GMFL: Efficient Global Masking for Federated LLM Fine-tuning

Xin Huang, Yan Hu, Yue-Jiao Gong, Xinglin Zhang*
School of Computer Science and Engineering, South China University of Technology
202421044809@mail.scut.edu.cn, huyan515601@163.com
gongyuejiao@gmail.com, zhxlinse@gmail.com

Abstract

Low-Rank Adaptation (LoRA) has emerged as a prominent solution to mitigate the communication and computation costs in federated fine-tuning of Large Language Models (LLMs). However, we observe that even within low-rank adapters, a substantial portion of parameters manifest negligible updates during federated training, leading to redundant communication and wasted local computation. To address this, we propose **GMFL**, a **plug-and-play** layer freezing mechanism designed to **seamlessly integrate** with existing federated fine-tuning frameworks. Specifically, the server monitors the global update magnitude of each LoRA layer to dynamically generate freezing masks. These masks are updated periodically with a fixed freezing rate, ensuring stable convergence by robustly identifying "saturated" layers. Theoretical analysis confirms the convergence of GMFL, where the freezing mechanism yields a bounded error that scales with client heterogeneity. Extensive experiments across multiple tasks (GLUE, Commonsense Reasoning, Math Reasoning and General Generation) demonstrate that GMFL reduces communication overhead and lowers computational costs while preserving the performance of the underlying federated fine-tuning methods. Our work provides a practical, versatile solution for deploying large-scale federated LLM fine-tuning in resource-constrained environments. Our code is available at: https://github.com/tunx-cyber/GMFL.

1 Introduction

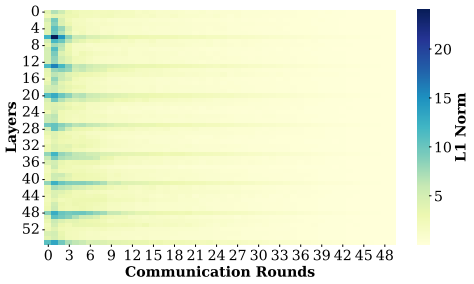
Large Language Models (LLMs), such as GPT-4 (OpenAI et al., 2024) and DeepSeek-R1 (DeepSeek-AI et al., 2025), have demonstrated remarkable capabilities across a wide range of tasks. Yet, adapting these general-purpose models to specialized downstream applications necessitates fine-tuning on high-quality, domain-specific data. In practice, such high-value data—ranging from medical records (Jung et al., 2025) to industrial codebases (Kumar and Chimalakonda, 2024)—are inherently fragmented across isolated data silos. Furthermore, centralizing this sensitive information is often infeasible due to strict privacy regulations and confidentiality concerns. To resolve the tension between the need for model adaptation and the imperative of privacy protection, federated fine-tuning has emerged as a promising solution, enabling the collaborative optimization of LLMs without exposing raw data.

Implementing full-parameter federated fine-tuning remains practically challenging due to the prohibitive communication bandwidth and local computational burdens imposed by the massive scale of LLMs (Wei et al., 2025). Prior efforts to mitigate these overheads, such as FedBAT (Li et al., 2024c), FedMRN (Li et al., 2024b), and FedMUD (Li et al., 2025), have made significant strides by employing gradient compression and low-rank decomposition. However, these approaches primarily target full-parameter training scenarios. To surmount these resource constraints, Parameter-Efficient Fine-Tuning (PEFT) techniques have been extensively incorporated into federated frameworks. Amidst a plethora of strategies—ranging from prompt-based (Lester et al., 2021) and selection-based (Liu et al., 2021; Li et al., 2024a; Pan et al., 2024) to adapter-based (Hu et al., 2021; Houlsby et al., 2019) methods—Low-Rank Adaptation (LoRA) (Hu et al., 2021) has distinguished itself as a dominant approach, owing to its superior trade-off between parameter efficiency and model performance.

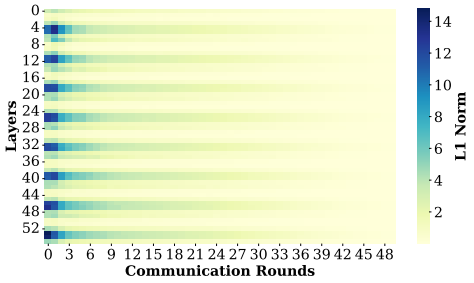
Despite the widespread adoption of LoRA in federated learning settings (FL-LoRA), our empirical analysis reveals a pronounced redundancy in the update dynamics. As illustrated in Figure 1, a substantial portion of LoRA layers manifest negligible variations throughout the training process,

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25293–25312



(a) Global updates of matrices A.



(b) Global updates of matrices B.

Figure 1: Global LoRA updates with MiniMind2-Small (Gong, 2024) on Alpaca-GPT4 dataset (Peng et al., 2023).

suggesting that continuously optimizing the full adapter yields diminishing returns. This insight motivates us to explore whether "saturated" layers can be strategically deactivated without sacrificing accuracy. To this end, we present **GMFL**, a **theoretically grounded, plug-and-play module** featuring a Global-update-Magnitude-based layer freezing mechanism for FL. By dynamically identifying and freezing inactive layers at periodic intervals, GMFL demonstrates **superiority in both practice and theory**: practically, it substantially lowers communication overhead and computational costs while maintaining model accuracy; theoretically, it is proven to reduce aggregation variance and ensure bounded loss deviation, thereby guaranteeing robust convergence when integrated into existing federated algorithms.

The main contributions of this paper are as follows:

- We propose **GMFL**, a **plug-and-play** module that reduces communication and computation by periodically freezing "saturated" layers based on global update magnitudes.
- We confirm the convergence of GMFL. Specifically, the proposed freezing mechanism is proven to produce a bounded error, and the overall convergence bound is a function of client heterogeneity.
- Extensive experiments across GLUE, Commonsense Reasoning, Math Reasoning and General Generation show that GMFL reduces communication overhead and lowers computational cost while preserving the performance of state-of-the-art FL-LoRAs.

2 Related Work

2.1 Parameter-Efficient Fine-Tuning (PEFT)

PEFT addresses the computational prohibitive nature of full-parameter fine-tuning by updating only a small subset of parameters. Existing methods largely fall into three categories: prompt-based, adapter-based, and selection-based approaches. While prompt-based methods(Lester et al., 2021) leverage learnable tokens, Adapter-based methods have gained dominance, with LoRA (Hu et al., 2021) becoming the de facto standard by employing low-rank decomposition to approximate weight updates. Recent variants like QLoRA (Dettmers et al., 2023), AdaLoRA (Zhang et al., 2023)and DoRA (Liu et al., 2024) further enhance its efficiency. Selection-based methods, such as AutoFreeze (Liu et al., 2021) and LiSA (Pan et al., 2024), focus on identifying and freezing converged layers to accelerate training. However, these methods typically operate on full-parameter settings or rely on local heuristics.

In our work, GMFL uniquely bridges these paradigms by applying a dynamic, global-magnitude-driven selection mechanism specifically to LoRA adapters, maximizing efficiency without compromising the collaborative learning capacity.

2.2 Federated Learning with LoRA

Integrating PEFT with Federated Learning (FL) has emerged as a promising solution for privacy-preserving LLM training (Wei et al., 2025). Early works like FedIT (Zhang et al., 2024) directly apply FedAvg to LoRA parameters. To mitigate the noise caused by inexact aggregation arising from the independent averaging of A and B matrices, methods like FLoRA (Wang et al., 2024) and FedEx-LoRA (Singhal et al., 2025) introduce stacked aggregation schemes or residual corrections, yet often at the cost of increased communication or memory overhead. Recently, sparse federated tuning has gained traction to further reduce overhead. Approaches such as FFA-LoRA (Sun et al., 2024) and

FedSA-LoRA (Guo et al., 2025) statically freeze specific matrices (e.g., fixing A or only transmitting A), while LoRA-A^2 (Koo et al., 2025) employs an alternating freezing strategy. However, these FL-LoRAs often rely on static rules, random selection (Park et al., 2025), or local signals, ignoring the global convergence status of different layers.

In contrast, GMFL serves as a plug-and-play freezing mechanism compatible with diverse federated fine-tuning frameworks. Unlike methods tied to specific architectures or rigid static rules, GMFL employs a data-driven approach based on global update magnitudes. This strategy effectively balances efficiency and model performance without requiring complex modifications to the underlying training protocol.

3 Motivation and Challenge

Motivation. The paradigm of adapting pre-trained LLMs to downstream tasks relies on the premise that the base model already possesses strong generalization capabilities. Consequently, fine-tuning typically entails only minimal parameter shifts rather than extensive re-learning. This raises a fundamental question: *Do specific network layers exhibit redundancy by undergoing negligible updates during fine-tuning?*

To investigate this in the context of FL, we tracked the layer-wise update magnitudes (L_1 norm) of LoRA adapters (Matrices A and B) throughout the training process. As visualized in Figure 1, the heatmaps reveal a striking sparsity pattern. Specifically, the vast majority of the heatmap areas remain in the low-value range (indicated by light colors) across all communication rounds, signaling that most parameters remain effectively stable. In contrast, noticeable updates are concentrated in only a small fraction of layers (indicated by dark colors). This observation implies that the standard practice of synchronizing all parameters constitutes significant resource waste. Continuously transmitting and computing gradients for these inactive layers consumes bandwidth and computational power without contributing effectively to model adaptation, motivating our design of a magnitude-based freezing mechanism.

Challenge. Translating the observation of update redundancy into an effective federated strategy is non-trivial. Designing a robust freezing mechanism must overcome three practical challenges:

1) In FL, local gradients often fail to reflect global importance. Relying on local importance for freezing decisions can lead to disjoint optimization paths and model divergence. Thus, identifying redundancy requires a unified global perspective.

2) While layer importance is dynamic, frequent mask updates (e.g., per-round switching) disrupt the optimization trajectory and cause oscillation. A mechanism is required to balance adaptation with structural stability.

3) Models may require rapid parameter shifts during early training. Premature or aggressive freezing impedes initial feature learning and leads to sub-optimal convergence, highlighting the need for a progressive schedule.

4 Proposed Method

We begin by systematically defining and framing the core issues before tackling the challenges presented above. The GMFL module is then employed to provide targeted solutions to these identified problems.

4.1 Problem Formulation

We consider a general federated fine-tuning scenario comprising N clients, indexed by i, where each client holds a private, Non-IID dataset $\mathcal{D}_i$. Let W_0 denote the frozen pre-trained weights of the LLM. Our goal is to collaboratively optimize the trainable parameters θ (i.e., the adapters), while keeping W_0 fixed. The global optimization problem is formulated as:

$$\min_{\theta} F(\theta) = \sum_{i=1}^{N} p_i F_i(\theta), \tag{1}$$

$$F_i(\theta) = \mathbb{E}_{\xi \sim \mathcal{D}_i}[\ell(W(\theta); \xi)], \tag{2}$$

where $p_i = \frac{|\mathcal{D}_i|}{\sum_{j=1}^{N} |\mathcal{D}_j|}$ represents the aggregation weight of client i, and $\ell(\cdot)$ denotes the loss function computed on input batch ξ. Here, $W(\theta)$ represents the effective weights of the fine-tuned model parameterized by the adapters θ.

To efficiently solve this problem, we adopt LoRA. Specifically, when fine-tuning a pre-trained weight matrix $W_0 \in \mathbb{R}^{k \times d}$ to obtain the adapted weight W, LoRA decomposes the update ΔW into two low-rank matrices $B \in \mathbb{R}^{k \times r}$ and $A \in \mathbb{R}^{r \times d}$:

$$W = W_0 + \Delta W = W_0 + BA, \tag{3}$$

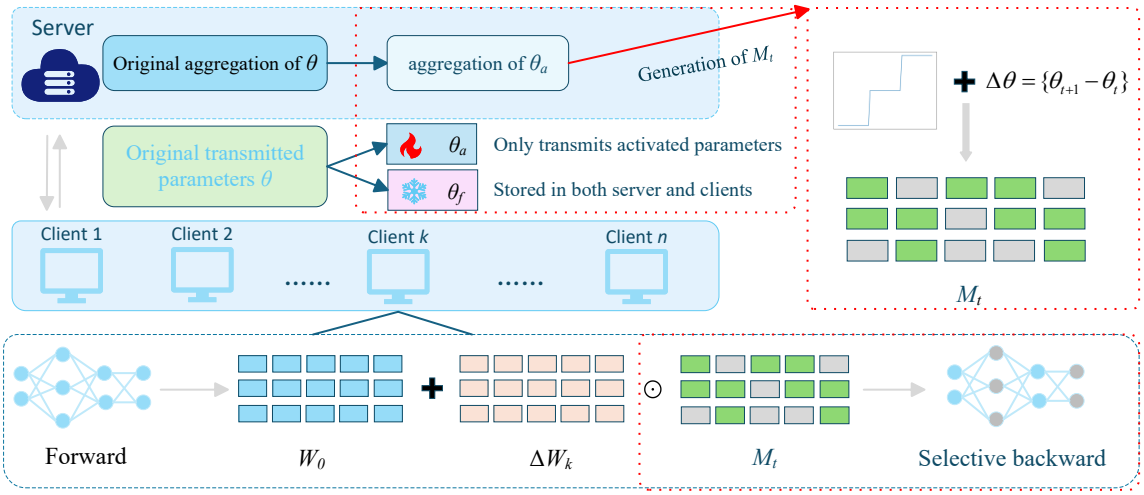


Figure 2: An overview of the proposed GMFL. As marked by red dashed blocks, GMFL integrates seamlessly into the general federated LLM fine-tuning framework with LoRA. It periodically freezes LoRA layers with smaller global update magnitude and only trains and transmits LoRA layers with larger global update magnitude.

where $r \ll \min(k, d)$ denotes the rank. In this formulation, W_0 remains frozen, while A and B contain the trainable parameters. Consequently, the learnable parameters θ consist of the collection of these low-rank matrices $\{A, B\}$ from all adapted layers. Following standard practice, A is initialized with a random Gaussian distribution and B with zero, ensuring that $\Delta W = 0$ at the beginning of training.

4.2 The Design of GMFL Module

To navigate the trade-off between communication efficiency and model performance, GMFL employs a dynamic, magnitude-based freezing strategy. The training process is structured as follows.

Warm-up Stage. To prevent the detrimental effects of premature freezing (referring to *Challenge 3*)), we initiate the training with a warm-up stage. In this phase, all original activated LoRA layers are jointly optimized. The primary goal of this stage is to allow the parameter update patterns to stabilize, thereby enabling the accurate identification of redundant layers before any freezing is applied. In each communication round t, the server aggregates the updates from sampled clients using standard Federated Averaging (FedAvg):

$$\theta_{t+1} = \sum_{i \in \mathcal{S}_t} w_i \theta_{t,i}, \tag{4}$$

where $w_i = \frac{p_i}{\sum_{j \in S_t} p_j}$ and θ_{t+1} represents the original aggregated global LoRA parameters. This ensures that the subsequent magnitude-based selection is performed on a reliable and representative model state.

Global Magnitude-Based Freezing. Following the warm-up, we introduce the freezing mechanism. To address the local-global discrepancy (referring to *Challenge 1*)), we deliberately avoid using local gradient norms, which are biased by Non-IID data. Instead, we assess layer importance based on the global update magnitude derived strictly from the aggregated server model. This ensures a unified global perspective when identifying redundancy.

First, we determine the current *freezing ratio* $\tau \in [0, 1]$ via a scheduling function $\tau = h(t)$. Then, to identify the redundant parameters, we quantify the global magnitude for each matrix. For the l-th layer at round t, the magnitudes $V_t^{\theta,l}$ are calculated as the L_1 norm of the global parameter variations:

$$V_t^{\theta,l} = \|\theta_{t+1}^l - \theta_t^l\|_1, \tag{5}$$

We then construct a unified set $\mathcal{V}_t$ collecting these magnitudes across all layers:

$$\mathcal{V}_t = \{V_t^{\theta,l}\}. \tag{6}$$

Next, we sort all elements in $\mathcal{V}_t$ in ascending order. The cut-off threshold V_τ is defined as the value at the τ-th percentile of this pooled set. In other words, the bottom τ fraction of the smallest updates in the entire model fall below V_τ.

The binary mask $M_t^{\theta,l}$ is generated by comparing each matrix's individual magnitude against this unified threshold. For any θ^l:

$$M_t^{\theta,l} = \begin{cases} 0, & \text{if } V_t^{\theta,l} < V_\tau \quad \text{(Frozen)}, \\ 1, & \text{otherwise} \quad \text{(Active)}. \end{cases} \tag{7}$$

Under this mechanism, matrices A and B within the same layer compete independently.

Consequently, clients only transmit and aggregate the non-frozen parameters. For each l, the global aggregation rule is adjusted as follows:

$$\theta_{t+1}^l = \sum_{i \in \mathcal{S}_t} w_i \theta_{t,i}^l, \text{ s.t. } M_t^{\theta,l} = 1, \tag{8}$$

For frozen parameters (where $M = 0$), the global model retains the value from the previous round without aggregation. The freezing ratio τ is not static. Tt is adaptively adjusted over communication rounds according to a schedule $\tau = h(t)$.

Periodic and Progressive Scheduling. To guarantee the stability of the sparse topology and avoid optimization oscillation (addressing *Challenge 2*), we implement a periodic updating mechanism. Instead of re-evaluating the freezing mask at every communication round—which may disrupt the convergence trajectory—we update the masks only at fixed intervals of E rounds. Within each period (i.e., when $t \pmod E \neq 0$), the mask remains constant, allowing the active parameters to adapt sufficiently to the current sparse structure.

Furthermore, recognizing that parameters gradually stabilize as training proceeds, we employ a progressive schedule to increase the freezing ratio over time. The scheduling function $h(t)$ is defined as a step-wise increasing function:

$$\tau(t) = \min\left(\tau_{\max},\ c + \left\lfloor \frac{t}{E} \right\rfloor \times \beta\right), \tag{9}$$

Here, c represents the initial freezing ratio (constant), E is the predefined period length, and β is the incremental freezing rate applied after each period. To prevent total freezing, we typically set an upper bound $\tau_{\max}$ (e.g., 0.9). This strategy ensures that the model progressively focuses computational resources on the most stubborn parameters, maximizing efficiency without compromising global convergence. Figure 2 illustrates the procedure for inserting the proposed GMFL module into the existing system.

5 Theoretical Analysis

Intuitively, GMFL reduces the number of trained parameters across clients, which is expected to decrease client model heterogeneity and thereby mitigate the impact of Non-IID data. However, this freezing strategy may evidently limit the model's expressive capacity. The precise nature of this trade-off is examined in the following analysis.

Under assumptions in Appendix D.1, we can get

Theorem 1 (Convergence of GMFL) *With learning rate $\eta_t = \frac{1}{\sqrt{T}}$, after T rounds:*

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla F(\theta_t)\|^2] \leq & \frac{2(F(\theta_0) - F^*)}{k\rho_{\min}\sqrt{T}} \\ & + \frac{Lk(G^2 + \sigma^2\omega)}{k\rho_{\min}\sqrt{T}} \\ & + \frac{2\bar{\mu}^2\zeta^2}{\rho_{\min}^2} + \frac{2L^2G^2k^2}{3\rho_{\min}^2 T}. \end{aligned}$$

Here, $F(\theta)$ denotes the global objective function, T is the total number of communication round, θ_t represents the model parameters at round t, and F^* is the infimum of $F(\theta)$. The key parameters are: k is the number of clients selected per round; L is the Lipschitz smoothness constant of F; G and σ denote the upper bounds on the norm and variance of stochastic gradients, respectively; ζ quantifies the degree of client heterogeneity; $\bar{\mu}$ is the heterogeneity suppression coefficient introduced by our method; and $\rho_{\min} \in [0, 1]$ represents the minimum gradient-mask alignment coefficient.

The proof is in Appendix D.

For a sufficiently large T, the convergence rate is dominated by:

$$\min_{t \in [T]} \mathbb{E}[\|\nabla F(\theta_t)\|^2] \leq O\left(\frac{1}{\rho_{\min}\sqrt{T}} + \frac{\bar{\mu}^2\zeta^2}{\rho_{\min}^2}\right). \tag{10}$$

As shown in Theorem 1 and Equation (10), compared to traditional FL whose convergence rate is dominated by $O\left(\frac{1}{\sqrt{T}} + \zeta^2\right)$, our method introduces two additional hyperparameters, ρ_{min} and $\bar{\mu}$. This highlights a trade-off inherent to our approach: excessively freezing parameters could substantially impair model performance. We will discuss this empirically in the experiments. However, as motivated in our introduction, LLMs do not require excessively large or frequent updates, as a portion of these updates is redundant. Therefore, in most practical scenarios, the positive benefits conferred by $\bar{\mu}$ outweigh the drawbacks associated with ρ_{min}.

Model	Parameters	Architecture
MiniMind2-Small	26M	Decoder-only
RoBERTa-base	125M	Encoder-only
LlaMa-2-7b	7B	Decoder-only

Table 1: Summary of the tested models in this paper.

Theorem 2 (Effect on Loss)

$$\Delta \ell_{diff} \approx \eta_t \|g\|^2 R_F,$$

where

$$R_F := \frac{\|g_F\|^2}{\|g\|^2} = \frac{\sum_{j \in F} g_j^2}{\sum_{j=1}^{d} g_j^2} \in [0, 1],$$

where F represents the frozen LoRA parameters, g_j denotes the gradient of j-th layer and d is the number of layers.

The proof is in Appendix E.

Ideally, if the set of frozen LoRA parameters (characterized by small updates) coincides exactly with the set of parameters that have consistently small gradient updates from every client, then this selective freezing strategy would have a small detrimental effect on client-side training.

6 Experiments

6.1 Experiments Setup

Models and Datasets. To ensure a comprehensive evaluation across different architectures, we assess our method on both **Natural Language Understanding (NLU)** and **Generation (NLG)** tasks. We employ RoBERTa-base (Liu et al., 2019) for NLU, evaluated on the GLUE benchmark. For NLG, we fine-tune LlaMa-2-7B (Touvron et al., 2023) across three distinct capabilities, utilizing 20k training samples for each: (1) **Arithmetic Reasoning**, fine-tuned on MetaMathQA (Yu et al., 2024) and evaluated on GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021); (2) **Commonsense Reasoning**, trained and evaluated using COMMONSENSE170K (Hu et al., 2023); and (3) **General Generation**, fine-tuned on Alpaca-GPT4 (Peng et al., 2023) and assessed via MT-Bench (Zheng et al., 2023). All experiments are conducted on a single NVIDIA A800 GPU with `bfloat16` precision, reporting the mean and standard deviation of three independent runs. See Appendix B.1 for further details. The tested model sizes in this paper are shown in Table 1.

Hyperparameter setting. In all experiments, we set $E = 20$ and $c = 0.1$. All models are optimized with AdamW with weight decay 0.01.

Baseline. We compare our method with the following SOTA algorithms: **FedIT** (Zhang et al., 2024), **FFA-LoRA** (Sun et al., 2024), **LoRA-A^2** (Koo et al., 2025) and **FedEx-LoRA** (Singhal et al., 2025). Detailed description of these algorithms are in Appendix B.2.

As a plug-and-play module, according to the performance of these algorithms, we implement our method based on three baseline frameworks with superior performance, FedIT, LoRA-A^2 and FedEx-LoRA. Thus our results are denoted as FedIT+GMFL, LoRA-A^2+GMFL and FedEx-LoRA+GMFL.

6.2 Natural Language Understanding

Setup and Implementation. We evaluate on five GLUE datasets (SST-2, CoLA, MNLI, RTE, QQP, MRPC) using RoBERTa-base (125M) (Liu et al., 2019). To simulate Non-IID data, training sets are partitioned via Dirichlet sampling with $\phi = 0.5$ where ϕ controls the level of data heterogeneity. Following the experimental setting in (Singhal et al., 2025), we apply LoRA to query and value matrices with two settings: $r = 4(\alpha = 8)$ and $r = 1(\alpha = 2)$. Rank 4 was found to be sufficient for GLUE tasks, while rank 1 was included to evaluate performance under a low-rank constraint. Training runs for 200 rounds using AdamW (lr=2e-4, cosine annealing schedules), sampling 2 clients per round. Total clients and local steps vary by dataset (see Appendix B.3). For GMFL, we grid-search $\beta \in [0.01, 0.09]$. Performance is measured by Matthews correlation for CoLA and accuracy for others. The best β for SST-2, CoLA, MNLI, RTE, QQP, MRPC is $0.02, 0.02, 0.08, 0.05, 0.04, 0.07$ respectively.

Main Results. As shown in Table 2, GMFL demonstrates robust compatibility across ranks ($r \in \{1, 4\}$). In heterogeneous settings, it reduces communication and computational costs while matching baseline performance. This empirically validates our theoretical analysis that the benefits of $\bar{\mu}$ outweigh the drawbacks of ρ_{min}. Additional evaluations are provided in Appendix C (Table 7).

6.3 Natural Language Generation

Implementation Details. We fine-tune LlaMa-2-7B on three distinct generation tasks, utilizing

rank	Method	SST-2	MNLI (mismatched)	MNLI (matched)	QQP	MRPC
r=4	FedIT	92.66±0.16	81.08±0.96	**80.66±0.79**	82.49±1.63	86.19±2.53
	FFA-LoRA	91.78±0.14	77.96±1.05	77.13±0.95	79.22±2.88	80.80±3.10
	LoRA-A^2	91.06±0.32	77.19±0.79	76.20±0.93	77.77±2.98	83.82±3.28
	FedEx-LoRA	92.58±0.72	80.76±0.37	80.28±0.49	79.20±5.34	85.62±2.78
	FedIT+GMFL	**92.92±0.11**↑	**81.10±0.87**↑	80.65±0.86↓	82.51±1.16↑	**86.19±1.70**↑
	LoRA-A^2+GMFL	91.62±0.41↑	79.90±0.64↑	79.25±0.34↑	79.45±2.75↑	83.01±2.61↓
	FedEx-LoRA+GMFL	92.77±0.83↑	80.56±0.80↓	79.98±0.88↓	**82.75±2.03**↑	71.00±3.18↓
r=1	FedIT	92.01±0.22	80.16±1.10	79.42±0.78	81.13±2.05	84.64±1.67
	FFA-LoRA	86.77±1.61	65.21±2.46	63.75±2.49	73.59±3.78	71.16±1.82
	LoRA-A^2	90.71±0.95	76.64±2.22	75.51±2.38	76.49±3.22	79.90±2.46
	FedEx-LoRA	92.47±0.70	**80.72±0.16**	79.75±0.17	81.57±0.96	**85.46±1.17**
	FedIT+GMFL	92.09±0.16↑	80.08±1.00↓	79.26±0.72 ↓	81.04±0.84↓	84.89±1.80↑
	LoRA-A^2+GMFL	90.94±0.57↑	77.53±1.05↑	76.42±0.92↑	77.97±2.55↑	79.49±4.27↓
	FedEx-LoRA+GMFL	**92.62±0.66**↑	80.63±0.44↓	**79.83±0.18**↑	**83.19±1.92**↑	85.38±1.63↓

Table 2: Results with RoBERTa-base on the GLUE benchmark datasets, comparing various federated LoRA methods at ranks $r = \{4, 1\}$. The best results for each dataset are indicated in **bold**, while the second-best results are highlighted using underline.

20k training samples for each. For **Arithmetic Reasoning**, we train on MetaMathQA (Yu et al., 2024) and evaluate on the GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) benchmarks. For **Commonsense Reasoning**, we utilize the COMMONSENSE170K dataset (Hu et al., 2023) for both training and evaluation. For **General Generation**, we train on Alpaca-GPT4 (Peng et al., 2023) and assess performance via MT-Bench (Zheng et al., 2023). To mitigate potential assessment bias, we employ both GPT-4o and DeepSeek-V3.2 as judge models. Since the above-mentioned datasets lack explicit labels, the Dirichlet distribution cannot be applied to simulate non-IID scenarios, therefore, we adopt an IID strategy to partition these datasets. Performance is measured by accuracy on commonsense and mathematical reasoning benchmarks and by MT-Bench scores (1–10) for general generation tasks.

LoRA adapters ($r = 32, \alpha = 64$) are applied to all linear layers (attention query/key/value/output and MLP modules). We use the AdamW optimizer (learning rate 5e-5, batch size 8) with cosine annealing for 200 communication rounds, performing 20 local update steps per round. The total client pool consists of 10 clients for Math Reasoning and 20 for other tasks, with $K = 2$ clients sampled per round. We perform a grid search for β over $\{0.02, 0.03, 0.05\}$ for math tasks and $\{0.05, 0.06, 0.07\}$ for others, reporting the best results across seeds.

Main Results. The average performance of Commonsense Reasoning, the average score of General Generation task and results of two Math Reasoning are presented in Table 3. Detailed experimental results are shown in Appendix C (Tables 8 and 9). GMFL demonstrates a superior trade-off between efficiency and performance: it reduces communication by approximately 20–30% while limiting performance degradation to within 0.1%–1%. Notably, our method slightly surpasses existing SOTA baselines on most datasets.

6.4 Communication and Memory Cost

Communication overhead. Let $\mathcal{C}_{\text{base}}$ denote the standard communication volume of LoRA parameters required by baseline methods (e.g., FedIT) across the entire training process. In contrast, GMFL significantly reduces this overhead through its magnitude-based progressive freezing.

Accounting for a 10% warm-up stage (where all parameters are active), the theoretical communication volume of LoRA parameters in our method is derived as:

$$\mathcal{C}_{\text{ours}} \approx [1 - 0.9(c + 5\beta)] \times \mathcal{C}_{\text{base}}, \qquad (11)$$

where the term $0.9(c+5\beta)$ represents the effective reduction achieved during the 90% freezing phase. For instance, with specific settings (e.g., $c = 0.1$ and $\beta = 0.05$), our method can drastically reduce communication costs while maintaining original model performance, as empirically demonstrated in Figure 3.

Memory overhead. We further analyze client-side memory requirements in Figure 4. By freezing a subset of parameters, we substantially reduce the

Method	COMMONSENSE170K$_{Avg}$	GSM8k	MATH	MT-Avg$_{GPT-4o}$	MT-Avg$_{DeepSeekV3.2}$
FedIT	74.81±0.33	40.11±0.73	5.77±0.10	3.61±0.05	3.94±0.11
FFA-LoRA	72.40±0.32	34.22±0.89	4.68±0.12	3.54±0.10	3.81±0.08
LoRA-A^2	73.47±0.27	35.56±0.33	4.71±0.20	3.60±0.03	3.91±0.12
FedEx-LoRA	75.24 ± 0.32	40.79±0.11	**6.22±0.13**	3.70±0.07	3.89±0.12
FedIT+GMFL	74.99±0.39↑	40.03±0.81↓	5.85±0.21↑	3.67±0.12↑	3.98±0.10↑
LoRA-A^2+GMFL	73.17±0.19↓	35.56±0.22↑	4.90±0.17↑	3.59±0.06↓	3.87±0.16↓
FedEx-LoRA+GMFL	**75.57±0.39**↑	**41.12±0.37**↑	6.14±0.06↓	**3.72±0.10**↑	**4.00±0.01**↑

Table 3: Main results of Natural Language Generation tasks. The best results for each dataset are indicated in **bold**, while the second-best results are highlighted using underline.

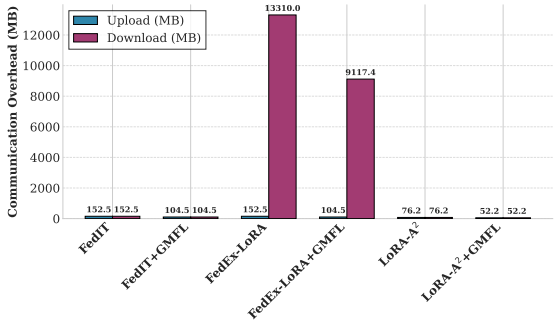


Figure 3: Average communication overhead with LlaMa-2-7b, where $c = 0.1, \beta = 0.05$.

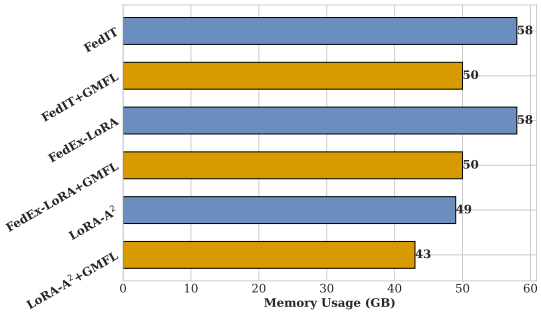


Figure 4: Average memory consumption with LlaMa-2-7b, where $c = 0.1, \beta = 0.05$.

memory footprint required for storing optimizer states, directly alleviating the computational burden. Note that the reported results represent the average consumption over the entire training process. Crucially, due to the progressive nature of our scheduling, GMFL requires less memory in later training stages as the freezing ratio increases. This progressive efficiency enables resource-constrained clients to sustain participation throughout the federated lifecycle, significantly enhancing the framework's practical deployability.

6.5 Hyperparameter Analysis

We investigate the sensitivity of β and c, the control knob that governs the trade-off between resource efficiency and model performance.

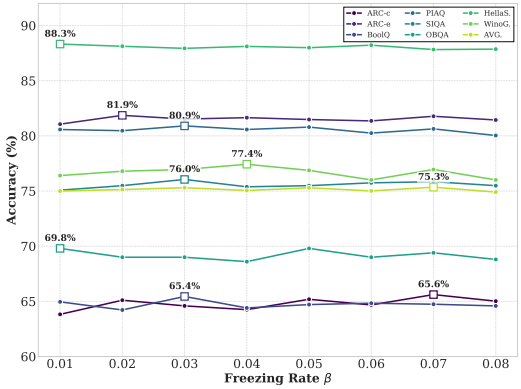


Figure 5: Sensitivity of β conducted on COMMONSENSE170K dataset.

As illustrated in Figure 5, the model performance exhibits remarkable stability across a wide range of β values. Extended analyses on NLU tasks are provided in Appendix F (Figure 7- Figure 13). Intriguingly, from Figure 5, we observe that a larger β occasionally yields superior accuracy compared to conservative settings. This empirical evidence reinforces our hypothesis regarding parameter redundancy: by pruning non-essential parameters, the optimization process focuses exclusively on the most informative parameters. This mechanism effectively acts as a regularizer, enhancing the global model's robustness against noise arising from data heterogeneity.

As indicated in Figure 6, model performance is sensitive to the initial freezing rate, with an optimal range yielding the best results. An excessively high rate has a detrimental effect, which motivated our design of a progressive freezing scheduler to circumvent this performance degradation.

6.6 Ablation Study: freezing A/B jointly vs independently

We explore evaluating A and B jointly using $\|BA\|$ as the importance criterion for FedIT. The comparison results for natural language generation tasks

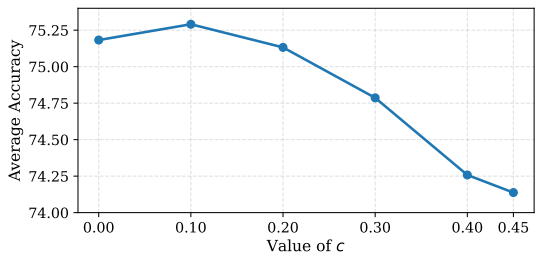


Figure 6: Sensitivity of c conducted on COMMONSENSE170K dataset.

Dateset	Jointly	Independently
ARC-Challenge	46.53 ± 2.97	**63.94±0.50**
ARC-Easy	69.19 ± 3.31	**80.78±0.34**
BoolQ	54.84 ± 6.25	**65.17±0.49**
PIQA	72.65 ±1.96	**79.85±0.58**
Social IQA	57.30 ± 3.21	**75.20±0.24**
OpenBookQA	47.53 ± 4.54	**68.47±0.66**
HellaSwag	60.06 ± 5.27	**87.54±0.34**
WinoGrande	52.28 ± 1.90	**76.37±0.27**
GSM8K	26.46 ± 0.33	**39.88±0.60**
MATH	3.41 ± 0.33	**5.68±0.03**

Table 4: Results of Natural Language Generation tasks for Ablation Study: freezing A/B jointly vs independently. The best results for each dataset are indicated in **bold**.

are presented in Table 4 and the comparison results for natural language understanding tasks are presented in Table 5. The results show that the independent strategy consistently outperforms joint evaluation. We attribute this to the fact that when A and B are multiplied, a large norm in one matrix can be canceled by a small norm in the other, leading to underestimated importance even when one adapter contains highly task-relevant information. Evaluating A and B independently thus provides a more reliable and stable importance criterion.

6.7 Dissusion

In our experiments, we follow the setup of (Ye et al., 2024), in which only a subset of client data is utilized during training. Consequently, certain layers that are critical for extracting features from unseen data may be frozen prematurely due to limited exposure, resulting in a slight performance gap compared to full-update baselines. Nevertheless, given the substantial reductions in computational and communication costs achieved through freezing, we consider this trade-off acceptable.

Dataset	Jointly$_{r=4}$	Independently$_{r=4}$	Jointly$_{r=1}$	Independently$_{r=1}$
SST-2	90.86±0.05	**92.13±0.44**	90.21±0.76	**91.78±0.80**
COLA	25.24±6.58	**75.39±4.45**	11.14±8.07	**77.60±1.92**
MNLI(mismatched)	74.53±2.20	**80.29±0.73**	69.81±2.78	**79.39±0.38**
MNLI(matched)	73.50±2.10	**79.80±0.23**	68.49±2.71	**78.29±0.36**
QQP	76.99±2.43	**80.89±1.88**	75.11±3.34	**79.24±2.14**
MRPC	76.31±4.94	**84.40±2.53**	73.12±2.69	**82.92±3.18**
RTE	57.28±4.30	**65.22±6.81**	56.56±4.70	**63.90±5.65**

Table 5: Results with RoBERTa-base on the GLUE datasets for Ablation Study: freezing A/B jointly vs independently. The best results for each dataset are indicated in **bold**.

7 Conclusion

In this paper, we propose GMFL, a communication-efficient and plug-and-play module for federated fine-tuning of LLMs. GMFL alleviates training and communication inefficiencies arising from redundant LoRA parameters in federated settings. Specifically, we introduce a dynamic freezing strategy guided by parameter update magnitudes, which selectively freezes parameters with negligible contributions. Theoretical analysis demonstrates that freezing such negligible updates effectively mitigates the adverse effects of client heterogeneity and guarantees convergence with only a small deviation in model performance. Extensive experimental results show that GMFL significantly reduces communication and memory overhead while maintaining competitive performance, with accuracy fluctuation within 1%.

8 Limitation

Our current implementation operates at the matrix level, freezing entire A or B matrices to avoid complex indexing overhead. While effective, exploring finer-grained sparsity (e.g., row-wise or neuron-level freezing) could potentially yield higher compression rates and reduce brittleness, which we leave for future work. Additionally, although GMFL demonstrates robust performance across diverse NLP tasks, its generalization to other modalities, such as computer vision or multimodal federated learning, remains to be verified.

Acknowledgments

This work was supported in part by the National Natural Science Foundations of China under Grant 62372185 and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515010131.

References

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics.

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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Jingyao Gong. 2024. Minimind: Train a tiny llm from scratch. https://github.com/jingyaogong/minimind.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. 2025. Selective aggregation for low-rank adaptation in federated learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

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. *ArXiv*, abs/1902.00751.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/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. *Preprint*, arXiv:2304.01933.

Jincheol Jung, Hongju Jeong, and Eui-Nam Huh. 2025. Federated learning and rag integration: A scalable approach for medical large language models. In *2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 0968–0973.

Jabin Koo, Minwoo Jang, and Jungseul Ok. 2025. Towards robust and efficient federated low-rank adaptation with heterogeneous clients. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 416–429, Vienna, Austria. Association for Computational Linguistics.

Jahnavi Kumar and Sridhar Chimalakonda. 2024. Code summarization without direct access to code - towards exploring federated llms for software engineering. In *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering*, EASE '24, page 100–109, New York, NY, USA. Association for Computing Machinery.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing*.

Sheng Li, Geng Yuan, Yue Dai, Youtao Zhang, Yanzhi Wang, and Xulong Tang. 2024a. Smartfrz: An efficient training framework using attention-based layer freezing. *Preprint*, arXiv:2401.16720.

Shiwei Li, Yingyi Cheng, Haozhao Wang, Xing Tang, Shijie Xu, Weihong Luo, Yuhua Li, Dugang Liu, Xiuqiang He, and Ruixuan Li. 2024b. Masked random noise for communication-efficient federated learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 3686–3694, New York, NY, USA. Association for Computing Machinery.

Shiwei Li, Xiandi Luo, Haozhao Wang, Xing Tang, Shijie Xu, Weihong Luo, Yuhua Li, Xiuqiang He, and Ruixuan Li. 2025. The panaceas for improving low-rank decomposition in communication-efficient federated learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 35536–35561. PMLR.

Shiwei Li, Wenchao Xu, Haozhao Wang, Xing Tang, Yining Qi, Shijie Xu, Weihong Luo, Yuhua Li, Xiuqiang He, and Ruixuan Li. 2024c. FedBAT: Communication-efficient federated learning via learnable binarization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 29074–29095. PMLR.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. *Preprint*, arXiv:2402.09353.

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. *Preprint*, arXiv:1907.11692.

Yuhan Liu, Saurabh Agarwal, and Shivaram Venkataraman. 2021. Autofreeze: Automatically freezing model blocks to accelerate fine-tuning. *Preprint*, arXiv:2102.01386.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Rui Pan, Xiang Liu, Shizhe Diao, Renjie Pi, Jipeng Zhang, Chi Han, and Tong Zhang. 2024. Lisa: Layerwise importance sampling for memory-efficient large language model fine-tuning. *Preprint*, arXiv:2403.17919.

Sangwoo Park, Seanie Lee, Byungjoo Kim, and Sung Ju Hwang. 2025. Fedrand: Enhancing privacy in federated learning with randomized lora subparameter updates. *ArXiv*, abs/2503.07216.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *Preprint*, arXiv:2304.03277.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *Preprint*, arXiv:1907.10641.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. 2019. Natural language understanding with the quora question pairs dataset. *Preprint*, arXiv:1907.01041.

Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. 2025. Fedex-lora: Exact aggregation for federated and efficient fine-tuning of foundation models. *Preprint*, arXiv:2410.09432.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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*, abs/2409.05976.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Shuyue Wei, Yongxin Tong, Zimu Zhou, Yi Xu, Jingkai Gao, Tongyu Wei, Tianran He, and Weifeng Lv. 2025. Federated reasoning LLMs: a survey. *Frontiers of Computer Science*, 19(12):1912613.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Rui Ye, Rui Ge, Xinyu Zhu, Jingyi Chai, Yaxin Du, Yang Liu, Yanfeng Wang, and Siheng Chen. 2024. Fedllm-bench: Realistic benchmarks for federated learning of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 111106–111130. Curran Associates, Inc.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models. *Preprint*, arXiv:2309.12284.

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

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *Preprint*, arXiv:2303.10512.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Design of GMFL Module

The pseudocode of the proposed GMFL is described in Algorithm 1.

Algorithm 1 GMFL

Require: Total rounds T, local steps k, clients $\{1, \ldots, N\}$, client weights $\{p_i\}$, initial LoRA parameters θ_0, initial $M_0 = \{1\}^d$, learning-rate scheduler $\{\eta_t\}$

Initialize global parameters θ_0
for $t = 0, 1, \ldots, T-1$ **do**
  Server samples client subset C_t
  for all $i \in C_t$ **in parallel do**
    $\theta_{t,i}^{(0)} \leftarrow \theta_t$
    for $e = 0, 1, \ldots, k-1$ **do**
      Sample mini-batch $\xi_{t,i}^e$
      $g_{t,i}^e \leftarrow \nabla \ell(\theta_{t,i}^e; \xi_{t,i}^e)$
      $g_{t,i}^e \leftarrow M_t \odot g_{t,i}^e$ {masked gradient}
      $\theta_{t,i}^{(e+1)} \leftarrow \theta_{t,i}^e - \eta_t \Delta$ {AdamW update or SGD update}
    end for
    Send $\theta_{t,i}^{(k)}$ to server
  end for
  Server aggregates:

$$\theta_{t+1} \leftarrow \sum_{i \in S_t} w_i \, \theta_{t,i}^{(k)}, \qquad w_i = \frac{p_i}{\sum_{j \in S_t} p_j}$$

  if $(t+1)\%E == 0$ **then**
    Server computes new global mask $M_t \in \{0,1\}^d$
  end if
end for
return θ_T

B Experiment Details

B.1 Details of Datasets

COMMONSENSE170K is a dataset combining eight Commonsense Reasoning datasets(Hu et al., 2023), as detailed below:

1. WinoGrande (Sakaguchi et al., 2019) involves filling in blanks with binary choices based on sentences that demand commonsense reasoning

2. HellaSwag (Zellers et al., 2019) asks the model to predict the most plausible continuation of a given context by selecting the correct ending from several options.

3. ARC Challenge or ARC-c (Clark et al., 2018) consists of multiple-choice science questions designed to challenge models with more complex reasoning, making them harder for methods that rely solely on co-occurrence patterns.

4. PIQA (Bisk et al., 2019) tests physical commonsense reasoning, where the task is to choose the best action from a set of options in a hypothetical situation.

5. BoolQ (Clark et al., 2019) focuses on yes/no question answering from naturally occurring queries.

6. ARC Easy or ARC-e(Clark et al., 2018) consists of grade-school-level multiple-choice science questions, providing a simpler set of tasks for testing models' basic reasoning abilities.

7. OBQA (Mihaylov et al., 2018) contains open-book, knowledge-intensive QA tasks requiring multi-hop reasoning to answer questions that involve integrating information from multiple sources.

8. SIQA (Sap et al., 2019) focuses on understanding human actions and predicting their social consequences, evaluating models' social commonsense reasoning.

MetaMathQA dataset (Yu et al., 2024) generates mathematical questions by rephrasing them from various perspectives without introducing additional knowledge. We evaluate this dataset on two benchmarks: GSM8K (Cobbe et al., 2021) which includes grade-school math word problems that require multi-step reasoning, and MATH(Hendrycks et al., 2021), which features challenging competition-level mathematics problems.

GLUE Benchmark is a diverse suite of tasks for evaluating natural language understanding capabilities. It includes datasets such as SST-2 (Socher et al., 2013) for sentiment analysis, MRPC (Dolan and Brockett, 2005) and QQP (Sharma et al., 2019) for paraphrase detection, CoLA for linguistic acceptability (Warstadt et al., 2019), RTE (Cer et al., 2017) and MNLI (Williams et al., 2018) for inference, and MNLI is a collection of texts from many different domains and styles. Therefore, it is divided into two dataset versions: matched and

mismatched. "Matched" means the training set and the test set come from the same source, while "mismatched" means the training set and the test set come from different sources. Due to its comprehensive coverage of NLU tasks, GLUE is widely used to assess models like RoBERTa. Each dataset is released under its own license.

Alpaca-GPT4(Peng et al., 2023) is generated via GPT-4(OpenAI et al., 2024) using Self-Instruct.

B.2 Details of Compared Algorithms

The followings are details of algorithms we compare with.

1. **FedIT**(Zhang et al., 2024): a classic method which applies vanilla federated averaging (FedAvg) to LoRA.

2. **FFA-LoRA**(Sun et al., 2024): it freezes the A matrices and trains only the B matrices, allowing for exact aggregation in a federated setting but losing the information of A matrices

3. **LoRA-A^2**(Koo et al., 2025): it alternately freezes the A matrices and the B matrices to address the data heterogeneity problem and allows exact aggregation in a federated setting

4. **FedEx-LoRA** (Singhal et al., 2025): A state-of-the-art method that applies a residual error term to frozen weights to mitigate aggregation noise.

B.3 Experiment Setting

The hyperparameter settings of GLUE Benchmark are shown in Table 6.

dataset	batch size	max steps	clients
SST-2	32	30	20
COLA	32	10	20
QQP	32	30	20
NMLI	32	30	20
MRPC	32	10	10
RTE	32	10	5

Table 6: Hyperparameter settings for GLUE Benchmark.

C More experimental results

The experimental results on COLA and RTE are presented in Table 7. It can be observed that data heterogeneity exerts a notable influence on various methods in our experiments. Nevertheless, with the integration of the GMFL module, most approaches demonstrate an ability to mitigate the effects induced by data heterogeneity, thereby further corroborating the analysis provided in Theorem 1.

Detailed results on the eight commonsense reasoning datasets from COMMONSENSE170K are summarized in Table 8. Overall, GMFL contributes to consistent performance improvements across baseline methods while simultaneously reducing communication and memory overhead. The outcomes on MT-Bench under two distinct judge models are displayed in Table 9. The findings reveal that by leveraging the GMFL module to selectively freeze non-essential LoRA layers, performance fluctuations can be effectively confined within a narrow range. In Natural Language Generation, since our experiments are conducted under an IID setting with minimal client heterogeneity, the primary factor influencing model performance is ρ. The experimental results demonstrate that ρ preserves the majority of the update energy, confirming our earlier assertion that some LoRA updates are non-essential.

D Proof of Theorem 1

D.1 Assumptions

Assumption 1 (Smoothness) *Each local function F_i is L-smooth: for all $\theta, \theta' \in \mathbb{R}^d$,*

$$\|\nabla F_i(\theta) - \nabla F_i(\theta')\| \leq L\|\theta - \theta'\|.$$

Equivalently,

$$F_i(x + \Delta) \leq F_i(x) + \langle \nabla F_i(x), \Delta \rangle + \frac{L}{2}\|\Delta\|^2.$$

Assumption 2 (Bounded Stochastic Gradient) *The stochastic gradient has bounded second moment: for all θ and i,*

$$\mathbb{E}\|g_{t,i}^{(e)}\|^2 \leq G^2.$$

Assumption 3 (Bounded Variance) *The variance of stochastic gradients is bounded: for all θ and i,*

$$\mathbb{E}\|g_{t,i}^{(e)} - \nabla F_i(\theta_{t,i}^{(e)})\|^2 \leq \sigma^2.$$

Assumption 4 (Bounded Client Heterogeneity) *The client heterogeneity is bounded: for all θ and i,*

$$\|\nabla F_i(\theta) - \nabla F(\theta)\|^2 \leq \zeta^2.$$

rank	Method	COLA	RTE
r=4	FedIT	47.01±6.16	65.70±7.44
	FFA-LoRA	39.24±8.50	60.77±6.26
	LoRA-A^2	38.67±8.02	57.52±5.58
	FedEx-LoRA	36.42±6.54	69.31±5.62
	FedIT+GMFL	**50.74±4.06**↑	65.59±7.06↓
	LoRA-A^2+GMFL	36.81±10.68↓	59.69±6.86↑
	FedEx-LoRA+GMFL	39.00±7.88↑	**71.00±3.18**↑
r=1	FedIT	43.04±6.79	62.94±6.01
	FFA-LoRA	7.14±7.36	53.31±0.85
	LoRA-A^2	26.03±12.52	59.69±6.76
	FedEx-LoRA	42.52±7.53	67.75±2.38
	FedIT+GMFL	**45.11±5.06**↑	63.30±6.70↑
	LoRA-A^2+GMFL	26.29±12.73↑	58.48±7.17↓
	FedEx-LoRA+GMFL	43.22±8.61↑	**68.11±3.42**↑

Table 7: Results with RoBERTa-base on the GLUE benchmark datasets, comparing various federated LoRA methods at ranks $r = \{4, 1\}$. The best results for each dataset are indicated in **bold**, while the second-best results are highlighted using underline.

Assumption 5 (Heterogeneity Suppression) *We decompose the client gradient as*

$$\nabla F_i(\theta) = \nabla F(\theta) + \delta_i(\theta),$$

where $\delta_i(\theta)$ captures client heterogeneity. There exists a constant $\mu_t \in [0, 1)$ such that for all clients i,

$$\|M_t \odot \delta_i(\theta_t)\| \leq \mu_t \|\delta_i(\theta_t)\|.$$

Assumption 6 (Gradient-Mask Alignment) *Define the alignment coefficient at round t as:*

$$\rho_t := \frac{\|M_t \odot \nabla F(\theta_t)\|^2}{\|\nabla F(\theta_t)\|^2} \in [0, 1].$$

We assume $\rho_t \geq \rho_{\min} > 0$ for all t, meaning the mask preserves a non-trivial fraction of gradient energy.

D.2 Detailed proof

One-Round Descent. By L-smoothness (Assumption 1), we have:

$$F(\theta_{t+1}) \leq F(\theta_t) + \langle \nabla F(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2.$$

Taking expectation conditioned on θ_t, define:

$$T_1 := \mathbb{E}[\langle \nabla F(\theta_t), \theta_{t+1} - \theta_t \rangle],$$
$$T_2 := \frac{L}{2}\mathbb{E}[\|\theta_{t+1} - \theta_t\|^2].$$

Global Update Expression with Weighted Aggregation. From the local update rule:

$$\theta_{t,i}^{(k)} = \theta_t - \eta_t \sum_{e=0}^{k-1} M_t \odot g_{t,i}^{(e)}.$$

Thus, the weighted aggregation gives:

$$\theta_{t+1} - \theta_t = -\eta_t \sum_{i \in S_t} w_i \sum_{e=0}^{k-1} M_t \odot g_{t,i}^{(e)}.$$

Bounding T_1.

$$T_1 = \mathbb{E}\left\langle \nabla F(\theta_t), -\eta_t \sum_{i \in S_t} w_i \sum_{e=0}^{k-1} M_t \odot g_{t,i}^{(e)} \right\rangle = -\eta_t \sum_{e=0}^{k-1} \mathbb{E}\left\langle \nabla F(\theta_t), \sum_{i \in S_t} w_i M_t \odot g_{t,i}^{(e)} \right\rangle.$$

Using the law of total expectation and the fact that $\mathbb{E}[g_{t,i}^{(e)} \mid \theta_t, \theta_{t,i}^{(e)}] = \nabla F_i(\theta_{t,i}^{(e)})$:

$$T_1 = -\eta_t \sum_{e=0}^{k-1} \mathbb{E}\left\langle \nabla F(\theta_t), \sum_{i \in S_t} w_i M_t \odot \nabla F_i(\theta_{t,i}^{(e)}) \right\rangle. \tag{12}$$

Now, we decompose $\nabla F_i(\theta_{t,i}^{(e)})$ using the client heterogeneity decomposition (Assumption 5):

$$\nabla F_i(\theta) = \nabla F(\theta) + \delta_i(\theta),$$

Method	ARC-c	ARC-e	BoolQ	PIQA
FedIT	64.33±0.35	81.27±0.59	65.14±0.52	79.80±0.49
FFA-LoRA	61.29±0.94	79.50±0.29	64.34±0.22	79.02±0.14
LoRA-A^2	62.66±0.36	80.06±0.62	64.30±0.68	79.31±0.30
FedEx-LoRA	63.96±0.81	81.87 ± 0.25	65.48 ± 0.14	79.60±0.32
FedIT+GMFL	**64.70±0.84**↑	81.31±0.38↑	65.14±0.37↑	80.11±0.38↑
LoRA-A^2+GMFL	61.21±0.28↓	80.22±0.60↑	64.98±0.24↑	79.09±0.51↓
FedEx-LoRA+GMFL	64.59±0.85↑	**82.10±0.38**↑	**65.69±0.16**↑	**80.12±0.57**↑
Method	SIQA	OBQA	HellaS.	WinoG.
FedIT	75.25±0.21	68.40±0.98	87.59±0.36	76.72±0.55
FFA-LoRA	72.24±0.31	66.60±1.61	83.69±0.45	72.48±0.62
LoRA-A^2	73.73±0.39	66.93±0.25	86.89±0.20	73.85±0.66
FedEx-LoRA	**75.38±0.07**	69.93 ± 2.10	88.53 ± 0.17	77.11 ± 0.23
FedIT+GMFL	75.35±0.44↑	68.93±0.93↑	87.69±0.19↑	76.66±0.39↓
LoRA-A^2+GMFL	73.25±0.66↓	67.13±0.68↑	85.98±0.22↓	73.51±1.19↓
FedEx-LoRA+GMFL	75.35±0.57↓	**70.33±1.80**↑	**88.86±0.26**↑	**77.48±0.32**↑

Table 8: Detailed experimental results of Commonsense Reasoning task, comparing various federated LoRA methods at rank r = 32. The best results for each dataset are indicated in **bold**, while the second-best results are highlighted using underline.

and write:

$$\nabla F_i(\theta_{t,i}^{(e)}) = \nabla F(\theta_t) + \delta_i(\theta_t) + [\nabla F_i(\theta_{t,i}^{(e)}) - \nabla F_i(\theta_t)].$$

Then, T_1 can be split into three parts:

$$T_1 = T_{1a} + T_{1b} + T_{1c},$$

where

$$T_{1a} = -\eta_t k \mathbb{E} \left\langle \nabla F(\theta_t), M_t \odot \nabla F(\theta_t) \right\rangle,$$

$$T_{1b} = -\eta_t k \mathbb{E} \left\langle \nabla F(\theta_t), M_t \odot \sum_{i \in S_t} w_i \delta_i(\theta_t) \right\rangle,$$

and

$$T_{1c} = -\eta_t \sum_{e=0}^{k-1} \mathbb{E} \left\langle \nabla F(\theta_t), M_t \odot \sum_{i \in S_t} w_i [\nabla F_i(\theta_{t,i}^{(e)}) - \nabla F_i(\theta_t)] \right\rangle. \tag{13}$$

Bounding T_{1a}. By Assumption 6 (Gradient-Mask Alignment):

$$\begin{aligned} \langle \nabla F(\theta_t), M_t \odot \nabla F(\theta_t) \rangle &= \|M_t \odot \nabla F(\theta_t)\|^2 \\ &= \rho_t \|\nabla F(\theta_t)\|^2 \\ &\geq \rho_{\min} \|\nabla F(\theta_t)\|^2. \end{aligned}$$

Thus,

$$T_{1a} \leq -\eta_t k \rho_{\min} \mathbb{E}[\|\nabla F(\theta_t)\|^2]. \tag{14}$$

Bounding T_{1b}. Using Cauchy-Schwarz and Young's inequality:

$$\begin{aligned} |T_{1b}| &\leq \eta_t k \mathbb{E} \left[\|\nabla F(\theta_t)\| \cdot \left\| M_t \odot \sum_{i \in S_t} w_i \delta_i(\theta_t) \right\| \right] \\ &\leq \eta_t k \left(\frac{\rho_{\min}}{4} \mathbb{E}[\|\nabla F(\theta_t)\|^2] \right. \\ &\quad \left. + \frac{1}{\rho_{\min}} \mathbb{E} \left[\left\| M_t \odot \sum_{i \in S_t} w_i \delta_i(\theta_t) \right\|^2 \right] \right). \end{aligned}$$

By Jensen's inequality and Assumption 5:

$$\begin{aligned} \left\| M_t \odot \sum_{i \in S_t} w_i \delta_i(\theta_t) \right\|^2 &\leq \sum_{i \in S_t} w_i \left\| M_t \odot \delta_i(\theta_t) \right\|^2 \\ &\leq \mu_t^2 \sum_{i \in S_t} w_i \|\delta_i(\theta_t)\|^2. \end{aligned}$$

From Assumption 4 (Bounded Client Heterogeneity): $\|\delta_i(\theta_t)\| = \|\nabla F_i(\theta_t) - \nabla F(\theta_t)\| \leq \zeta$, so:

$$\left\| M_t \odot \sum_{i \in S_t} w_i \delta_i(\theta_t) \right\|^2 \leq \mu_t^2 \zeta^2.$$

Therefore,

$$|T_{1b}| \leq \frac{\eta_t k \rho_{\min}}{4} \mathbb{E}[\|\nabla F(\theta_t)\|^2] + \frac{\eta_t k \mu_t^2 \zeta^2}{\rho_{\min}}. \tag{15}$$

judge	Method	MT-1	MT-2
GPT-4o	FedIT	4.42±0.09	2.80±0.04
	FFA-LoRA	4.30±0.11	2.78±0.11
	LoRA-A^2	4.35±0.05	2.85±0.02
	FedEx-LoRA	**4.54±0.14**	2.85±0.06
	FedIT+GMFL	4.51±0.14↑	2.83±0.14↑
	LoRA-A^2+GMFL	4.35±0.11↓	2.82±0.03↓
	FedEx-LoRA+GMFL	4.53±0.11↓	**2.91±0.10**↑
DeepSeek-V3.2	FedIT	4.72±0.17	**3.16±0.13**
	FFA-LoRA	4.61±0.15	3.01±0.19
	LoRA-A^2	4.67±0.08	3.15±0.16
	FedEx-LoRA	4.73±0.09	3.06±0.18
	FedIT+GMFL	4.85±0.13↑	3.12±0.11↓
	LoRA-A^2+GMFL	4.75±0.23↑	2.95±0.10↓
	FedEx-LoRA+GMFL	**4.88±0.05**↑	3.11±0.07↑

Table 9: Detailed experimental results results of General Generation task which is evaluated by MT-Bench, comparing various federated LoRA methods at rank r = 32. The best results for each dataset are indicated in **bold**, while the second-best results are highlighted using underline.

Bounding T_{1c}. Again, using Cauchy-Schwarz and Young's inequality:

$$\begin{aligned}|T_{1c}| \leq \eta_t \sum_{e=0}^{k-1} \mathbb{E}\,[\|\nabla F(\theta_t)\| \cdot \\ \left\| M_t \odot \sum_{i\in S_t} w_i [\nabla F_i(\theta_{t,i}^{(e)}) - \nabla F_i(\theta_t)] \right\| \Bigg] \\ \leq \eta_t \sum_{e=0}^{k-1} \Bigg(\frac{\rho_{\min}}{4} \mathbb{E}[\|\nabla F(\theta_t)\|^2] \\ + \frac{1}{\rho_{\min}} \mathbb{E}\left[\left\| \sum_{i\in S_t} w_i [\nabla F_i(\theta_{t,i}^{(e)}) - \nabla F_i(\theta_t)] \right\|^2 \right] \Bigg).\end{aligned}$$

By Jensen's inequality and L-smoothness (Assumption 1):

$$\begin{aligned}\left\| \sum_{i\in S_t} w_i [\nabla F_i(\theta_{t,i}^{(e)}) - \nabla F_i(\theta_t)] \right\|^2 \\ \leq \sum_{i\in S_t} w_i \left\| \nabla F_i(\theta_{t,i}^{(e)}) - \nabla F_i(\theta_t) \right\|^2 \\ \leq L^2 \sum_{i\in S_t} w_i \left\| \theta_{t,i}^{(e)} - \theta_t \right\|^2.\end{aligned}$$

From the update rule and Assumption 2:

$$\begin{aligned}\left\| \theta_{t,i}^{(e)} - \theta_t \right\|^2 = \eta_t^2 \left\| \sum_{s=0}^{e-1} M_t \odot g_{t,i}^{(s)} \right\|^2 \\ \leq \eta_t^2 e^2 G^2.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\left\| \sum_{i\in S_t} w_i [\nabla F_i(\theta_{t,i}^{(e)}) - \nabla F_i(\theta_t)] \right\|^2 \right] \\ \leq L^2 \eta_t^2 e^2 G^2.\end{aligned}$$

Then,

$$\begin{aligned}|T_{1c}| \leq \frac{\eta_t k \rho_{\min}}{4} \mathbb{E}[\|\nabla F(\theta_t)\|^2] \\ + \frac{L^2 G^2 \eta_t^3}{\rho_{\min}} \sum_{e=0}^{k-1} e^2.\end{aligned}$$

Since $\sum_{e=0}^{k-1} e^2 = \frac{(k-1)k(2k-1)}{6} \leq \frac{k^3}{3}$, we have:

$$|T_{1c}| \leq \frac{\eta_t k \rho_{\min}}{4} \mathbb{E}[\|\nabla F(\theta_t)\|^2] + \frac{L^2 G^2 \eta_t^3 k^3}{3\rho_{\min}}. \tag{16}$$

Combining T_{1a}, T_{1b}, and T_{1c}. From ((14)), ((15)), and ((16)):

$$\begin{aligned}T_1 \leq -\eta_t k \rho_{\min} \mathbb{E}[\|\nabla F(\theta_t)\|^2] \\ + \frac{\eta_t k \rho_{\min}}{2} \mathbb{E}[\|\nabla F(\theta_t)\|^2] \\ + \frac{\eta_t k \mu_t^2 \zeta^2}{\rho_{\min}} + \frac{L^2 G^2 \eta_t^3 k^3}{3\rho_{\min}}.\end{aligned}$$

By simplifying this inequality, we obtain:

$$\begin{aligned}T_1 \leq -\frac{\eta_t k \rho_{\min}}{2} \mathbb{E}[\|\nabla F(\theta_t)\|^2] \\ + \frac{\eta_t k \mu_t^2 \zeta^2}{\rho_{\min}} + \frac{L^2 G^2 \eta_t^3 k^3}{3\rho_{\min}}.\end{aligned} \tag{17}$$

Bounding T_2.

$$\begin{aligned}
&\mathbb{E}[\|\theta_{t+1}-\theta_t\|^2] \\
&= \eta_t^2 \mathbb{E}\left[\left\|\sum_{i\in S_t} w_i \sum_{e=0}^{k-1} M_t \odot g_{t,i}^{(e)}\right\|^2\right] \\
&\le \eta_t^2 \mathbb{E}\left[\left(\sum_{e=0}^{k-1}\left\|\sum_{i\in S_t} w_i M_t \odot g_{t,i}^{(e)}\right\|\right)^2\right] \\
&\le \eta_t^2 k \sum_{e=0}^{k-1} \mathbb{E}\left[\left\|\sum_{i\in S_t} w_i M_t \odot g_{t,i}^{(e)}\right\|^2\right].
\end{aligned}$$

For each e, we have:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{i\in S_t} w_i M_t \odot g_{t,i}^{(e)}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\sum_{i\in S_t} w_i M_t \odot \nabla F_i(\theta_{t,i}^{(e)})\right\|^2\right] \\
&+ \mathbb{E}\left[\sum_{i\in S_t} w_i^2 \|M_t \odot (g_{t,i}^{(e)} - \nabla F_i(\theta_{t,i}^{(e)}))\|^2\right] \\
&\le \mathbb{E}\left[\left(\sum_{i\in S_t} w_i \|M_t \odot \nabla F_i(\theta_{t,i}^{(e)})\|\right)^2\right] \\
&+ \sigma^2 \sum_{i\in S_t} w_i^2.
\end{aligned}$$

By Assumption 2 and Jensen's inequality, $\|\nabla F_i(\theta)\| \le G$, so:

$$\mathbb{E}\left[\left(\sum_{i\in S_t} w_i \|M_t \odot \nabla F_i(\theta_{t,i}^{(e)})\|\right)^2\right] \le G^2.$$

Let $\omega = \max_{S_t} \sum_{i\in S_t} w_i^2$. Then:

$$\mathbb{E}\left[\left\|\sum_{i\in S_t} w_i M_t \odot g_{t,i}^{(e)}\right\|^2\right] \le G^2 + \sigma^2\omega.$$

Therefore:

$$\mathbb{E}[\|\theta_{t+1}-\theta_t\|^2] \le \eta_t^2 k^2 (G^2 + \sigma^2\omega).$$

$$T_2 \le \frac{L}{2}\eta_t^2 k^2 (G^2 + \sigma^2\omega). \tag{18}$$

One-Round Recursion. Combining ((17)) and ((18)):

$$\begin{aligned}
\mathbb{E}[F(\theta_{t+1})] \le \mathbb{E}[F(\theta_t)] &- \frac{\eta_t k \rho_{\min}}{2}\mathbb{E}[\|\nabla F(\theta_t)\|^2] \\
&+ \frac{\eta_t k \mu_t^2 \zeta^2}{\rho_{\min}} + \frac{L^2 G^2 \eta_t^3 k^3}{3\rho_{\min}} \\
&+ \frac{L}{2}\eta_t^2 k^2 (G^2 + \sigma^2\omega).
\end{aligned}$$

Telescoping and Convergence Analysis. Assume a constant learning rate $\eta_t = \eta$ for all t. Summing from $t = 0$ to $T-1$:

$$\begin{aligned}
&\sum_{t=0}^{T-1} \frac{\eta k \rho_{\min}}{2}\mathbb{E}[\|\nabla F(\theta_t)\|^2] \\
&\le F(\theta_0) - F^* \\
&+ \sum_{t=0}^{T-1}\left(\frac{\eta k \mu_t^2 \zeta^2}{\rho_{\min}} + \frac{L^2 G^2 \eta^3 k^3}{3\rho_{\min}}\right. \\
&\left.+ \frac{L}{2}\eta^2 k^2 (G^2 + \sigma^2\omega)\right).
\end{aligned} \tag{19}$$

Dividing both sides by $\frac{\eta k \rho_{\min} T}{2}$:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\theta_t)\|^2] \le &\frac{2(F(\theta_0)-F^*)}{\eta k \rho_{\min} T} \\
&+ \frac{2\bar{\mu}^2\zeta^2}{\rho_{\min}^2} \\
&+ \frac{2L^2 G^2 \eta^2 k^2}{3\rho_{\min}^2} \\
&+ \frac{L\eta k (G^2 + \sigma^2\omega)}{\rho_{\min}},
\end{aligned} \tag{20}$$

where $\bar{\mu}^2 = \frac{1}{T}\sum_{t=0}^{T-1}\mu_t^2$.

Choose $\eta = \frac{1}{\sqrt{T}}$:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\theta_t)\|^2] \le &\frac{2(F(\theta_0)-F^*)}{k\rho_{\min}\sqrt{T}} \\
&+ \frac{2\bar{\mu}^2\zeta^2}{\rho_{\min}^2} \\
&+ \frac{Lk(G^2 + \sigma^2\omega)}{\rho_{\min}\sqrt{T}} \\
&+ \frac{2L^2 G^2 k^2}{3\rho_{\min}^2 T}.
\end{aligned} \tag{21}$$

D.3 Interpretation

As $T \to \infty$, the average squared gradient norm is bounded by:

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\theta_t)\|^2] \le \frac{2\bar{\mu}^2\zeta^2}{\rho_{\min}^2}.$$

This shows that the algorithm converges to a neighborhood of a stationary point, with the size of the neighborhood determined by the client heterogeneity ζ^2, the average mask heterogeneity suppression factor $\bar{\mu}^2$, and the gradient-mask alignment $\rho_{\min}$.

E Proof of Theorem 2

E.1 Problem Setup

Consider a loss function $\ell(\theta)$ with parameters $\theta \in \mathbb{R}^d$. The gradient is $g(\theta) = \nabla \ell(\theta)$. We analyze the effect of freezing parameters with small update magnitudes.

Masking Operation. Define a binary mask $M \in \{0,1\}^d$ that selects parameters to freeze based on update magnitude:

$$M_j = \begin{cases} 0 & \text{if } |g_j| \text{ is among the smallest } \tau \cdot d \\ 1 & \text{otherwise} \end{cases}$$

where $\tau \in (0,1)$ is the freezing ratio. Let $g_A = M \odot g$ (active gradients) and $g_F = (1-M) \odot g$ (frozen gradients).

E.2 Update Comparison

Full gradient update:

$$\theta_{\text{full}} = \theta - \eta_t g.$$

Masked update:

$$\theta_{\text{mask}} = \theta - \eta_t M \odot g = \theta - \eta_t g_A.$$

E.3 Loss Change Difference

Using first-order Taylor expansion:

$$\ell(\theta') \approx \ell(\theta) + \langle g, \theta' - \theta \rangle.$$

Loss after full update:

$$\ell(\theta_{\text{full}}) \approx \ell(\theta) - \eta_t \langle g, g \rangle = \ell(\theta) - \eta_t \|g\|^2.$$

Loss after masked update:

$$\ell(\theta_{\text{mask}}) \approx \ell(\theta) - \eta_t \langle g, g_A \rangle = \ell(\theta) - \eta_t \|g_A\|^2.$$

Difference in loss change:

$$\begin{aligned} \Delta \ell_{\text{diff}} &= \ell(\theta_{\text{mask}}) - \ell(\theta_{\text{full}}) \\ &\approx \eta_t (\|g\|^2 - \|g_A\|^2) \\ &= \eta_t \|g_F\|^2. \end{aligned} \tag{22}$$

E.4 Relative Importance of Frozen Parameters

Define the relative importance of frozen parameters:

$$R_F := \frac{\|g_F\|^2}{\|g\|^2} = \frac{\sum_{j \in F} g_j^2}{\sum_{j=1}^d g_j^2} \in [0,1].$$

Then:

$$\Delta \ell_{\text{diff}} \approx \eta_t \|g\|^2 R_F.$$

If R_F is small, the impact is small.

F Extended Hyperparameter Analysis Results

We show the sentivity of β on NLU tasks in Figure 7 through Figure 13. In most cases, model performance remains stable when β falls within the range of 0.02 to 0.07. This observation provides further evidence for our hypothesis that a subset of LoRA layers contributes minimally to overall model performance and freezing these layers does not significantly impact the original model performance.

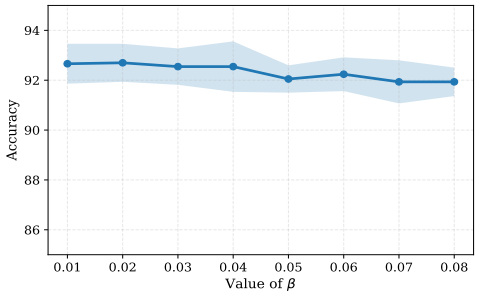


Figure 7: Sensitivity of β on SST-2.

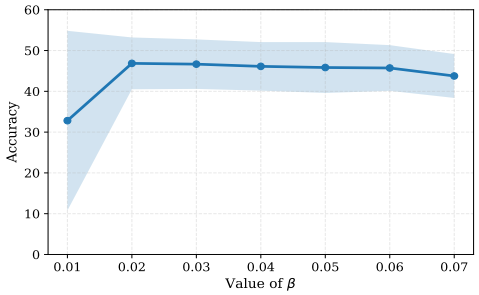


Figure 8: Sensitivity of β on COLA.

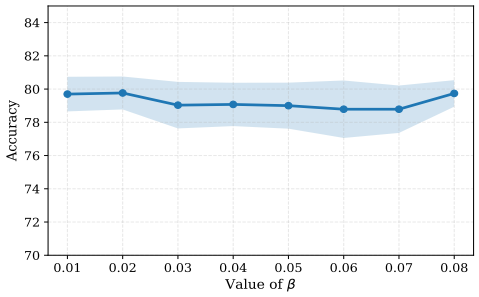


Figure 9: Sensitivity of β on MNLI (matched).

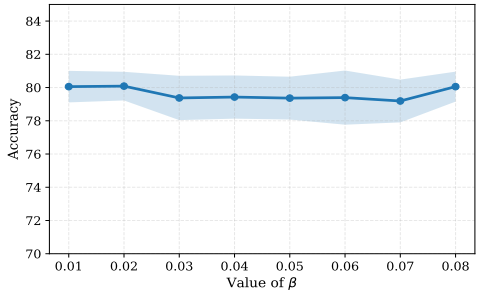


Figure 10: Sensitivity of β on MNLI (mismatched).

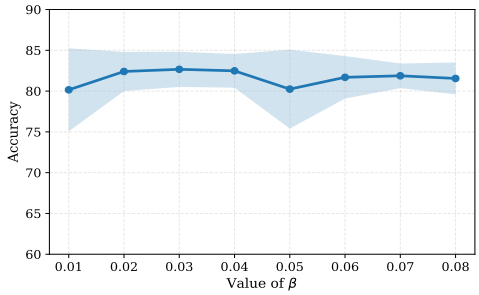


Figure 11: Sensitivity of β on QQP.

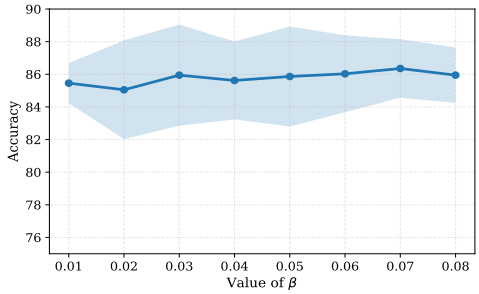


Figure 12: Sensitivity of β on MRPC.

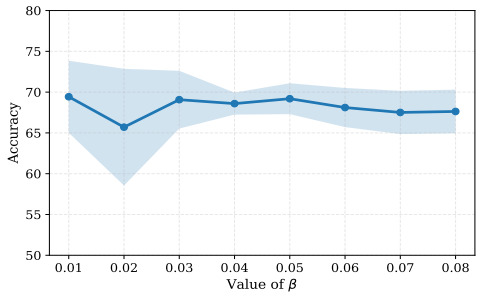


Figure 13: Sensitivity of β on RTE.