
FLoRA: Federated Fine-Tuning Large Language Models with Heterogeneous Low-Rank Adaptations

Ziyao Wang¹, Zheyu Shen¹, Yexiao He¹, Guoheng Sun¹ Hongyi Wang^{2,3}

Lingjuan Lyu⁴, Ang Li¹

1. University of Maryland, College Park

2. Rutgers University 3. GenBio.ai 4. Sony Research

{ziyaow, zyshen, yexiaohe, ghsun, angliice}@umd.edu,
hongyi.wang.001@rutgers.edu, lingjuansmile@gmail.com

Abstract

The rapid development of Large Language Models (LLMs) has been pivotal in advancing AI, with pre-trained LLMs being adaptable to diverse downstream tasks through fine-tuning. Federated learning (FL) further enhances fine-tuning in a privacy-aware manner by utilizing clients' local data through in-situ computation, eliminating the need for data movement. However, fine-tuning LLMs, given their massive scale of parameters, poses challenges for clients with constrained and heterogeneous resources in FL. Previous methods employed low-rank adaptation (LoRA) for efficient federated fine-tuning but utilized traditional FL aggregation strategies on LoRA adapters. These approaches led to mathematically inaccurate *aggregation noise*, reducing fine-tuning effectiveness and failing to address heterogeneous LoRAs. In this work, we first highlight the mathematical incorrectness of LoRA aggregation in existing federated fine-tuning methods. We introduce a new approach called FLoRA that enables federated fine-tuning on heterogeneous LoRA adapters across clients through a novel stacking-based aggregation method. Our approach is noise-free and seamlessly supports heterogeneous LoRA adapters. Extensive experiments demonstrate FLoRA's superior performance in both homogeneous and heterogeneous settings, surpassing state-of-the-art methods. We envision this work as a milestone for efficient, privacy-preserving, and accurate federated fine-tuning of LLMs. Our code is available at <https://github.com/ATP-1010/FederatedLLM>.

1 Introduction

The Large Language Models (LLMs) have shown remarkable performance on various tasks, such as chatbots [1], virtual assistants [4], search engines [11], and healthcare [20; 18]. However, adapting pre-trained LLMs (*e.g.*, Llama 2 [22]) to downstream tasks requires tremendous computation resources to fine-tune all the model parameters. To mitigate this issue, a variety of parameter-efficient fine-tuning (PEFT) methods have been proposed. One of the most widely used PEFT methods is low-rank adaptation (LoRA) [10]. As shown in the top of Figure 1, LoRA adds a parallel branch of trainable adapters \mathbf{A} and \mathbf{B} to compute the model update $\Delta\mathbf{W}$, where the ranks of \mathbf{A} and \mathbf{B} are much smaller than the pre-trained model parameter \mathbf{W} . When applying LoRA for fine-tuning, only \mathbf{A} and \mathbf{B} are updated while the entire \mathbf{W} is frozen, thereby significantly reducing the GPU memory consumption.

Fine-tuning LLMs requires ample data for adaptation to specific downstream tasks [13; 7]. Often, this data is dispersed across a multitude of devices, raising privacy concerns. For instance, aggregating medical data from hospitals for centralized LLM fine-tuning poses significant challenges. Consequently, to facilitate fine-tuning without compromising private data, federated learning (FL) becomes essential, enabling LLM fine-tuning across distributed clients while preserving data privacy [16; 27; 33; 24]. In this work, we focus on federated fine-tuning, enabling distributed clients to collaboratively fine-tune LLMs for adaptation to downstream tasks while preserving data privacy.

Prior work, FedIT, proposed a federated fine-tuning method [28], integrating LoRA with FedAvg [16]. In each FL round of FedIT, clients fine-tune LoRA modules using their local data and then send the fine-tuned modules to the server. The server averages all the local LoRA modules to obtain a global LoRA. Since only the weights of the LoRA modules are fine-tuned and communicated, FedIT effectively reduces both computation and communication costs.

However, FedIT faces two key issues. **First, the naive averaging of local LoRA modules in FedIT introduces noise to the global model update.** Specifically, FedIT averages local \mathbf{A} and \mathbf{B} independently, which introduces mathematical errors to the global LoRA. In short,

The cause of aggregation noise:

$$\underbrace{\sum \mathbf{A} \times \sum \mathbf{B}}_{\text{FedIT}} \neq \underbrace{\sum \mathbf{A} \times \mathbf{B}}_{\text{mathematically correct}} .$$

We will elaborate on this issue in Section 2 with theoretical analysis. Such an inaccurate aggregation will hinder convergence, leading to higher fine-tuning costs. **Second**, due to the heterogeneous data distribution [31; 12] and heterogeneous hardware resources, clients need to adapt LoRA ranks [30] according to the system and data heterogeneity. However, **FedIT cannot aggregate local LoRAs with heterogeneous ranks.**

In this work, we present FLORA, an aggregation-noise-free federated fine-tuning method that supports heterogeneous LoRAs. Specifically, as shown in Figure 2, we propose to **stack** the local LoRA modules \mathbf{A}_k and \mathbf{B}_k separately to construct the global LoRA modules \mathbf{A} and \mathbf{B} , where \mathbf{A}_k and \mathbf{B}_k denote the corresponding LoRA modules on the k -th client. This stacking method is theoretically proven to be accurate for the aggregation of local LoRA modules (Section 3.1). Additionally, it can naturally accommodate heterogeneous LoRA settings (Section 3.2), since stacking does not require the local LoRA modules to have identical ranks across clients. The noise-free aggregation of FLORA accelerates convergence, which will in turn improve the overall computation and communication efficiency of federated fine-tuning. Furthermore, FLORA can effectively cater to heterogeneous data and computational resources across clients, where heterogeneous ranks are applied. Our key contributions are summarized as follows:

- We propose FLORA, a federated fine-tuning algorithm based on LoRA that can perform noise-free aggregation of local LoRA modules. Theoretical analysis shows that FLORA eliminates the meaningless intermediate term in the global model update, leading to faster convergence and improved performance.
- The proposed stacking mechanism for aggregating LoRA modules supports heterogeneous LoRA ranks across clients, accommodating data and system heterogeneity in realistic settings. This encourages the broader participation of clients with heterogeneous data and resources in federated fine-tuning.
- We use FLORA to fine-tune LLaMA, Llama2 [21] and TinyLlama [29] on four benchmarks for two downstream tasks. Results show that FLORA surpasses state-of-the-art methods for both homogeneous and heterogeneous settings.

2 Preliminaries

Fine-tuning LLMs with LoRA. LoRA [10] uses two decomposed low-rank matrices to represent the update of the target module:

$$\mathbf{W}' = \mathbf{W} + \Delta \mathbf{W} = \mathbf{W} + \mathbf{BA}, \quad (1)$$

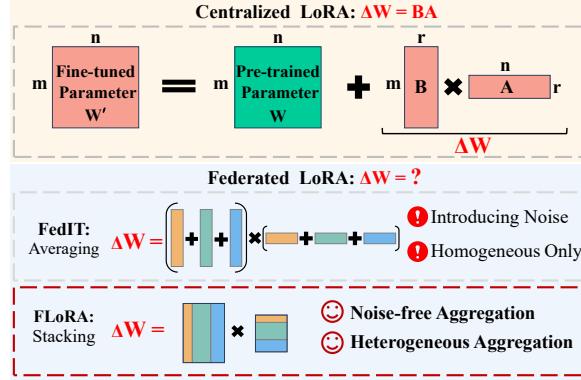


Figure 1: The overview of LoRA, FedIT, and our FLORA. The top row shows how LoRA updates the model in centralized fine-tuning. The middle and bottom rows show the global model updating strategies in FedIT and our FLORA respectively.

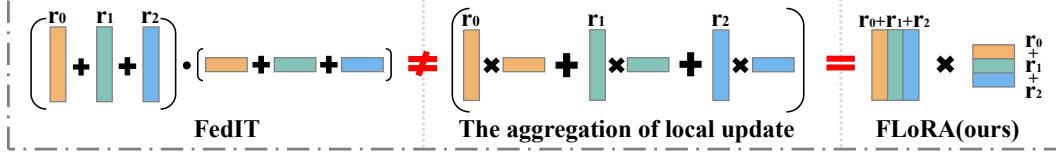


Figure 2: Module stacking in FLORA is a noise-free aggregation for LoRA, while the module averaging in FedIT cannot accurately aggregate the local updates.

where $\mathbf{W} \in \mathbb{R}^{m \times n}$ and $\mathbf{W}' \in \mathbb{R}^{m \times n}$ denote the pre-trained and fine-tuned parameters of target modules (*e.g.*, attention modules), respectively. \mathbf{A} and \mathbf{B} are low-rank decomposition of $\Delta\mathbf{W}$, where $\mathbf{A} \in \mathbb{R}^{r \times n}, \mathbf{B} \in \mathbb{R}^{m \times r}$, such that $\Delta\mathbf{W} = \mathbf{BA}$ with the identical dimensions as \mathbf{W} and \mathbf{W}' . The rank of LoRA, denoted by r , is typically significantly smaller than m and n , leading to dramatic parameter reduction of $\Delta\mathbf{W}$. During the fine-tuning phase, LoRA optimizes matrices \mathbf{A} and \mathbf{B} instead of directly updating \mathbf{W} , thus achieving substantial savings in GPU memory usage. For example, in the context of the LLaMA-7B [21], the original dimension of attention modules is $\mathbf{W} \in \mathbb{R}^{4096 \times 4096}$, setting the LoRA rank to 16 reduces the decomposed matrices to $\mathbf{A} \in \mathbb{R}^{16 \times 4096}$ and $\mathbf{B} \in \mathbb{R}^{4096 \times 16}$. This approach decreases the number of trainable parameters to merely 0.78% of the entire parameter of the pre-trained model, offering a significant GPU memory footprint reduction.

FedIT: Averaging Homogeneous LoRA. The most widely used FL algorithm, *i.e.*, FedAvg [16], aggregates all the local model updates by weighted averaging to update the global model in each communication round:

$$\mathbf{W}' = \mathbf{W} + \sum_{k=1}^K p_k \Delta\mathbf{W}_k = \mathbf{W} + \Delta\mathbf{W} \quad (2)$$

where \mathbf{W}' and \mathbf{W} denote the global model parameters before and after a communication round. $\Delta\mathbf{W}_k$ represents the local model update from the k -th client, with p_k being the corresponding scaling factor that is typically weighted by the local data size, and $\Delta\mathbf{W}$ represents the global model update.

FedIT [28] directly integrates FedAvg with LoRA to enable federated fine-tuning, where each client fine-tunes LoRA modules with a homogeneous rank. Specifically, the clients download the pre-trained LLM from the server, locally initialize and fine-tune the LoRA modules, and then send the updated LoRA modules to the server. The server updates the global LoRA modules \mathbf{A} and \mathbf{B} by independently applying weighted averaging across all local modules \mathbf{A}_k and \mathbf{B}_k :

$$\mathbf{A} = \sum_{k=1}^K p_k \mathbf{A}_k, \quad \mathbf{B} = \sum_{i=0}^K p_k \mathbf{B}_k. \quad (3)$$

This aggregation of FedIT is almost the same as FedAvg except that only the LoRA modules are trained and communicated. However, such a naive aggregation introduces additional issues for federated fine-tuning. First, each single module \mathbf{A} or \mathbf{B} is not the model update, and only \mathbf{BA} represents the model update. Thus, averaging \mathbf{A}_k and \mathbf{B}_k *independently* to compute the aggregated gradients will introduce noises to the global model update. Here we use a simple example to explain how the noise is generated, and we assume that two clients are applying FedIT to perform federated fine-tuning. In a communication round, the two clients train $\mathbf{A}_0, \mathbf{B}_0$ and $\mathbf{A}_1, \mathbf{B}_1$ respectively. The local model updates $\Delta\mathbf{W}_0$ and $\Delta\mathbf{W}_1$ are the product of corresponding LoRA modules:

$$\Delta\mathbf{W}_k = \mathbf{B}_k \mathbf{A}_k, k \in \{0, 1\}. \quad (4)$$

According to Equation 2, the expected global model update $\Delta\mathbf{W}$ can be obtained by weighted averaging $\Delta\mathbf{W}_0$ and $\Delta\mathbf{W}_1$:

$$\Delta\mathbf{W} = p_0 \Delta\mathbf{W}_0 + p_1 \Delta\mathbf{W}_1 = p_0 \mathbf{B}_0 \mathbf{A}_0 + p_1 \mathbf{B}_1 \mathbf{A}_1. \quad (5)$$

However, according to Equation 3, FedIT aggregates \mathbf{A} and \mathbf{B} independently:

$$\begin{aligned} \Delta\mathbf{W} &= \mathbf{BA} = (p_0 \mathbf{B}_0 + p_1 \mathbf{B}_1)(p_0 \mathbf{A}_0 + p_1 \mathbf{A}_1) \\ &= p_0^2 \mathbf{B}_0 \mathbf{A}_0 + p_1^2 \mathbf{B}_1 \mathbf{A}_1 + \underline{p_0 p_1 (\mathbf{B}_0 \mathbf{A}_1 + \mathbf{B}_1 \mathbf{A}_0)}. \end{aligned} \quad (6)$$

The global model update in Equation 6 differs from the expected one in Equation 2, mainly due to the underlined intermediate term that is obtained by the cross-product of LoRA modules from

different clients. This intermediate term introduces unexpected noise in the model aggregation. As the number of clients increases, this noisy term becomes much larger than the real global updates, significantly slowing down the fine-tuning progress. In addition, FedIT applies the scaling factor p_k to both \mathbf{A}_k and \mathbf{B}_k , resulting in a p_k^2 coefficient for the local model update $\Delta \mathbf{W}_k$, exacerbating the error in LoRA aggregation. As Figure 2 illustrates, the averaging algorithm in FedIT is an inaccurate aggregation method, leading to slower convergence and higher computation costs.

The other deficiency of FedIT is that it cannot support aggregation on heterogeneous LoRA modules. The local data in FL may exhibit significant heterogeneity across clients [31; 12]. If a client configures a higher rank than the actual one required by the local data complexity, this may result in overfitting. Conversely, if the rank is too small, it may lack the necessary generalization capacity to effectively learn from the local dataset (Figure 4). Moreover, the heterogeneous computational resource across clients also requires heterogeneous rank deployment, *e.g.*, clients with smaller memory can only afford to train LoRA modules with smaller ranks. AdaLoRA [30] has been proposed to adapt LoRA ranks based on available computation resources. Therefore, deploying heterogeneous ranks across clients is a pressing requirement for accommodation to data and system heterogeneity. However, according to Equation 3, FedIT is only able to aggregate LoRA modules with the homogeneous rank.

3 Proposed Method: FLoRA

3.1 Stacking-based Noise-free Aggregation

Motivated by the aforementioned problem, we propose a novel aggregation mechanism that accurately computes global model update $\Delta \mathbf{W}$ by aggregating local LoRA modules and effectively supports the heterogeneous LoRA. According to matrix multiplication principles and the model update rule in LoRA (*i.e.*, Equation 1), the element at position (x, y) of the model update $\Delta \mathbf{W}$ is computed as the sum of the products of corresponding elements from the x -th column of \mathbf{B} and the y -th row of \mathbf{A} :

$$\delta_{xy} = \sum_{i=0}^r a_{yi} b_{xi}, \quad (7)$$

where δ_{xy} represents the element at position (x, y) in $\Delta \mathbf{W}$. a_{yi}, b_{xi} are the elements at positions (y, i) and (x, i) in \mathbf{A} and \mathbf{B} , respectively. According to Equation 3.1, the model update in LoRA can be expressed as the sum of the products of the corresponding rows of \mathbf{A} and the columns of \mathbf{B} .

To illustrate this concept further, let us consider a simplified example where the dimensions of LoRA modules are given by $\mathbf{A} \in \mathbb{R}^{2 \times 3}$ and $\mathbf{B} \in \mathbb{R}^{3 \times 2}$. As described in Equation 8, \mathbf{A} and \mathbf{B} can be decomposed to two sub-matrices with rank $r = 1$, and the product of \mathbf{A} and \mathbf{B} then are computed as the sum of the products of two respective sub-matrices:

$$\mathbf{BA} = \begin{bmatrix} b_{00}, b_{01} \\ b_{10}, b_{11} \\ b_{20}, b_{21} \end{bmatrix} \cdot \begin{bmatrix} a_{00}, a_{10}, a_{20} \\ a_{01}, a_{11}, a_{21} \end{bmatrix} = \begin{bmatrix} b_{00} \\ b_{10} \\ b_{20} \end{bmatrix} \cdot [a_{00}, a_{10}, a_{20}] + \begin{bmatrix} b_{01} \\ b_{11} \\ b_{21} \end{bmatrix} \cdot [a_{01}, a_{11}, a_{21}] \quad (8)$$

To address the aggregation challenge from an alternative perspective, let us consider the scenario where we have multiple pairs of LoRA modules, $\mathbf{A}_k, \mathbf{B}_k$, optimized by the clients. Each pair satisfies the dimensions $\mathbf{A}_k \in \mathbb{R}^{r_k \times n}$ and $\mathbf{B}_k \in \mathbb{R}^{m \times r_k}$. Similar to Equation 8, the sum of the products of these module pairs is the product of the stacked modules, *i.e.*, $\sum_{k=1}^K \mathbf{B}_k \mathbf{A}_k = \mathbf{BA}$, where \mathbf{B} represents the *stacking* of all \mathbf{B}_k modules aligned through dimension m and \mathbf{A} is the *stacking* of all \mathbf{A}_k aligned through dimension n . Figure 2 visually illustrates this concept, where the orange, green, and blue rectangles symbolize $\mathbf{A}_k, \mathbf{B}_k$, and their respective products. The aggregation of three products mirrors the product of the stacked \mathbf{B} and \mathbf{A} from all \mathbf{B}_k and \mathbf{A}_k pairs trained by clients. This mechanism demonstrates that, in the context of federated fine-tuning, we can achieve a noise-free aggregation of local updates by simply stacking the local LoRA modules. This process also avoids transmitting the full model parameters, thus reducing communication costs.

To facilitate our discussion, we introduce the stacking operation symbolized by “ \oplus ” to denote the module aggregation as depicted in Figure 2. This operation is mathematically defined as:

$$\mathbf{A} = \mathbf{A}_0 \oplus \mathbf{A}_1 \oplus \mathbf{A}_2, \quad \mathbf{B} = \mathbf{B}_0 \oplus \mathbf{B}_1 \oplus \mathbf{B}_2, \quad (9)$$

$$\mathbf{A}_k \in \mathbb{R}^{r_k \times n}, \mathbf{A} \in \mathbb{R}^{(r_0+r_1+r_2) \times n}, \mathbf{B}_k \in \mathbb{R}^{m \times r_k}, \mathbf{B} \in \mathbb{R}^{m \times (r_0+r_1+r_2)}.$$

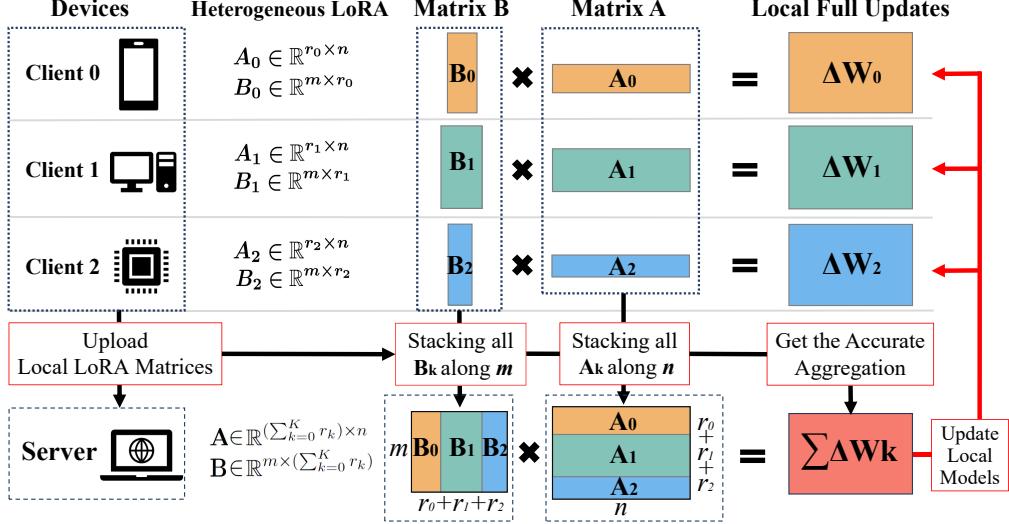


Figure 3: FLORA workflow. The local LoRA modules are initialized and optimized each round, and stacked by the server to obtain the global LoRA modules. The global modules are then sent back to clients to update local models.

In Equation 9, " \oplus " indicates that for \mathbf{A} , each subsequent module is vertically stacked below the preceding one, whereas for \mathbf{B} , each module is horizontally stacked to the right of the one before it.

We can now formalize our conclusion regarding the aggregation of LoRA modules. The sum of the products of K LoRA module pairs is equivalent to the product of their stacked matrices:

$$\sum_{k=0}^K \mathbf{B}_k \mathbf{A}_k = (\mathbf{B}_0 \oplus \dots \oplus \mathbf{B}_K)(\mathbf{A}_0 \oplus \dots \oplus \mathbf{A}_K) \quad (10)$$

This foundational principle will guide the design of FLORA, as it allows for the efficient and effective aggregation of local updates without the transmission of entire model parameters.

3.2 FLORA: Stacking-based Federated Fine-tuning for Heterogeneous LoRA

The stacking-based aggregation facilitates not only the accurate aggregation of LoRA modules but also inherently supports the heterogeneous LoRA ranks. This approach imposes no constraints on the ranks of each local LoRA module as long as each client fine-tunes the same pre-trained model, *i.e.*, they share the same dimension m and n .

By employing the stacking-based aggregation mechanism, we introduce FLORA, an approach designed to facilitate federated fine-tuning of LLMs with heterogeneous LoRA. Let us use a concrete example to illustrate the key steps of applying FLORA, where K heterogeneous clients are involved in fine-tuning an LLM, and the pre-trained parameters are denoted by \mathbf{W} .

Initialization. The server first disseminates the pre-trained model parameters \mathbf{W} to all K clients. Then, the clients initialize their local LoRA modules based on the complexity of local data and available local resources. The adaptation of LoRA ranks is beyond the scope of this paper, but existing work like AdaLoRA [30] can facilitate the rank adjustment.

Local Fine-tuning. Following initialization, the clients train their local LoRA modules with the local data for several iterations. Then, the clients send the local LoRA modules back to the server. Note that the clients initialize local LoRA modules *each* round before local fine-tuning.

Stacking-based LoRA Aggregation. Upon receiving the heterogeneous LoRA modules from participating clients, the server proceeds to aggregate them by stacking all \mathbf{B}_k and \mathbf{A}_k according to Equation 10, resulting in the global $\mathbf{A} \in \mathbb{R}^{(\sum_{k=0}^K r_k) \times n}$ and $\mathbf{B} \in \mathbb{R}^{m \times (\sum_{k=0}^K r_k)}$. The aggregation

process of FLORA can be described as follows:

$$\begin{aligned} \mathbf{A} &= p_0 \mathbf{A}_0 \oplus p_1 \mathbf{A}_1 \oplus \dots \oplus p_K \mathbf{A}_K, \quad \mathbf{B} = \mathbf{B}_0 \oplus \mathbf{B}_1 \oplus \mathbf{B}_2 \oplus \dots \oplus \mathbf{B}_K \\ \mathbf{A}_k &\in \mathbb{R}^{r_k \times n}, \mathbf{B}_k \in \mathbb{R}^{m \times r_k}, \quad \mathbf{A} \in \mathbb{R}^{(\sum_{k=0}^K r_k) \times n}, \mathbf{B} \in \mathbb{R}^{m \times (\sum_{k=0}^K r_k)}, \end{aligned} \quad (11)$$

where p_k represents the scaling factor for each local update, determined by the relative size of the local data to the global data:

$$p_k = \frac{\text{len}(D_k)}{\text{len}(\sum_{k=0}^K D_k)}. \quad (12)$$

Note that the scaling factor p_k should be only applied to one of \mathbf{A}_k and \mathbf{B}_k to avoid squaring the factor in the final model update \mathbf{BA} . This method ensures a noise-free aggregation mechanism as described in Equation 10.

Update Local Models. After each round of noise-free aggregation, the server redistributes the updated global LoRA modules \mathbf{A} and \mathbf{B} back to the clients. The clients then proceed to update the local models using \mathbf{BA} and continue the fine-tuning. Using the stacking approach, the dimensions of updated global LoRA modules \mathbf{A} and \mathbf{B} are larger than those of FedIT, potentially leading to larger communication overhead in each round. However, empirical observations indicate that federated fine-tuning typically requires only a limited number of communication rounds to achieve satisfactory results, as detailed in Section 4. In addition, it is important to note that the LoRA modules \mathbf{A} and \mathbf{B} constitute a small fraction of the overall size of the pre-trained model, which is distributed to clients during the initialization phase. Thus, the additional communication overhead of the stacking approach is negligible and does not significantly impact the efficiency of federated fine-tuning.

4 Experiments

The key features of FLORA are (i) noise-free aggregation and (ii) support for heterogeneous LoRA modules. In this section, we verify these key features across various LLM fine-tuning tasks. We first study the performance of FLoRA and compare it against FedIT under homogeneous settings to demonstrate the advantages of noise-free aggregation [28]. Then, we examine performance in a synthetic heterogeneous setup and compare FLORA with a vanilla *zero-padding* method. Finally, we conduct ablation studies on the scaling factor, the heterogeneity of LoRA ranks, and the extra communication overhead of FLoRA.

4.1 Experiment Setup

Models, Datasets and Experiment Settings. We employ three Llama-based models with different scales in our experiments: TinyLlama with 1.1 billion parameters [29], and the 7 billion parameter versions of Llama [21] and Llama2 [22], evaluating FLORA across different model capacities. Following the configurations in the original LoRA paper [10], the LoRA modules are applied to the self-attention layers only.

We use the Databricks-dolly-15k [28] instruction dataset, Alpaca dataset [19], and Wizard dataset [14] for the question-answering (QA) task, and Wizard and ShareGPT for the chat assistant task. We evaluate the federated fine-tuned models on MMLU [8] for the QA task and MT-bench [32] for the chat assistant task, respectively. We sample 10 clients uniformly at random following the non-IID setting in FedIT [28]. The other experimental configurations are elaborated in Appendix A.

Baselines. We compare FLoRA with four baselines. (1) **FedIT**: It is the SOTA federated fine-tuning method [28] that integrates LoRA with FedAvg. We only apply FedIT to homogeneous LoRA experiments as it does not support heterogeneous LoRA. (2) **Zero-Padding**: It is an approach that enables FedIT to support heterogeneous LoRA [3]. It extends all the heterogeneous local ranks to the maximum rank among the clients and pads their remaining parts by 0. (3) **Centralized Fine-tuning**: we compare FLORA with centralized LoRA with the same hyperparameters and configurations. (4) **Standalone**: the client fine-tunes the pre-trained model locally without federations.

Table 1: Comparison of FLORA with baselines on MMLU and MT-bench. "Homo" represents the settings with homogeneous LoRA ranks, and "Heter" denotes those with heterogeneous LoRA ranks.

Foundation model	Strategy	Fine-tuning algorithm	MMLU			MT-bench	
			Dolly	Alpaca	Wizard	Wizard	ShareGPT
TinyLlama	Centralized	LoRA	27.99	28.03	29.13	2.34	2.79
	Homo	FedIT	16.35	30.02	42.51	2.92	2.55
	Heter	FLoRA	30.80	31.92	43.87	3.13	2.77
Llama	Centralized	LoRA	35.91	29.18	31.68	4.38	3.99
	Homo	FedIT	29.67	29.41	33.43	3.07	3.73
	Heter	FLoRA	30.99	29.85	34.26	4.21	3.93

4.2 Experiment Results

Homogeneous LoRA. We first evaluate the performance of FLORA with homogeneous LoRA. Specifically, all the clients share the identical LoRA rank of 16. As Table 1 depicts, FLORA achieves consistently better performance than FedIT across all the evaluated models and tasks. This is evident in the MT-bench scores for both TinyLlama and Llama models, where FLORA’s performance exceeds that of FedIT by at least 0.2. A notable example is the MT-bench score for the Llama model fine-tuned with Wizard dataset, where FLORA scores 4.21, surpassing FedIT’s 3.07. On the MMLU test set, FLORA outperforms FedIT in all the settings. For example, considering the TinyLlama model fine-tuned with Dolly, FLORA nearly doubles the accuracy achieved by FedIT. While FedIT occasionally matches the performance of FLORA, as observed with the Alpaca dataset on MMLU, the performance gap is marginal. Interestingly, in several scenarios, the performance of FLORA not only outpaces FedIT but also exceeds the performance achieved by the centralized fine-tuning. This phenomenon, observed in the TinyLlama model fine-tuned with the Alpaca and Wizard datasets, suggests that the smaller data volume on clients for federated fine-tuning may help mitigate overfitting, thereby enhancing model generalization. The experiment results of the Llama2 model are presented in Appendix A, which reveal the same trend as that in TinyLlama and Llama. The consistent observations across the three models demonstrate that FLORA consistently outperforms FedIT in the homogeneous LoRA setting.

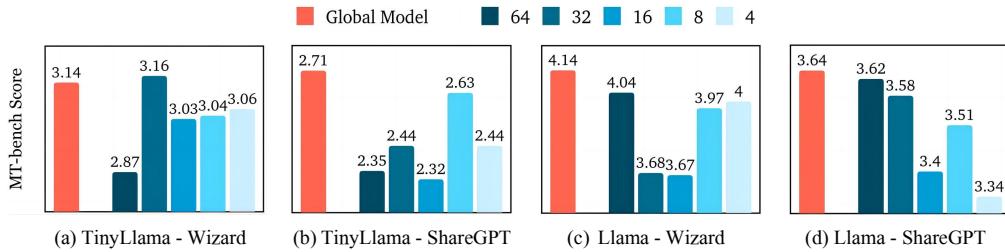


Figure 4: Standalone experiment results. The red bars represent the global model performance and the blue bars represent the local model performance with varying LoRA ranks.

Heterogeneous LoRA. Compared with FedIT, a distinctive strength of FLORA lies in its inherent capability to accommodate heterogeneous LoRA configurations. In the heterogeneous LoRA settings, we apply varied local LoRA ranks, *i.e.*, [64, 32, 16, 16, 8, 8, 4, 4, 4, 4], to 10 clients, simulating a realistic scenario where clients have heterogeneous computational resources. As Table 1 and Table 4 illustrate, FLORA not only adapts to heterogeneous ranks without performance degradation but also maintains consistency with the results observed in most homogeneous settings. This contrasts sharply with the performance of FedIT, where the application of zero-padding significantly degrades its performance on MMLU and MT-bench. It reveals that zero-padding exacerbates FedIT’s inherent noise issues in the aggregation process, posing significant challenges in managing fine-tuning performance. For example, by applying the zero-padding method, the MMLU accuracy of Llama model fine-tuned with Alpaca dataset dramatically drops to 7.97%. The results demonstrate that FLORA

not only accommodates heterogeneous LoRA ranks effectively but also sustains robust training performance compared to baseline methods. It efficiently facilitates the participation of devices with varied computational capacities in heterogeneous federated fine-tuning tasks. Additionally, FLORA can be seamlessly integrated with AdaLoRA [30], which dynamically adjusts the LoRA rank on the clients, the results are presented in Appendix A.

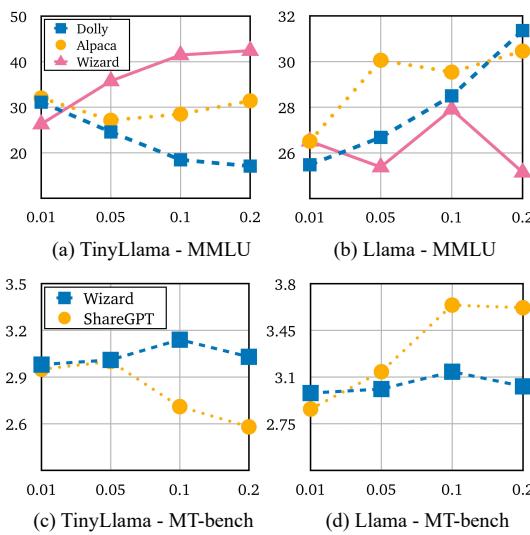


Figure 5: The impact of the scaling factor on FLORA. The x-axis is the scaling factor, and the y-axis represents the MMLU accuracy for (a)-(b) and the MT-bench score for (c)-(d). The results of Llama2 are in Appendix A.

plying FLORA to Dolly and Alpaca shows the optimal performance with a scaling factor of 0.2. These observations suggest that the choice of an appropriate scaling factor is highly dependent on specific datasets and model characteristics, underscoring the necessity for a tailored approach in federated fine-tuning.

The Impact of Heterogeneous LoRA Ranks. Although the above results demonstrate FLORA effectively enables the federated fine-tuning with heterogeneous LoRA, it is worth further investigating how the federated fine-tuning improves the local models with various ranks. Motivated by this, we evaluate MT-bench scores for local models with LoRA ranks of 64, 32, 16, 8, and 4, presenting the results in Figure 4. Global model scores are shown in red bars, while local models are in blue, with deeper shades indicating higher ranks. The results show that the global model outperforms all local models, except for a case with the TinyLlama model fine-tuned on the Wizard dataset, where the client with rank 32 slightly exceeds the global model. This demonstrates FLORA’s ability to synthesize knowledge from diverse clients effectively.

Regarding the LoRA rank’s impact, a rank of 8 consistently yields strong performance across various models and datasets. However, performance diverges at extreme ranks; for instance, the TinyLlama model fine-tuned on Wizzard with the LoRA rank of 64 underperforms the ones with smaller ranks, but the Llama model with the rank of 64 excels the counterparts with smaller ranks. This also demonstrates the heterogeneous rank deployment across clients is a realistic setting. These observations suggest a potential positive correlation between optimal LoRA rank and model capacity, motivating further exploration in future research.

5 Discussion

The Communication Overhead of FLORA. As discussed in Section 3, the server needs to send global LoRA modules to the clients in FLORA, potentially raising concerns about increased commu-

nication overhead. To quantify this, we compare the communicated parameters of full fine-tuning, FedIT, and FLORA over three communication rounds.

As Figure 6 shows, although FLORA transmits slightly more parameters than FedIT, it still significantly reduces the overhead compared to full fine-tuning. This is due to the fact that the primary communication load in federated fine-tuning, especially with large models, is the initial full model parameter transmission. Subsequent rounds primarily involve smaller updates (e.g., LoRA matrices). Thus, even though FLORA introduces additional communication for these updates, the overall impact on total communication costs remains marginal, making it comparable to FedIT’s costs. Despite the minor communication increase compared to FedIT, FLORA enhances fine-tuning effectiveness and supports heterogeneous LoRA ranks, making it a preferable solution in federated fine-tuning.

The Privacy Preservation of FLORA. The requirement of FLORA to stack the LoRA modules uploaded by all clients introduces a potential privacy concern, as malicious clients might infer the LoRA matrices of other clients through the global LoRA modules sent from the server. To address this issue, we split all the local LoRA modules into sub-modules with rank=1 and then stack the sub-modules together in random order. This approach prevents malicious clients from recovering the local LoRA modules from other clients. In addition, FLORA is also compatible with standard privacy mechanisms such as encryption [2] and differential privacy (DP) [25], aligning it with the privacy-preserving nature of FL.

6 Related Work

Parameter-efficient Fine-tuning of LLMs. Parameter-efficient fine-tuning (PEFT) aims to reduce the number of trainable parameters. BitFit [26] fine-tunes only the biases while achieving similar accuracy with full fine-tuning. Other works such as [9] and [17] apply transfer learning that adds pre-trained adapter layers between transformer blocks. LoRA [10] adopts the product of two low-rank matrices to represent the gradient in full fine-tuning, which achieves memory-efficient fine-tuning. AdaLoRA [30] optimizes LoRA by adaptively allocating the parameter budget, which enhances the flexibility of LoRA. There are also many works regarding optimizing LoRA in various aspects [5; 6; 15].

Federated Fine-tuning of LLMs. Federated fine-tuning aims to extract knowledge from multiple on-device datasets while preserving data privacy. FedIT [28] leverages the FL framework for fine-tuning LLMs. It uses LoRA as the local fine-tuning strategy. However, concerns related to the deficiency in supporting heterogeneous LoRA limit its utilization. [3] tries to solve this problem by zero-padding the local LoRA modules. However, this padding process causes additional computing overhead. Besides, it separately averages **A** and **B** modules, introducing noise to the global model.

7 Conclusion

In this work, we identified the limitations in current federated fine-tuning methods (*e.g.*, FedIT), and the challenges of applying federated fine-tuning in realistic settings, *i.e.*, the heterogeneous LoRA ranks across clients. To overcome these practical challenges and broaden the applicability of federated fine-tuning, we introduced FLORA to enable the accurate aggregation on heterogeneous LoRA modules using the proposed stack-based LoRA aggregation mechanism. Our extensive experiments demonstrate that FLORA outperforms the SOTA method in both homogeneous and heterogeneous LoRA settings. Moreover, our inspiring results provide valuable insights for future research in the federated fine-tuning of large language models in a lightweight and accurate manner.

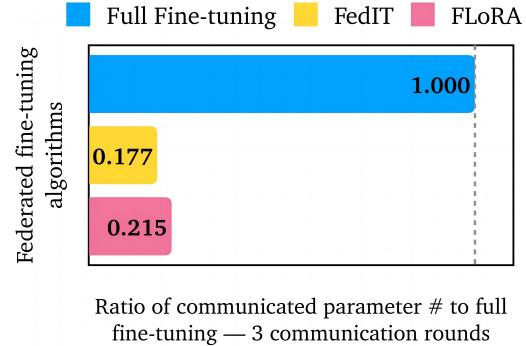


Figure 6: The ratio of communicated parameter numbers to full fine-tuning.

References

- [1] Desirée Bill and Theodor Eriksson. Fine-tuning a llm using reinforcement learning from human feedback for a therapy chatbot application, 2023.
- [2] Yansong Chang, Kai Zhang, Junqing Gong, and Haifeng Qian. Privacy-preserving federated learning via functional encryption, revisited. *IEEE Transactions on Information Forensics and Security*, 18:1855–1869, 2023.
- [3] Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, Matt Barnes, and Gauri Joshi. Heterogeneous lora for federated fine-tuning of on-device foundation models. In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*, 2023.
- [4] Xin Luna Dong, Seungwhan Moon, Yifan Ethan Xu, Kshitiz Malik, and Zhou Yu. Towards next-generation intelligent assistants leveraging llm techniques. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5792–5793, 2023.
- [5] Shwai He, Liang Ding, Daize Dong, Miao Zhang, and Dacheng Tao. Sparseadapter: An easy approach for improving the parameter-efficiency of adapters. *arXiv preprint arXiv:2210.04284*, 2022.
- [6] Shwai He, Run-Ze Fan, Liang Ding, Li Shen, Tianyi Zhou, and Dacheng Tao. Mera: Merging pretrained adapters for few-shot learning. *arXiv preprint arXiv:2308.15982*, 2023.
- [7] Yexiao He, Ziyao Wang, Zheyu Shen, Guoheng Sun, Yucong Dai, Yongkai Wu, Hongyi Wang, and Ang Li. Shed: Shapley-based automated dataset refinement for instruction fine-tuning. *arXiv preprint arXiv:2405.00705*, 2024.
- [8] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020.
- [9] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.
- [10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [11] Dominique Kelly, Yimin Chen, Sarah E Cornwell, Nicole S Delellis, Alex Mayhew, Sodiq Onaolapo, and Victoria L Rubin. Bing chat: The future of search engines? *Proceedings of the Association for Information Science and Technology*, 60(1):1007–1009, 2023.
- [12] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.
- [13] Fei Liu, Yiming Yao, Ping Guo, Zhiyuan Yang, Xi Lin, Xialiang Tong, Mingxuan Yuan, Zhichao Lu, Zhenkun Wang, and Qingfu Zhang. A systematic survey on large language models for algorithm design. *arXiv preprint arXiv:2410.14716*, 2024.
- [14] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*, 2023.
- [15] Yuren Mao, Yuhang Ge, Yijiang Fan, Wenyi Xu, Yu Mi, Zhonghao Hu, and Yunjun Gao. A survey on lora of large language models. *arXiv preprint arXiv:2407.11046*, 2024.
- [16] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

- [17] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020.
- [18] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- [19] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- [20] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.
- [21] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [22] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [23] Ziyao Wang, Jianyu Wang, and Ang Li. Fedhyper: A universal and robust learning rate scheduler for federated learning with hypergradient descent. *arXiv preprint arXiv:2310.03156*, 2023.
- [24] Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. Openfedllm: Training large language models on decentralized private data via federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6137–6147, 2024.
- [25] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.
- [26] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021.
- [27] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021.
- [28] Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Guoyin Wang, and Yiran Chen. Towards building the federated gpt: Federated instruction tuning. *arXiv preprint arXiv:2305.05644*, 2023.
- [29] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model, 2024.
- [30] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023.
- [31] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.
- [32] 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. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.
- [33] Weiming Zhuang, Jian Xu, Chen Chen, Jingtao Li, and Lingjuan Lyu. Coala: A practical and vision-centric federated learning platform. *arXiv preprint arXiv:2407.16560*, 2024.

A Additional Experiments and Setup Details

A.1 Environments, Datasets and Metric

Computer Resources. We use a 256GB AMD EPYC 7763 64-Core Processor on Linux v4.18.0 to run the experiments. For LoRA fine-tuning on all the models, we use 4 NVIDIA RTX A6000 GPUs.

Dolly dataset. The Dolly dataset is an open-source dataset with 15k text samples generated by Databricks employees. The topics include brainstorming, classification, closed QA, generation, information extraction, open QA, and summarization [28].

Alpaca dataset. The Alpaca dataset contains 52K instruction-following data used for fine-tuning the Alpaca model [19]. This dataset is believed to be diverse enough for fine-tuning LLMs.

Wizard dataset. The Wizard dataset we use is the training data of the WizardLM model. It includes 70k pairs of instructions and outputs. The Wizard dataset generally features more complex instructions compared to the other datasets. Its fine-tuning results are typically better, which has been confirmed by our experiments, especially those evaluated by the MT-bench scores.

ShareGPT dataset. The ShareGPT dataset is a collection of approximately 52,000 conversations scraped via the ShareGPT API. The conversations in ShareGPT include both user prompts and responses from ChatGPT. In our experiments, we split the conversation dataset into question-answering pairs.

MMLU test set. The MMLU dataset is a widely used question-and-answer dataset in LLM fine-tuning. It has 14,024 questions in 57 different subjects, which can evaluate the logical reasoning capabilities of LLMs. We selected 1444 samples from the dataset for a quick and comprehensive evaluation.

MT-bench evaluation. MT-bench is a set of challenging multi-turn open-ended questions for evaluating chat assistants [32]. It evaluates the performance of LLMs by using the GPT-4 API to score the LLM-generated conversations. LLMs that behave more like GPT-4 will receive higher scores.

A.2 Hyperparameter Details

In all our experiments, the learning rate of fine-tuning is set to 0.0003; the batch size is 128 and the micro batch size is 16. Due to the large dataset and model sizes selected, federated fine-tuning consumes significant computational resources and time. Therefore, we opted for fewer fine-tuning rounds (even just one round) to ensure that we could observe enough data. Additionally, the MMLU dataset is prone to overfitting on these large datasets, resulting in a decrease in accuracy. Therefore, fewer training rounds ensure the effectiveness of the observed phenomena. Table 2 shows the fine-tuning rounds and local epochs we selected.

Table 2: The communication rounds and local epochs on each experiment setting. The Rounds column represents the number of communication rounds and the Epochs column represents the number of local fine-tuning epochs in each round.

Foundation	Datasets	Rounds	Epochs
TinyLlama	Dolly	3	1
	Alpaca	3	1
	Wizard	3	1
	ShareGPT	1	1
Llama	Dolly	3	3
	Alpaca	3	3
	Wizard	1	1
	ShareGPT	1	1
Llama2	Wizard	1	1
	ShareGPT	1	1

Table 3: The performance of FLORA + AdaLoRA. AdaLoRA can reduce the rank while preserving the fine-tuning effectiveness.

Foundation model	Fine-tuning algorithm	Sum of local ranks	MT-bench score
TinyLlama	FLORA	160	3.13
	FLORA+AdaLoRA	120	3.14
Llama	FLORA	160	4.21
	FLORA+AdaLoRA	131	4.10
Llama2	FLORA	160	4.17
	FLORA+AdaLoRA	140	4.25

Table 4: Compare FLORA with baselines in Llama2.

Strategy	Fine-tuning algorithm	Wizard	ShareGPT
Centralized	LoRA	4.24	3.99
Homo	FedIT	4.03	3.87
	FLORA	4.22	3.96
Heter	Zero-Padding	4.01	3.70
	FLORA	4.17	3.91

A.3 Supplementary Experiment Results

Integrating FLORA with AdaLoRA All the observations about the impact of rank on the model performance, despite being influenced by data heterogeneity, still manage to reveal the importance of selecting an appropriate LoRA rank for a specific task. Thus, some algorithms such as AdaLoRA [30] are designed to adaptively adjust the LoRA rank to optimize the model performance and save computational resources. With our support for heterogeneous LoRA, we can flexibly utilize AdaLoRA with adaptive LoRA ranks. We conducted corresponding experiments to demonstrate that we can use AdaLoRA to further improve the efficiency of federated fine-tuning. We implement AdaLoRA on each client to adjust LoRA modules during local fine-tuning. The results are shown in Table 3. The "Sum of local ranks" column means the sum of all local LoRA rank values *after* fine-tuning. Since our FLORA does not adjust the rank, its value is 160, the same as the initial value. On the other hand, AdaLoRA dynamically adjusts the rank to maximize training effectiveness and minimize rank values to save resources. From Table 3, we can see that AdaLoRA on TinyLlama and Llama reduced the sum of local ranks to 120 and 131 from 160 respectively. We further conclude that FLORA+AdaLoRA can further reduce the trainable parameter count while ensuring comparable or even improved performance compared to simply using LoRA on the clients. Our support for such rank adaptation further demonstrates the effectiveness and applicability of the FLORA approach.

Experiment results of Llama2. Due to the inherently strong performance of Llama2, the improvement in the QA dataset is not significant. Therefore, we fine-tuned Llama2 using the Wizard and ShareGPT datasets. Overall, Llama2 exhibits similar experimental results to TinyLlama and Llama. Table 4 shows the comparison between FLORA and our baselines. In the homogeneous and heterogeneous settings, the MT-bench scores of Wizard and ShareGPT all surpass those in FedIT and Zero-Padding. As for the impact of scaling factors in Figure 7, Llama2 has a similar trend to the Llama-7b model, in which higher scaling factors exhibit better fine-tuning performance.

B Convergence Analysis

In this section, we demonstrate the convergence of FLORA following the standard convergence analysis in [12]. The FedAvg algorithm exhibits convergence to the global optimum at a rate of $O(1/T)$ for non-IID (independent and identically distributed) data under full client participation. This convergence is based on four assumptions mentioned in [12]:

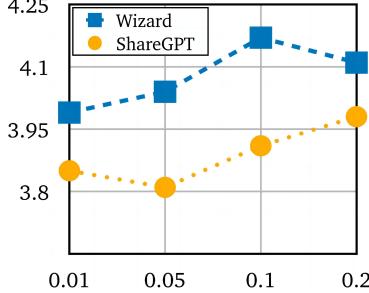


Figure 7: The impact of scaling factor on Llama2 model.

Assumption 1. Each local objective function is L- smooth, that is, for all x and y , $F_k(x) \leq F_k(y) + (x - y)^T \nabla F_k(y) + \frac{L}{2} \|x - y\|_2^2$

Assumption 2. Each local objective function is μ - strongly convex that is, for all x and y , $F_k(x) \geq F_k(y) + (x - y)^T \nabla F_k(y) + \frac{\mu}{2} \|x - y\|_2^2$

Assumption 3. The variance of stochastic gradients in each client is bounded: $\mathbb{E} \|\nabla F_k(W_k^{(t)}, \xi_k^{(t)}) - \Delta W_k^{(t)}\|^2 \leq \sigma_k^2$ for $k = 1, \dots, K$, where $\xi_k^{(t)}$ is the subset of training data randomly sampled from k -th client.

Assumption 4. The expected squared norm of stochastic gradients is uniformly bounded: $\mathbb{E} \|\nabla F_k(W_k^{(t)}, \xi_k^{(t)})\|^2 \leq G^2$ for all $k = 1, \dots, K$ and $t = 1, \dots, T$, where $\xi_k^{(t)}$ is the subset of training data randomly sampled from k -th client.

For the convergence analysis of FLORA, we introduce an additional assumption 5 tailored to the specific dynamics of LoRA fine-tuning and its relation to traditional SGD-based full fine-tuning:

Assumption 5. (Unbiased LoRA Gradient). The updates applied to LoRA modules by each client serve as unbiased estimators of the gradient that would be directly computed on the base model through SGD: $\mathbf{B}_k^{(t+1)} \mathbf{A}_k^{(t+1)} - \mathbf{B}_k^{(t)} \mathbf{A}_k^{(t)} = \eta^{(t)} \nabla F_k(\mathbf{W}_k^{(t)} | \xi_k^{(t)})$. Note that we define the model parameter in t -th round by $\mathbf{W}^{(t)}$.

Theorem 1. Based on Assumptions 1-5, we choose $k = \frac{L}{\mu}$, $\gamma = \max\{8k, E\}$. The local learning rate $\frac{\alpha_k}{r_k} = \frac{2}{\mu(\gamma+t)}$. Then, we can deduce that the expectation of the fine-tuning error in FLORA can be bounded by:

$$\delta^{(T)} \leq \frac{2k}{\gamma+T} \left(\frac{M}{\mu} + 2L \|\mathbf{W}^{(1)} - \mathbf{W}^*\|^2 \right), \quad (13)$$

where $\delta^{(T)}$ is the fine-tuning error in T -th round. $\delta^{(T)}$ and M are defined as follows:

$$\begin{aligned} \delta^{(T)} &= \mathbb{E}[F(\mathbf{w}^{(T-1)} + \mathbf{B}^{(T)} \mathbf{A}^{(T)})] - F^*, \\ M &= \sum_{k=1}^K p_k^2 \sigma_k^2 + 6L\Gamma + 8(E-1)^2 G^2. \end{aligned} \quad (14)$$

where L , μ , σ_k , and G are defined by the assumptions 1-4. Γ is defined by $\Gamma = F * - \sum_{k=1}^K p_k F_k^*$ for quantifying the degree of non-iid. This theorem posits that as the number of rounds T approaches infinity, the expectation of the fine-tuning error $\delta^{(T)}$ converges to zero. In contrast, FedIT deviates from the FedAvg model updating rule as depicted in Equation 2, introducing non-gradient noises through its averaging process. Therefore, it fails to achieve convergence at the rate of $O(1/T)$. While this deviation does not invalidate FedIT's utility in federated fine-tuning, it significantly impairs its convergence rate and overall effectiveness.

C Limitation

Our approach has the limitation that the server sends the stacked LoRA modules to the client, thereby increasing the communication costs. We discussed this limitation both theoretically and

experimentally in Section 3 and Section 4, respectively. We believe that the increase in communication overhead is acceptable under the premise of improving fine-tuning effectiveness and accelerating convergence. In addition, due to constraints on computational resources and time, we only utilized Llama models in the experimental section. We aim to observe experimental phenomena of different types of LLM federated fine-tuning in future research and derive more general principles.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The paper has a clear abstract and an introduction section.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The Limitation section is presented in the Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All proofs are provided.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Experimental settings are provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is on GitHub. We already have a discussion with other researchers about the code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Details are in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The results are reproducible with low errors.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Information on GPUs and resources is in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Conducted.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: It is only a technical paper, which has no social impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: All the models and datasets used are open-sourced.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Models and datasets are cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No human-involved evaluations.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.