter type independently, ensuring robustness to outlier clients and misaligned updates. This enables stable and personalized aggregation without sacrificing task-specific semantics.

Weiszfeld's algorithm provides a theoretically grounded and computationally efficient way to compute the geometric median, making it ideal for ABM aggregation in heterogeneous FL. By leveraging this algorithm in FedReFT+, we ensure robustness in aggregation and improve both convergence and personalization in non-i.i.d. federated environments.

F Ablation Study

F.1 Comparison of Aggregation method on different task

We use only 20% of the training data from the COMMONSENSE170K dataset, split among three clients, and evaluate the models using the SIQA task. Figure 3 shows that Geometric Median ABM aggregation outperforms all other approaches. Similarly, we split 50% of the MATH10K dataset among five clients, train each client for only five local epochs, and evaluate the results using the GSM8K evaluation set. Additionally, we use RTK GLUE for the natural language understanding (NLU) task.

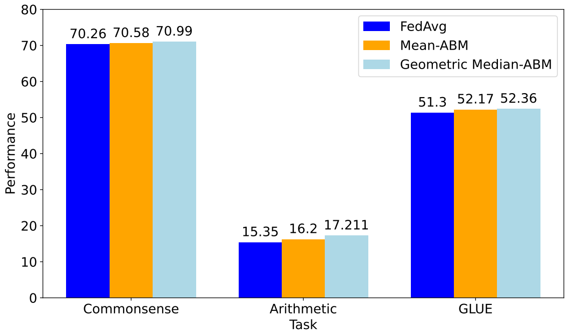

Figure 3: **Comparison of Aggregation Strategies Across Tasks.** Performance of FedAvg, Mean-ABM, and Geometric Median-ABM on three benchmark task groups: Commonsense Reasoning, Arithmetic Reasoning, and GLUE. Geometric Median-ABM consistently outperforms FedAvg and Mean-ABM, highlighting its robustness and effectiveness in heterogeneous federated settings.

Table 13: Trainable Intervention Parameters across Models (in Millions) in FedReFT+

Model	**Total P(M)**	**# TP(M)**	**TP%**
LLaMa-1.1B	1100.05	0.72	0.0655
LLaMA 7B	6,738.42	2.10	0.0311
LLaMA-2 7B	6,738.42	2.10	0.0311
LLaMA-3 8B	8,030.27	2.10	0.0261
LlaMa-2-13B	13,015.86	6.55	0.0503
RoBERTa Large	355.36	0.0492	0.0138

G Communication Efficiency

Table 14: We vary LLaMA model sizes with $C = 3$ clients following the DT design, alongside a centralized LoReFT baseline. As model capacity increases, we observe notable performance gains, with the largest model approaching the accuracy of the centralized setting. First four experiments on the Standalone centralize setup and later four experiments on the FL setup.

Method	**BoolQ**	**PIQA**	**HellaS.**
LLaMa 7b	69.30	84.4	93.1
LLaMa-2 7B	71.10	83.8	94.3
LLaMa-3 7B	75.1	90.2	96.3
Tiny LLaMa 1B	63.83	49.18	46.03
LLaMa 7B	65.84	77.75	67.64
LLaMa-2 7B	68.93	74.81	77.73
LLaMa-3 7B	72.60	85.85	89.85

As shown in Table 13, FedReFT+ is communication and computationally efficient as it uses only a very small percentage of trainable parameters (TP) compared to the total model parameters. For example, in LLaMA-7B and LLaMA-2 7B, only 0.0311% of the total parameters are trained. In RoBERTa Large, this number is even smaller, at just 0.0138%. Even for large models like LLaMA-2-13B, the trainable portion remains as low as 0.0503%. This shows that FedReFT+ is highly parameter-efficient. Despite using such a small fraction of parameters, FedReFT+ still achieves strong performance, as discussed in the experimental analysis section 3. This highlights the benefit of using FedReFT+ in resource-constrained or communication-limited federated learning settings.

G.1 Additional Experimental Validation

In this section, we also conducted some additional experiments to show the robustness of FedReFT+ in different setups. Appendix Table 14 depicts these.