Align Before Aggregation: Basis-Consistent Federated LoRA under Heterogeneous Ranks

Pengpeng Qiao[1] **Yang Cao**[1] **Lingling Zhang**[2*] **Guo Cheng**[3]
Junwei Chen[1] **Manjiang Yu**[4] **Wei Yang Bryan Lim**[5] **Masatoshi Yoshikawa**[6]

[1]Institute of Science Tokyo [2]Capital Normal University
[3]Beijing Institute of Technology [4]The University of Queensland
[5]Nanyang Technological University [6]Osaka Seikei University

{peng2qiao,214wadechen}@gmail.com, cao@c.titech.ac.jp, 7089@cnu.edu.cn
guo_cheng@bit.edu.cn, manjiang.yu@uq.edu.cn
bryan.limwy@ntu.edu.sg, yoshikawa-mas@g.osaka-seikei.ac.jp

Abstract

Federated fine-tuning of large language models (LLMs) under heterogeneous client capacities is increasingly important, and low-rank adaptation (LoRA) with heterogeneous ranks makes it practical. Existing server-side aggregators compose local matrix updates ΔW_i from client-specific LoRA factors (A_i, B_i) before aggregating them. Although these matrix updates lie in the same weight space $\mathbb{R}^{m\times n}$ and can be averaged directly, the LoRA factors that generate them may expressed client-specific local basis. In the typical compose–aggregate–factorize–dispatch loop, this basis inconsistency can make the rank-r_i reference prefix dispatched to lower-rank clients unstable after server-side refactorization. In this work, we propose **FedAbA**, a basis-consistent federated LoRA aggregation framework for heterogeneous ranks. FedAbA first extracts each client's local basis, aligns it to the corresponding prefix of the global reference basis, and constructs a basis-consistent reconstruction of each local matrix update before aggregation. It then aggregates these reconstructions with weights that combine client data size and consistency score, followed by exact coefficient-space SVD refactorization and gauge fixing before rank-prefix dispatch. Our theoretical analysis clarifies the role of basis consistency and supports the proposed consistency-aware aggregation mechanism. Extensive experiments show that FedAbA outperforms representative federated LoRA baselines, with the largest gains under the strongest rank heterogeneity.

1 Introduction

Large pretrained models, especially large language models (LLMs) [8, 7, 13], have become a central backbone for a wide range of downstream applications [24, 33]. In many realistic deployments, however, the data needed to adapt these models is decentralized and privacy-sensitive. Federated learning (FL) offers a natural solution by allowing distributed clients to collaboratively fine-tune a shared model without exposing raw local data [29, 28, 30, 43]. Since full-parameter federated fine-tuning is prohibitively expensive in both communication and on-device computation, low-rank adaptation (LoRA) [16] has emerged as a practical parameter-efficient choice for federated fine-tuning of LLMs. In practice, clients also differ substantially in resources [1, 11, 19, 34], which naturally leads them to adopt different ranks r_i, yielding *heterogeneous-rank* federated LoRA.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

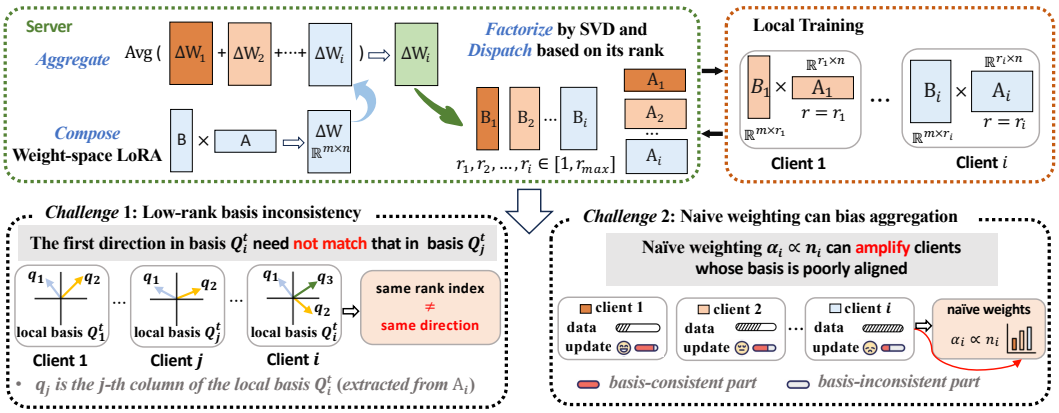

Figure 1: Illustration of federated LoRA under heterogeneous ranks. Compose–aggregate–factorize–dispatch training process (top). Two challenges: basis inconsistency and biased aggregation (bottom).

Why heterogeneous ranks are difficult. A typical server-side aggregation pipeline in federated LoRA follows a compose–aggregate–factorize–dispatch loop, as shown in Figure 1. Client i uploads LoRA factors (A_i, B_i). The server composes the local matrix update $\Delta W_i = sB_iA_i \in \mathbb{R}^{m\times n}$, averages these local matrix updates across clients, factorizes the averaged matrix into new global LoRA factors, and dispatches the rank-r_i reference prefix of those global factors back to client i. The difficulty is not the size of ΔW_i: regardless of r_i, every ΔW_i lies in the same matrix space and can be averaged as a matrix. The difficulty arises from what happens after this matrix average, because the server repeatedly moves between matrix updates and low-rank factors, while the factors uploaded by different clients can be learned in client-specific local basis. This mismatch comes from the non-uniqueness of low-rank factorization. For any invertible matrix $R \in \mathbb{R}^{r_i\times r_i}$,

$$B_iA_i = (B_iR)(R^{-1}A_i), \qquad A_i \in \mathbb{R}^{r_i\times n}, \qquad B_i \in \mathbb{R}^{m\times r_i}.$$

Thus, even two clients that learn similar matrix updates can encode them in different local basis. This ambiguity is less harmful when all clients use the same rank, because each client receives the full global LoRA factors and no rank-prefix truncation is imposed. Under heterogeneous ranks, however, lower-rank clients receive only the rank-r_i reference prefix of the global LoRA factors. After local training, the first uploaded direction of client i need not correspond to the same reference direction as the first uploaded direction of client j. When the server aggregates local matrix updates, refactorizes via SVD, and redistributes rank-r_i reference prefixes, the prefix coordinates received by lower-rank clients may change their semantics across rounds. This makes it important to define aggregation relative to an explicit global reference basis. We refer to this phenomenon as *basis inconsistency*.

Furthermore, basis inconsistency also affects how clients should be weighted. The typical aggregation strategy uses data-size weighting. Under heterogeneous-rank LoRA, however, a client with many samples can still be poorly aligned with the global reference basis. Giving such a client a large weight may amplify components that are poorly compatible with the reference basis and are later affected by reconstruction or refactorization. This suggests that server aggregation should reflect not only how much data a client has, but also how consistent its local basis is with the global reference basis.

Motivation: Align before aggregation. To address these challenges, we propose FedAbA, a *basis-consistent* federated LoRA aggregation framework that aligns client updates to a shared global reference basis before aggregation. Using the orthonormal basis induced by the previous-round global right factor as the reference basis, the server extracts each client's local basis from its uploaded right factor, aligns it to the corresponding reference prefix via orthogonal Procrustes, and reconstructs the local matrix update on that prefix. This reconstruction is essential: merely rotating local LoRA factors leaves the composed update ΔW_i unchanged and therefore cannot alter server aggregation. The alignment step also produces a *consistency score* that measures how well the client's local basis matches the reference prefix. FedAbA then aggregates basis-consistent reconstructions using weights that combine client data size and consistency score. Finally, the averaged basis-consistent global update is refactorized by an exact coefficient-space SVD, avoiding a full $m \times n$ SVD, and a lightweight gauge-fixing step removes arbitrary sign flips and near-degenerate ordering changes before the first r_i components are sent back to client i. Our contributions are summarized as follows:

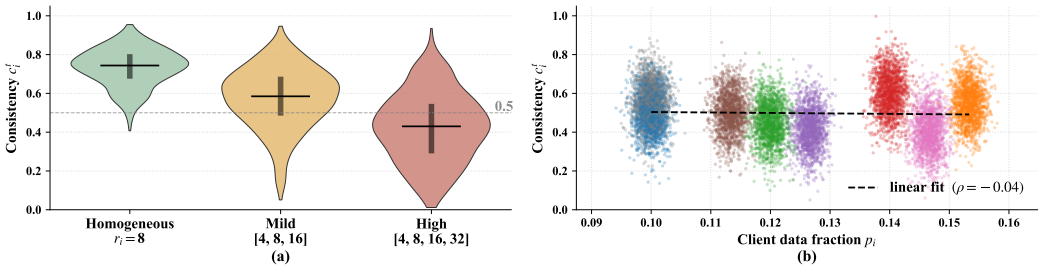

Figure 2: (a) Distribution of c_i^t at t=10 shifts downward as rank heterogeneity grows. (b) Under strong heterogeneity, c_i^t is weakly correlated with the data-size weight p_i.

- We identify that the main challenge in heterogeneous-rank federated LoRA is the lack of explicit basis reconciliation against a global reference basis in the compose–aggregate–factorize–dispatch loop, rather than a dimensional mismatch in the local matrix updates. We also show that factor-space rotations alone cannot resolve this issue.
- We propose **FedAbA**, a server-side aggregation framework that reconstructs each local matrix update under a shared global reference basis before aggregation, uses both data size and consistency score in aggregation, and enables an exact coefficient-space SVD refresh, rather than a full-update SVD, with gauge fixing before rank-prefix dispatch.
- We provide theoretical analysis showing how the consistency score characterizes basis compatibility and supports consistency-aware aggregation under heterogeneous ranks. Experiments show that FedAbA consistently outperforms representative federated LoRA baselines.

2 Method

2.1 Problem Setup

We consider federated fine-tuning of LLMs under heterogeneous client LoRA ranks. A server coordinates N clients. Client i owns a private dataset $\mathcal{D}_i$ of size n_i and uses LoRA rank $r_i \in [1, r_{\max}]$. At round t, an active subset $\mathcal{S}^t \subseteq \{1, \dots, N\}$ of clients participates. We describe the method for a single LoRA-inserted layer and omit the layer index.

Let $W_0 \in \mathbb{R}^{m\times n}$ be the frozen pretrained weight. Client i maintains local LoRA factors $A_i^t \in \mathbb{R}^{r_i\times n}$ and $B_i^t \in \mathbb{R}^{m\times r_i}$, producing the local matrix update $\Delta W_i^t = s\, B_i^t A_i^t \in \mathbb{R}^{m\times n}$, where s is the LoRA scaling factor. Unless otherwise specified, both A_i and B_i are trainable during local fine-tuning, so the row space of A_i^t and the corresponding local basis may change across rounds. The server maintains global LoRA factors $A_g^t \in \mathbb{R}^{r_{\max}\times n}$ and $B_g^t \in \mathbb{R}^{m\times r_{\max}}$, and distributes the rank-r_i prefix $A_g^t[1\!:\!r_i, :]$ and $B_g^t[:, 1\!:\!r_i]$ to client i.

Under heterogeneous ranks, lower-rank clients receive only the rank-r_i prefix of the global factors, so the per-coordinate semantics of those prefixes are determined by how the global basis is ordered across rounds. Aggregating client contributions on a shared, explicit global reference basis is therefore central to controlling cross-round prefix stability. We next quantify how severe this basis inconsistency is in practice (Section 2.2) before describing FedAbA (Section 2.3). Table 6 summarizes the notation.

2.2 Motivation

The basis-inconsistency phenomenon described in Section 1 is naturally quantified by the alignment between each client's local basis and the global reference basis it should be measured against. For client i at round t, let $Q_i^t \in \mathbb{R}^{n\times r_i}$ be the local basis obtained from a thin QR decomposition of the uploaded right factor $(A_i^t)^\top$, and let $Q_{\mathrm{ref},i}^{t-1} \in \mathbb{R}^{n\times r_i}$ denote the reference prefix taken from the global reference basis $Q_g^{t-1} = (A_g^{t-1})^\top$. We define the *consistency score*

$$c_i^t = \frac{1}{r_i}\operatorname{tr}(\Sigma_i^t) \;\in\; [0,1], \qquad (Q_i^t)^\top Q_{\mathrm{ref},i}^{t-1} = U_i^t \Sigma_i^t (V_i^t)^\top, \tag{1}$$

which is the average principal-angle cosine between the local basis and the reference prefix ($c_i^t = 1$ means perfect alignment). This is the same quantity that FedAbA later uses for both basis alignment

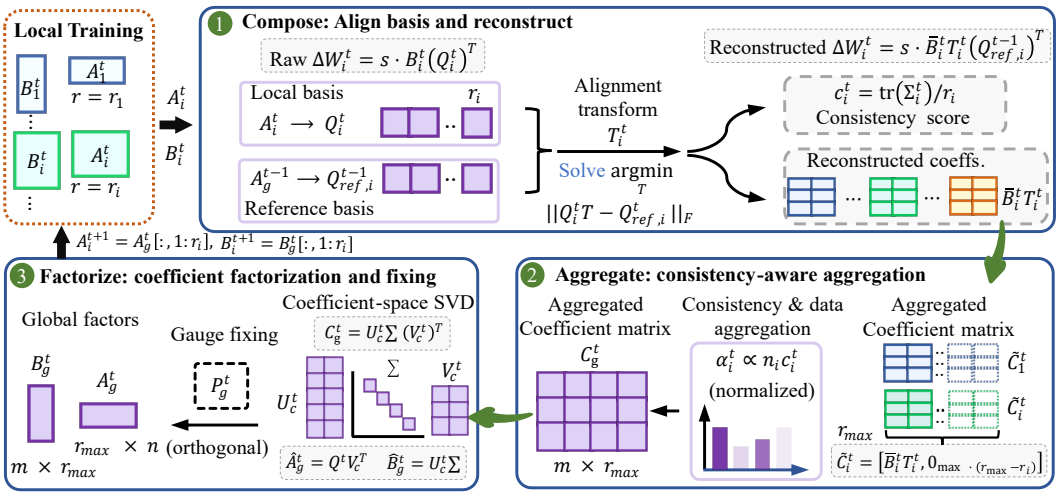

Figure 3: Overview of FedAbA.

and aggregation weighting (Section 2.3); we therefore measure it directly on a representative federated LoRA training backbone. Note that $c_i^t < 1$ generically because (i) A_i is locally trainable, so the local basis a client uploads has drifted from the reference prefix it received at the start of the round, and (ii) the server's SVD-based refactorization can refresh the global basis between rounds. To probe how this consistency behaves in practice, we fine-tune LLAMA-2-7B with N=8 clients on DOLLY-15K using FlexLoRA [1] as the training backbone (see Appendix B for details).

Observation 1 (rank heterogeneity sharply degrades basis consistency). Figure 2(a) shows the distribution of c_i^t over all clients and adapted layers at round t=10 under the three regimes. Even in the homogeneous case the median sits well below 1, and as rank heterogeneity grows the entire distribution shifts downward, with a substantial mass of c_i^t values falling below 0.5 under high heterogeneity. By Theorem 3.3, a smaller c_i^t directly translates into a larger reconstruction residual $\sqrt{2r_i(1-c_i^t)}$, so naive aggregation under heterogeneous ranks operates on increasingly inconsistent local bases. This motivates explicitly aligning each client's local basis to the reference prefix *before* aggregation, rather than relying on post-hoc server-side SVD to reconcile mismatched contributions.

Observation 2 (c_i^t varies sharply across clients and is weakly correlated with data size). Figure 2(b) plots c_i^t against the data-size weight p_i under the high-heterogeneity regime. The two quantities are weakly correlated ($|\rho| < 0.2$), and clients with the largest p_i are not necessarily those with the largest c_i^t. Pure data-size weighting can therefore amplify contributions from clients whose local basis is poorly aligned with the reference prefix, exactly the case in which the reconstruction residual is largest. This motivates a *consistency-aware* aggregation weight that combines p_i with c_i^t.

Together, these observations point to two design choices that FedAbA makes in Section 2.3: (i) before aggregation, replace each client's local basis with the corresponding reference prefix via an orthogonal Procrustes alignment, so that all client contributions are expressed under a shared global reference basis; and (ii) weight client contributions by $\alpha_i^t \propto p_i\, c_i^t$, so that data-rich but basis-misaligned clients no longer dominate the aggregate. Fig. 3 illustrates the overall workflow of FedAbA.

2.3 Basis-Consistent Federated LoRA Aggregation

At the start of round t, the server holds the previous-round global LoRA factors (A_g^{t-1}, B_g^{t-1}). We define the *global reference basis* as $Q_g^{t-1} := (A_g^{t-1})^\top \in \mathbb{R}^{n \times r_{\max}}$, which has orthonormal columns by construction. For client i, the corresponding *reference prefix* is $Q_{\mathrm{ref},i}^{t-1} := Q_g^{t-1}[:, 1 : r_i] \in \mathbb{R}^{n \times r_i}$. All clients in the same round are therefore aligned to prefixes of the same global reference basis.

FedAbA proceeds in three steps: (i) extract and align each client's local basis; (ii) construct a basis-consistent local matrix update under the reference prefix; and (iii) aggregate these basis-consistent updates with consistency-aware weights and refactorize the aggregate for rank-prefix dispatch.

Step 1: Local basis extraction and alignment

The server does not align the raw right factor A_i^t directly, because A_i^t entangles basis directions with coefficient magnitudes. Instead, it extracts the local basis by a thin QR decomposition of $(A_i^t)^\top$:

$$(A_i^t)^\top = Q_i^t (C_i^t)^\top, \qquad Q_i^t \in \mathbb{R}^{n\times r_i}, \quad (Q_i^t)^\top Q_i^t = I_{r_i}, \quad C_i^t \in \mathbb{R}^{r_i\times r_i}. \tag{2}$$

Substituting $A_i^t = C_i^t (Q_i^t)^\top$ into $\Delta W_i^t = s B_i^t A_i^t$ gives

$$\Delta W_i^t = s\, \bar{B}_i^t (Q_i^t)^\top, \qquad \bar{B}_i^t := B_i^t C_i^t \in \mathbb{R}^{m\times r_i}. \tag{3}$$

Here Q_i^t is the local basis, while $\bar{B}_i^t$ contains the corresponding coefficients. This representation separates the basis to be aligned from the coefficients to be preserved. Given Q_i^t and $Q_{\mathrm{ref},i}^{t-1}$, the server computes the orthogonal Procrustes alignment

$$T_i^t = \arg\min_{T\in\mathcal{O}(r_i)} \|Q_i^t T - Q_{\mathrm{ref},i}^{t-1}\|_F^2, \tag{4}$$

where $\mathcal{O}(r_i)$ is the orthogonal group. Writing $(Q_i^t)^\top Q_{\mathrm{ref},i}^{t-1} = U_i^t \Sigma_i^t (V_i^t)^\top$, the solution is $T_i^t = U_i^t (V_i^t)^\top$. This alignment removes the best within-rank rotation between the local basis and the reference prefix while preserving coefficient norms.

The singular values $\sigma_{i,j}^t$ of $(Q_i^t)^\top Q_{\mathrm{ref},i}^{t-1}$ are the cosines of the principal angles between the two bases, and the alignment step recovers the same consistency score introduced in Section 2.2: $c_i^t = \frac{1}{r_i}\mathrm{tr}(\Sigma_i^t) = \frac{1}{r_i}\sum_{j=1}^{r_i} \sigma_{i,j}^t \in [0,1]$. A larger c_i^t indicates better consistency between the local basis and the reference prefix; FedAbA reuses c_i^t for aggregation weighting in **Step 3**.

Step 2: Basis-consistent reconstruction

A central design decision of FedAbA is that the server does *not* merely rotate local factors. For any orthogonal $T \in \mathcal{O}(r_i)$, $\bar{B}_i^t T (Q_i^t T)^\top = \bar{B}_i^t (Q_i^t)^\top$. Thus, factor rotation leaves the composed matrix update unchanged. Aggregating rotated factors would therefore produce the same matrix aggregate as aggregating the raw ΔW_i^t. Proposition 3.1 formalizes this invariance.

Instead, FedAbA constructs a basis-consistent local matrix update by replacing the local basis with the reference prefix:

$$\Delta\widetilde{W}_i^t = s\, \bar{B}_i^t T_i^t (Q_{\mathrm{ref},i}^{t-1})^\top. \tag{5}$$

This construction uses all r_i aligned coefficient columns of client i, but expresses them on the reference prefix. It does not select a subset of local basis directions.

Let $E_i^t := Q_i^t T_i^t - Q_{\mathrm{ref},i}^{t-1}$. Then the raw local matrix update decomposes as

$$\Delta W_i^t = \Delta\widetilde{W}_i^t + s\, \bar{B}_i^t T_i^t (E_i^t)^\top. \tag{6}$$

The difference $\Delta W_i^t - \Delta\widetilde{W}_i^t$ is the reconstruction residual caused by replacing the local basis with the reference prefix. This residual is not a collection of unused basis directions. It quantifies the difference between the raw local update and its reference-basis reconstruction. In FedAbA, only $\Delta\widetilde{W}_i^t$ is aggregated. Theorem 3.3 shows that $\|\Delta W_i^t - \Delta\widetilde{W}_i^t\|_F \le s\|\bar{B}_i^t\|_2 \sqrt{2r_i(1-c_i^t)}$.

Step 3: Consistency-aware aggregation and refactorization

The server combines the data-size weight $p_i^t = n_i / \sum_{j\in\mathcal{S}^t} n_j$ with the consistency score c_i^t:

$$\alpha_i^t = \frac{p_i^t\, c_i^t}{\sum_{j\in\mathcal{S}^t} p_j^t\, c_j^t} = \frac{n_i\, c_i^t}{\sum_{j\in\mathcal{S}^t} n_j\, c_j^t}. \tag{7}$$

We assume $Z^t = \sum_{j\in\mathcal{S}^t} p_j^t\, c_j^t$ is positive. The case $Z^t = 0$ occurs only when all active clients have $c_i^t = 0$, so the consistency scores provide no weighting signal; we then fall back to $\alpha_i^t = p_i^t$. The aggregated basis-consistent update is $\Delta\bar{W}_g^t = \sum_{i\in\mathcal{S}^t} \alpha_i^t\, \Delta\widetilde{W}_i^t$.

Equivalently, define the zero-padded coefficient matrix $\widetilde{C}_i^t = \left[\bar{B}_i^t T_i^t,\ 0_{m\times(r_{\max}-r_i)}\right] \in \mathbb{R}^{m\times r_{\max}}$. Then we can rewrite

$$\Delta\bar{W}_g^t = s\, C_g^t (Q_g^{t-1})^\top, \qquad C_g^t = \sum_{i\in\mathcal{S}^t} \alpha_i^t \widetilde{C}_i^t. \tag{8}$$

This coefficient-space view makes the heterogeneous-rank structure explicit: client i contributes coefficient columns associated with the first r_i basis directions of the global reference basis. To redistribute the aggregate in rank-$r_{\max}$ LoRA form, the server performs an exact coefficient-space SVD refresh. Since $C_g^t \in \mathbb{R}^{m\times r_{\max}}$, it suffices to compute

$$C_g^t = U_c^t \Sigma_c^t (V_c^t)^\top, \tag{9}$$

where $V_c^t \in \mathbb{R}^{r_{\max}\times r_{\max}}$ is orthogonal and zero singular values are padded if necessary. Then

$$\Delta\bar{W}_g^t = s\, U_c^t \Sigma_c^t (Q_g^{t-1} V_c^t)^\top. \tag{10}$$

The provisional global factors are $\widehat{A}_g^t = (Q_g^{t-1}V_c^t)^\top$, $\widehat{B}_g^t = U_c^t\Sigma_c^t$. They exactly represent the aggregated basis-consistent update: $s\widehat{B}_g^t\widehat{A}_g^t = \Delta\bar{W}_g^t$. Thus, the server decomposes only the coefficient matrix $C_g^t \in \mathbb{R}^{m\times r_{\max}}$ rather than materializing and decomposing the full $m \times n$ update.

Gauge fixing. The SVD in Eq. (9) can introduce arbitrary sign flips and swaps among nearly equal singular values. These changes do not affect the composed matrix update, but they can relabel the factor indices sent to clients. We thus use a lightweight gauge-fixing step before rank-prefix dispatch.

Let $\tau_{\mathrm{svd}} > 0$ be a relative-gap threshold. After sorting singular values in descending order, we form contiguous near-degenerate blocks by splitting at adjacent relative gaps larger than τ_{svd}. Within each block, provisional directions in $\widehat{Q}_g^t = (\widehat{A}_g^t)^\top$ are matched to the corresponding columns of Q_g^{t-1} by maximum absolute cosine similarity, with signs chosen to make the matched cosines nonnegative. This yields a block-wise signed permutation matrix P_g^t. The final global factors are

$$A_g^t = P_g^t \widehat{A}_g^t, \qquad B_g^t = \widehat{B}_g^t (P_g^t)^\top. \tag{11}$$

Since P_g^t is orthogonal, $sB_g^tA_g^t = s\widehat{B}_g^t\widehat{A}_g^t = \overline{\Delta W}^t{}_g$. Thus, gauge fixing changes only the factor representation, not the composed global matrix update.

Remark 2.1 (Span of the global basis). The SVD refresh changes the global basis but not its span in the default FedAbA algorithm. After gauge fixing,

$$Q_g^t = (A_g^t)^\top = Q_g^{t-1} V_c^t (P_g^t)^\top = Q_g^{t-1} O_g^t, \qquad O_g^t \in \mathcal{O}(r_{\max}).$$

Therefore, $\mathrm{span}(Q_g^t) = \mathrm{span}(Q_g^{t-1})$. By induction, the span of the global basis remains the warm-start span throughout training. FedAbA should therefore be understood as a basis-consistent aggregation method within a warm-started global basis.

2.4 Workflow and Complexity

Algorithm 1 summarizes the training workflow of FedAbA. FedAbA requires a non-degenerate initial global reference basis. Since a zero LoRA update does not define meaningful basis directions, we initialize the first global LoRA factors by a warm-start phase: the server performs several standard federated LoRA aggregation rounds without basis-consistent reconstruction, computes a rank-$r_{\max}$ truncated SVD of the resulting aggregated matrix update, and uses the resulting factors as (A_g^0, B_g^0).

The client-side training protocol and communication are unchanged relative to standard federated LoRA fine-tuning. All additional operations are performed on the server. For each active client, thin QR of $(A_i^t)^\top$ costs $O(nr_i^2)$. Forming $(Q_i^t)^\top Q_{\mathrm{ref},i}^{t-1}$ and solving the Procrustes problem costs $O(nr_i^2 + r_i^3)$. Computing $\bar{B}_i^t = B_i^t C_i^t$ and $\bar{B}_i^t T_i^t$ costs $O(mr_i^2)$. Coefficient aggregation costs $O(|\mathcal{S}^t| mr_{\max})$. The final SVD is performed on $C_g^t \in \mathbb{R}^{m\times r_{\max}}$, costing $O(mr_{\max}^2 + r_{\max}^3)$, rather than on a full $m \times n$ matrix. Communication complexity remains $O(r_i(m+n))$ per client.

3 Theoretical Analysis

This section analyzes the aggregation of FedAbA. We first show why factor rotation alone cannot resolve basis inconsistency (Section 3.1), then quantify the reconstruction through the consistency

score and justify consistency-aware weighting (Section 3.2), and finally give an idealized convergence interpretation (Section 3.3). Throughout, we use the basis–coefficient form $\Delta W_i^t = s\bar{B}_i^t(Q_i^t)^\top$ and the basis-consistent reconstruction $\Delta\widetilde{W}_i^t = s\bar{B}_i^t T_i^t (Q_{\mathrm{ref},i}^{t-1})^\top$. Full proofs are in Appendix E.

3.1 Basis Alignment Requires Reconstruction, Not Rotation

A natural first attempt is to rotate each client's LoRA factors by the Procrustes solution before aggregation. This operation is invisible after composing the matrix update.

Proposition 3.1 (Factor rotation invariance). *For $T \in \mathcal{O}(r_i)$, $s\bar{B}_i^t T (Q_i^t T)^\top = s\bar{B}_i^t (Q_i^t)^\top = \Delta W_i^t$.*

FedAbA therefore replaces the aligned local basis with the reference prefix, as in Eq. (5) (proof in Appendix E.1). The reconstructed updates then admit a single shared-basis representation.

Proposition 3.2 (Shared-basis representation). *Let $\widetilde{C}_i^t = [\bar{B}_i^t T_i^t, 0_{m\times(r_{\max}-r_i)}]$. Then $\Delta\widetilde{W}_i^t = s\widetilde{C}_i^t (Q_g^{t-1})^\top$ and $\Delta\bar{W}_g^t = s\big(\sum_{i\in\mathcal{S}^t}\alpha_i^t \widetilde{C}_i^t\big)(Q_g^{t-1})^\top$, so $\mathrm{rank}(\Delta\bar{W}_g^t) \le r_{\max}$.*

Consequently (proof in Appendix E.2), the coefficient-space SVD refresh in Eq. (10) *exactly* refactorizes $\Delta\bar{W}_g^t$, and the gauge fixing in Eq. (11) only re-labels factor coordinates. Both steps are bias-free; their formal statements (exact refresh, gauge invariance, and span invariance within the warm-started global basis) are deferred to Appendix E.3.

3.2 Reconstruction Residual and Consistency-Aware Weighting

Let $(Q_i^t)^\top Q_{\mathrm{ref},i}^{t-1} = U_i^t \Sigma_i^t (V_i^t)^\top$ and recall $c_i^t = \frac{1}{r_i}\mathrm{tr}(\Sigma_i^t)$. The next theorem connects this score to both the basis alignment quality and the reconstruction residual.

Theorem 3.3 (Reconstruction residual is controlled by basis consistency). $\min_{T\in\mathcal{O}(r_i)} \|Q_i^t T - Q_{\mathrm{ref},i}^{t-1}\|_F^2 = 2r_i(1-c_i^t)$, *and* $\|\Delta W_i^t - \Delta\widetilde{W}_i^t\|_F \le s\|\bar{B}_i^t\|_2\sqrt{2r_i(1-c_i^t)}$.

Thus c_i^t is exactly the average principal-angle cosine between local basis and reference prefix (proof in Appendix E.4). The residual vanishes when $c_i^t = 1$ and grows monotonically as alignment degrades.

Assumption 3.4 (Uniform coefficient-radius bound). There exists $\kappa_t > 0$ such that $s\|\bar{B}_i^t\|_2\sqrt{r_i} \le \kappa_t$ for every $i \in \mathcal{S}^t$.

Theorem 3.5 (Consistency-aware weighting tightens a reconstruction bound). *Let $b_i^t := \kappa_t\sqrt{2(1-c_i^t)}$. Under Assumption 3.4, $\|\Delta W_i^t - \Delta\widetilde{W}_i^t\|_F \le b_i^t$. If $\sum_{j\in\mathcal{S}^t} p_j^t\, c_j^t > 0$, the consistency-aware weights $\alpha_i^t = \frac{p_i^t c_i^t}{\sum_j p_j^t c_j^t}$ satisfy $\sum_i \alpha_i^t b_i^t \le \sum_i p_i^t b_i^t$.*

Remark 3.6 (Variational view). Theorem 3.5 (proof in Appendix E.5) gives a bounded-error justification. Appendix E.6 shows that α_i^t is also the unique maximizer of a KL-regularized log-consistency objective relative to p, providing a complementary variational characterization.

3.3 Idealized Convergence Interpretation

We finally show how the reconstruction step enters a standard smooth non-convex stochastic descent analysis. Define $G^t := \sum_{i\in\mathcal{S}^t}\alpha_i^t \Delta W_i^t$ and $\widetilde{G}^t := \Delta\bar{W}_g^t$. By Proposition 3.2, $\mathrm{rank}(\widetilde{G}^t) \le r_{\max}$, so the SVD refresh and gauge fixing preserve $\widetilde{G}^t$ exactly and contribute no additional bias beyond the reconstruction step itself.

Assumption 3.7 (Idealized smooth stochastic descent model). F is L-smooth, $\mathbb{E}[G^t \mid W^t] = \nabla F(W^t)$, and $\mathbb{E}[\|G^t - \nabla F(W^t)\|_F^2 \mid W^t] \le \sigma^2$.

Theorem 3.8 (Convergence with reconstruction bias). *Let $e^t := \widetilde{G}^t - G^t$ and consider $W^{t+1} = W^t - \eta\widetilde{G}^t$ with $\eta \le 1/(4L)$. Under Assumption 3.7,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(W^t)\|_F^2 \le \frac{4(F(W^0)-F^\star)}{\eta T} + 4L\eta\sigma^2 + \frac{2+4L\eta}{T}\sum_{t=0}^{T-1}\mathbb{E}\|e^t\|_F^2, \tag{12}$$

and $\|e^t\|_F \le \sum_{i\in\mathcal{S}^t}\alpha_i^t\, s\|\bar{B}_i^t\|_2\sqrt{2r_i(1-c_i^t)}$.

Table 1: Performance comparison of six baselines across various models and tasks.

Method	NLU, RoBERTa-large				NLG-style, LLaMA-2-7B				NI, LLaMA-1B	
	MRPC F1	CoLA MCC	QNLI ACC	RTE ACC	Wizard (MMLU)	Dolly-15K (MMLU)	Alpaca (MMLU)	Alpaca (MT-Bench)	Natural Instructions v2 Rouge-L	Natural Instructions v2 Rouge-1
HetLORA	86.6	59.8	93.8	67.5	34.5	43.8	32.3	3.57	57.0	65.3
FLoRA	87.2	61.8	93.5	71.8	35.4	44.9	34.6	3.80	57.1	65.3
LoRA-A^2	88.9	62.5	93.8	79.5	35.6	45.4	35.6	4.03	57.4	65.4
FlexLoRA	88.7	62.4	93.8	78.3	36.2	45.9	36.2	4.13	57.6	65.5
ECLoRA	89.7	62.9	93.9	82.3	36.3	46.1	36.5	4.15	57.7	65.5
Fed-PLoRA	89.5	63.1	94.0	81.8	36.6	47.2	37.3	4.23	58.1	65.8
FedAbA	**90.3**	**63.6**	**94.2**	**83.5**	**37.1**	**47.8**	**37.8**	**4.32**	**58.4**	**66.1**

The bound (proof in Appendix E.7) isolates FedAbA's reconstruction effect via a measurable bias term controlled by the consistency scores.

4 Experiment

4.1 Setup

Datasets and metrics. We evaluate FedAbA on three benchmarks. For *natural language understanding* (NLU), we use four GLUE tasks [31]: MRPC, CoLA, QNLI, and RTE. For *NLG-style instruction tuning*, we train on Dolly-15K [9], Alpaca [27], and WizardLM [36], and evaluate with MMLU or MT-Bench as specified in Table 1. For *natural instructions* (NI), we use Natural Instructions v2 [32] and report ROUGE-L/ROUGE-1. For NLU and NLG-style experiments, we construct 10-client federated settings with Dirichlet(0.5) partitioning. For NI, each client corresponds to one task; we sample 100 training clients and use a participation rate of 0.1, i.e., 10 clients per round.

Baselines. We compare FedAbA against six baselines for heterogeneous-rank federated LoRA: FlexLoRA [1], HetLoRA [6], ECLoRA [26], LoRA-A^2 [19], FLoRA [34] and Fed-PLoRA [42].

Implementation Details. For NLU/NLG-style/NI experiments, we use RoBERTa-large/LLaMA-2-7B/LLaMA-1B-dj-refine-150B as the frozen backbone and train for 100/50/100 communication rounds, respectively. LoRA is inserted into attention projection layers. For NLU and NLG-style, client ranks are assigned as [4, 4, 8, 8, 16, 16, 32, 32, 64, 64]. For NI, client ranks are balanced over 4, 8, 16, and 32. All methods use AdamW with learning rate 2×10^{-4}. The batch size is 8 for GLUE and NLG-style experiments, and 4 or 8 for Natural Instructions. The maximum sequence length is 128 for GLUE and 512 for the LLaMA-based experiments. We uses 5 warm-start rounds.

4.2 Performance

Exp-1: Main Results. Table 1 compares FedAbA with heterogeneous-rank LoRA baselines on NLU, NLG-style, and NI benchmarks. FedAbA obtains the best reported results across the three task families. The gains are most visible in settings where rank-prefix dispatch is more sensitive to basis drift, while margins are smaller on tasks where most methods already perform strongly. HetLoRA and FLoRA handle rank mismatch through padding or stacking, and FlexLoRA and ECLoRA aggregate reconstructed updates without explicit basis reconciliation. FedAbA improves this pipeline by making aggregation basis-consistent before refactorization. We further compare FedAbA and Fed-PLoRA using three independent matched runs on six selected metrics. Appendix F.1 reports the mean and sample standard deviation across runs; FedAbA achieves a higher mean on all six metrics.

Exp-2: Results on different-rank clients. Table 2 evaluates the dispatched rank-r_i prefixes on each client's local test split. FedAbA improves all rank groups, with the largest gains at $r = 4$ and $r = 8$. This is consistent with our motivation: low-rank clients receive only short global prefixes and are most sensitive to cross-round basis drift. The smaller gain at higher ranks is also expected because high-rank clients receive a larger portion of the global adapter. These results show that FedAbA improves not only global average performance but also the stability and utility of low-rank prefixes. We also compare FedAbA and Fed-PLoRA on a fixed subset containing the 20% of Natural Instructions clients with the lowest average consistency in the default FedAbA run. FedAbA achieves higher average ROUGE-L and ROUGE-1 on this subset (Appendix F.2).

Table 2: Performance comparison across different LoRA ranks on Dolly and Natural Instructions v2.

Method	NLG-style, LLaMA-2-7B, Dolly-15K						NI, LLaMA-1B, Natural Instructions v2				
	Overall (Rouge-L)	$r=4$	$r=8$	$r=16$	$r=32$	$r=64$	**Overall**	$r=4$	$r=8$	$r=16$	$r=32$
HetLORA	29.2	27.6	28.5	29.4	30.2	31.0	57.0	54.8	56.5	57.8	59.1
FLoRA	28.7	26.8	27.9	28.8	30.1	30.9	57.1	54.2	56.1	58.0	59.5
LoRA-A^2	30.0	28.5	29.4	30.1	30.8	31.4	57.4	55.0	56.7	58.2	59.8
FlexLoRA	30.2	28.0	29.0	30.4	31.3	32.0	57.6	55.1	56.9	58.4	60.0
ECLoRA	30.2	28.2	29.1	30.4	31.4	31.9	57.7	55.5	57.2	58.6	60.2
Fed-PLoRA	31.3	29.7	30.8	31.5	32.1	32.6	58.1	56.6	57.9	58.8	60.5
FedAbA	**32.4**	**31.2**	**32.0**	**32.5**	**33.0**	**33.4**	**58.4**	**57.4**	**58.6**	**59.0**	**60.6**

Table 3: Performance comparison across different rank heterogeneity regimes on NLU, NLG, and NI.

Task Family	Rank Regime	Rank Setting	HetLoRA	FLoRA	LoRA-A^2	FlexLoRA	ECLoRA	Fed-PLoRA	FedAbA
NLU Avg.	Homo.	$[8]\times 10$	76.33	77.98	80.00	79.80	80.80	80.75	**80.97**
	Mild	$\{4,8,16\}$	76.65	78.43	80.73	80.17	81.35	81.28	**81.72**
	High	$\{4,8,16,32,64\}$	76.93	78.58	81.18	80.80	82.20	82.10	**82.90**
NLG Avg. (MMLU)	Homo.	$[8]\times 10$	36.37	37.70	38.07	38.77	38.86	39.14	**39.30**
	Mild	$\{4,8,16\}$	36.60	38.17	38.57	39.00	39.16	39.51	**39.83**
	High	$\{4,8,16,32,64\}$	36.87	38.30	38.87	39.43	39.63	40.37	**40.90**
NI Avg. (ROUGE-L/1)	Homo.	$r=8$	61.05	61.10	61.00	61.40	61.40	61.55	**61.60**
	Mild	$\{4,8,16\}$	61.15	61.25	61.20	61.50	61.45	61.70	**61.90**
	High	$\{4,8,16,32\}$	61.15	61.20	61.40	61.55	61.60	61.95	**62.25**

Table 4: System efficiency comparison across baselines. Lower is better for all metrics

Method	Comm. / Round (N=10)	Effective Rank	Aggregation Time	Server Memory
HetLoRA	520 MB	64	0.15s	67 MB
FLoRA	2.86 GB	248	0.12s	260 MB
LoRA-A^2	931 MB	64	0.35s	67 MB
FlexLoRA	520 MB	64	130s	2.15 GB
ECLoRA	520 MB	64	4.0s	4.3 GB
Fed-PLoRA	931 MB	64	0.35s	67 MB
FedAbA	**520 MB**	**64**	2.0s	190 MB

Table 5: Ablation study of FedAbA components across NLU, NLG, and NI tasks.

Variant	NLU Avg.	NLG MMLU	Alpaca MT-Bench	NI R-L	NI R-1
w/o basis consistency	82.02	39.95	4.15	57.8	65.6
rotation-only	82.07	40.00	4.16	57.9	65.6
w/o Procrustes	82.32	40.20	4.20	58.0	65.8
w/o c_i^t weighting	82.42	40.40	4.25	58.1	65.9
w/o data size	82.62	40.55	4.27	58.2	66.0
w/o gauge fixing	82.52	40.50	4.26	58.2	65.9
FedAbA	**82.90**	**40.90**	**4.32**	**58.4**	**66.1**

Exp-3: Effect of Rank Heterogeneity. Table 3 compares all methods under three rank regimes. In the homogeneous regime, rank-prefix truncation is less exposed because clients use the same LoRA rank. As rank heterogeneity grows, the methods separate more clearly. HetLoRA and FLoRA remain limited by padding or stacking, while FlexLoRA and ECLoRA remain competitive but lack explicit reference-basis reconciliation before aggregation. FedAbA achieves the best performance in all regimes, with its largest margin under high heterogeneity. This trend is consistent with Theorems 3.3 and 3.5: basis-consistent reconstruction and consistency-aware weighting become more useful when local bases are less consistent with the reference prefix.

Exp-4: Basis diagnostics. Fig. 4(a) evaluates basis stability under high rank heterogeneity. For each evaluated client, layer, and round, we form $M_i^t = (Q_i^t)^\top Q_{\mathrm{ref},i}^{t-1}$; its singular values are principal-angle cosines between the uploaded local basis and the reference prefix. We report average consistency c_i^t, selected prefix-depth scores $S_k^t = \mathbb{E}[\sigma_k(M_i^t) \mid r_i \geq k]$, and the sign/order flip rate across consecutive global bases. FedAbA yields higher alignment scores and a lower flip rate than baselines.

Exp-5: Sensitivity Analysis. Figure 4(b)–(d) studies sensitivity under high rank heterogeneity. Performance improves with a short warm-start and saturates around T_w=5. For consistency weighting, we use the generalized rule $\alpha_i^t \propto p_i^t (c_i^t)^{\gamma_c}$, where $\gamma_c = 1$ is the default setting; moderate consistency weighting performs best. FedAbA is also stable across the tested gauge-fixing thresholds τ_{svd}.

Exp-6: System Efficiency. Table 4 compares per-round system cost. FedAbA keeps the same rank-prefix communication as HetLoRA, FlexLoRA, and ECLoRA, and avoids the stacked-rank expansion of FLoRA and auxiliary communication of LoRA-A^2/Fed-PLoRA. FedAbA is more expensive than purely factor-space baselines, but substantially cheaper than dense reconstruction-based methods (FlexLoRA and ECLoRA). Its efficiency comes from aggregating coefficient matrices and performing SVD on $C_g^t \in \mathbb{R}^{m \times r_{\max}}$, rather than materializing and decomposing a full $m \times n$ update.

Exp-7: Ablation Study Table 5 ablates each component on the high-heterogeneity setting, and the order of the drops mirrors the design hierarchy of the method. Removing basis-consistent

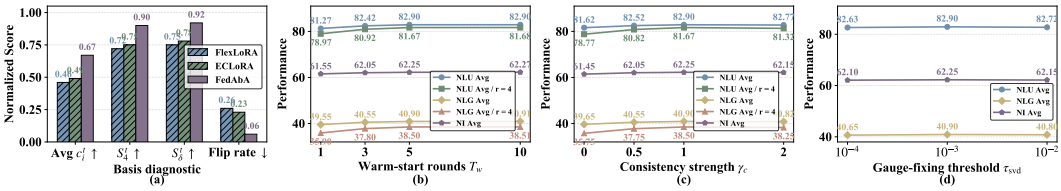

Figure 4: Sensitivity analysis under high rank heterogeneity. (a) Basis diagnostics comparing FlexLoRA, ECLoRA. (b)–(d) Sensitivity of FedAbA to warm-start rounds T_w, consistency weighting strength γ_c, and gauge-fixing threshold τ_{svd}, respectively.

reconstruction entirely (*w/o basis consistency*) causes the largest drop, identifying reconstruction as the central component. Replacing reconstruction with rotation only (*rotation-only*) recovers almost none of the gain, which empirically confirms Proposition 3.1 and rules out factor-rotation strategies as a substitute. Replacing Procrustes with generic orthonormalization loses the optimal alignment guaranteed by Theorem 3.3. Removing c_i^t or the data-size term from the weights shows that the bound-tightening property of Theorem 3.5 requires both factors. Disabling gauge fixing reintroduces the sign-flip pathology visible in Figure 4(a). Each component thus addresses a distinct, theoretically identified source of basis inconsistency.

Exp-8: Additional Backbones. We additionally evaluate FedAbA on Qwen3-4B and LLaMA-2-13B using the Dolly-15K protocol. FedAbA improves MMLU scores over Fed-PLoRA by 0.7 points on both backbones. Appendix F.3 provides the full comparison and experimental setting.

5 Conclusion

We proposed FedAbA, a basis-consistent aggregation method for heterogeneous-rank federated LoRA. By aligning client updates before aggregation and stabilizing rank-prefix dispatch, FedAbA achieves superior performance compared to SOTA heterogeneous Federated LoRA methods.

Acknowledgments and Disclosure of Funding

This work is supported by JSPS KAKENHI 26K21227, JST PRESTO JPMJPR23P5, JST CREST JPMJCR21M2, JST NEXUS JPMJNX25C4, NSFC under Grant 62302043, the Interdisciplinary Project of Capital Normal University under Grant 2026JCZD02, and A*STAR under its Japan-Singapore Joint Call: JST-A*STAR 2024 under Grant R2416IR139.

References

[1] Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. *Advances in Neural Information Processing Systems*, 37:14457–14483, 2024.

[2] Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. Fedalt: Federated fine-tuning through adaptive local training with rest-of-world lora. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 19728–19736, 2026.

[3] Chang Che, Ziqi Wang, Pengwan Yang, Cheems Wang, Hui Ma, and Zenglin Shi. Lora in lora: Towards parameter-efficient architecture expansion for continual visual instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 19978–19986, 2026.

[4] Junwei Chen, Pengpeng Qiao, Wei Yang Bryan Lim, Manjiang Yu, and Yang Cao. Privacy-preserving llm preference alignment: A systematic survey. *TechRxiv*, 2026.

[5] Tianran Chen, Jiarui Chen, Baoquan Zhang, Zhehao Yu, Shidong Chen, Rui Ye, Xutao Li, and Yunming Ye. Sensitivity-aware efficient fine-tuning via compact dynamic-rank adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9655–9664, 2025.

[6] Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, and Gauri Joshi. Heterogeneous lora for federated fine-tuning of on-device foundation models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 12903–12913, 2024.

[7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023.

[8] Ngo Cong-Lem, Ali Soyoof, and Diki Tsering. A systematic review of the limitations and associated opportunities of chatgpt. *International Journal of Human–Computer Interaction*, 41 (7):3851–3866, 2025.

[9] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instructiontuned llm. 2023.

[10] Chenghao Fan, Zhenyi Lu, Sichen Liu, Chengfeng Gu, Xiaoye Qu, Wei Wei, and Yu Cheng. Make lora great again: Boosting lora with adaptive singular values and mixture-of-experts optimization alignment. In *International Conference on Machine Learning*, pages 15804–15832. PMLR, 2025.

[11] Xiuwen Fang and Mang Ye. Towards robust parameter-efficient fine-tuning for federated learning. In *Advances in Neural Information Processing Systems*, 2025.

[12] Zhidong Gao, Zhenxiao Zhang, Yuanxiong Guo, and Yanmin Gong. Federated adaptive fine-tuning of large language models with heterogeneous quantization and lora. In *Ieee infocom 2025-ieee conference on computer communications*, pages 1–10. IEEE, 2025.

[13] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

[15] Jiangpeng He, Zhihao Duan, and Fengqing Zhu. Cl-lora: Continual low-rank adaptation for rehearsal-free class-incremental learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30534–30544, 2025.

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[17] Xinhao Huang, You-Liang Huang, and Zeyi Wen. Sola: Leveraging soft activation sparsity and low-rank decomposition for large language model compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17494–17502, 2025.

[18] Minsoo Kim, Sihwa Lee, Wonyong Sung, and Jungwook Choi. Ra-lora: Rank-adaptive parameter-efficient fine-tuning for accurate 2-bit quantized large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15773–15786, 2024.

[19] Jabin Koo, Minwoo Jang, and Jungseul Ok. 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, 2025.

[20] Xun Li, Lei Liu, Yuzhou Liu, Yu Zhao, Peng Zhang, and Huaxiao Liu. Multimodal fusion for android malware detection based on large pre-trained models. *IEEE Transactions on Software Engineering*, 51(5):1569–1590, 2025.

[21] Zequan Liu, Jiawen Lyn, Wei Zhu, Xing Tian, and Yvette Graham. Alora: Allocating low-rank adaptation for fine-tuning large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 622–641, 2024.

[22] Yijun Lu, Zihan Fang, Pengpeng Qiao, Zheng Lin, Jing Yang, Yuxin Zhang, Lip Yee, Zhe Chen, and Jun Luo. Conflict-aware federated fine-tuning of large language models with mixture-of-experts. In *2026 6th International Conference on Computer Science and Blockchain (CCSB)*, pages 430–434. IEEE, 2026.

[23] Yulong Mao, Kaiyu Huang, Changhao Guan, Ganglin Bao, Fengran Mo, and Jinan Xu. Dora: Enhancing parameter-efficient fine-tuning with dynamic rank distribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11662–11675, 2024.

[24] Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. Evaluation and benchmarking of llm agents: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6129–6139, 2025.

[25] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72, 2025.

[26] Wanyi Ning, Jingyu Wang, Qi Qi, Haifeng Sun, Daixuan Cheng, Cong Liu, Lei Zhang, Zirui Zhuang, and Jianxin Liao. Federated fine-tuning on heterogeneous loras with error-compensated aggregation. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

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

[28] Pengpeng Qiao, Kangfei Zhao, Bei Bi, Zhiwei Zhang, Ye Yuan, and Guoren Wang. Feed: Towards personalization-effective federated learning. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 1779–1791. IEEE, 2024.

[29] Zhiwei Tang, Yanmeng Wang, and Tsung-Hui Chang. z-signfedavg: A unified stochastic sign-based compression for federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15301–15309, 2024.

[30] Shujuan Tian, Jie Wu, Wei Li, Yanchun Li, Han Wang, Pengpeng Qiao, and Young-June Choi. Meta-pfl: Meta-based personalized federated learning against gradient inversion attacks. *IEEE Transactions on Dependable and Secure Computing*, 2026.

[31] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355, 2018.

[32] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 5085–5109, 2022.

[33] Zhuo Wang, Wen Wu, Guoqing Wang, Guangze Ye, and Zhenxiao Cheng. Metaeval: Measuring the discrimination of benchmarks for efficient llm evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33773–33781, 2026.

[34] Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *Advances in Neural Information Processing Systems*, 37:22513–22533, 2024.

[35] Fei Wu, Jia Hu, Geyong Min, and Shiqiang Wang. Adaptive rank allocation for federated parameter-efficient fine-tuning of language models. *IEEE Transactions on Computers*, 2026.

[36] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.

[37] Yunlu Yan, Chun-Mei Feng, Wangmeng Zuo, Rick Siow Mong Goh, Yong Liu, and Lei Zhu. Federated residual low-rank adaptation of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[38] Yaming Yang, Dilxat Muhtar, Yelong Shen, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Weiwei Deng, Feng Sun, Qi Zhang, et al. Mtl-lora: Low-rank adaptation for multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22010–22018, 2025.

[39] Dacao Zhang, Kun Zhang, Shimao Chu, Le Wu, Xin Li, and Si Wei. More: A mixture of low-rank experts for adaptive multi-task learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1311–1324, 2025.

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

[41] Ruiyi Zhang, Rushi Qiang, Sai Ashish Somayajula, and Pengtao Xie. Autolora: Automatically tuning matrix ranks in low-rank adaptation based on meta learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5048–5060, 2024.

[42] Zikai Zhang, Rui Hu, and Jiahao Xu. Heterogeneous federated fine-tuning with parallel one-rank adaptation. In *The Fourteenth International Conference on Learning Representations*, 2026.

[43] Xiangkai Zhu, Yeyu Yan, Pengpeng Qiao, Tingrui Pei, Yanchun Li, and Saiqin Long. Fedrahi: Reliability-aware hierarchical collaboration for federated graph foundation models. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6958–6968, 2026.

A Algorithms

The algorithm of proposed FedAbA is summarized in Algorithm 1.

Table 6: Notation used throughout the paper.

Symbol	Meaning
$N, \mathcal{S}^t$	total number of clients; active client set at round t
$n_i, p_i^t = n_i / \sum_{j\in\mathcal{S}^t} n_j$	client i's sample count; data-size weight
$r_i, r_{\max}$	client i's LoRA rank; maximum rank used by the server
(A_i^t, B_i^t)	client i's **LoRA factors** at round t, shapes (r_i, n) and (m, r_i)
$\Delta W_i^t = sB_i^t A_i^t$	client i's **local matrix update**
(A_g^t, B_g^t)	**global LoRA factors** at round t, shapes $(r_{\max}, n)$ and $(m, r_{\max})$
$Q_g^{t-1} = (A_g^{t-1})^\top$	**global reference basis** (round t aggregates against this)
$Q_{\mathrm{ref},i}^{t-1} = Q_g^{t-1}[:, 1{:}r_i]$	client i's **reference prefix**
Q_i^t	client i's **local basis** (thin-QR factor of $(A_i^t)^\top$)
$\bar{B}_i^t = B_i^t C_i^t$	coefficient matrix in the basis–coefficient form $\Delta W_i^t = s\bar{B}_i^t (Q_i^t)^\top$
T_i^t	orthogonal Procrustes **alignment matrix**
$c_i^t \in [0,1]$	**consistency score** (between local basis and reference prefix)
α_i^t	**consistency-aware aggregation weight**, $\alpha_i^t \propto p_i\, c_i^t$
$\Delta\widetilde{W}_i^t$	**basis-consistent reconstruction** of ΔW_i^t
$\Delta\bar{W}_g^t$	**aggregated basis-consistent global update**

Algorithm 1 FedAbA: Basis-Consistent Federated LoRA Aggregation

We assume $Z^t = \sum_j p_j\, c_j^t > 0$; the degenerate fallback $\alpha_i^t = p_i$ when $Z^t = 0$ is omitted for clarity.
Warm-start: obtain (A_g^0, B_g^0) by several standard federated LoRA aggregation rounds and rank-$r_{\max}$ SVD.
for round $t = 1, 2, \ldots, T$ **do**
  Server sends $A_g^{t-1}[1:r_i, :]$ and $B_g^{t-1}[:, 1:r_i]$ to each active client $i \in \mathcal{S}^t$.
  Each client locally fine-tunes both LoRA factors and uploads (A_i^t, B_i^t).
  Set $Q_g^{t-1} = (A_g^{t-1})^\top$.
  for each $i \in \mathcal{S}^t$ **do**
    Compute thin QR $(A_i^t)^\top = Q_i^t (C_i^t)^\top$ and set $\bar{B}_i^t = B_i^t C_i^t$.
    Set $Q_{\mathrm{ref},i}^{t-1} = Q_g^{t-1}[:, 1:r_i]$.
    Compute SVD $(Q_i^t)^\top Q_{\mathrm{ref},i}^{t-1} = U_i^t \Sigma_i^t (V_i^t)^\top$; set $T_i^t = U_i^t (V_i^t)^\top$.
    Compute $c_i^t = \frac{1}{r_i}\mathrm{tr}(\Sigma_i^t)$.
    Form $\widetilde{C}_i^t = [\bar{B}_i^t T_i^t,\, 0_{m\times(r_{\max}-r_i)}]$.
  end for
  For all $i \in \mathcal{S}^t$: $\alpha_i^t = \frac{p_i^t\, c_i^t}{\sum_{j\in\mathcal{S}^t} p_j^t\, c_j^t}$.
  Aggregate $C_g^t = \sum_{i\in\mathcal{S}^t} \alpha_i^t \widetilde{C}_i^t$.
  Compute coefficient-space SVD $C_g^t = U_c^t \Sigma_c^t (V_c^t)^\top$; set $\widehat{A}_g^t = (Q_g^{t-1} V_c^t)^\top$, $\widehat{B}_g^t = U_c^t \Sigma_c^t$.
  Compute block-wise signed permutation P_g^t via SVD gauge fixing (Section 2.3); set $A_g^t = P_g^t \widehat{A}_g^t$, $B_g^t = \widehat{B}_g^t (P_g^t)^\top$.
end for

B Details of the Motivation Experiments

Setup. Both observations in Section 2.2 use the same federated LoRA training backbone. We fine-tune LLaMA-2-7B with LoRA adapters (rank-scaled by α/r with α=16) inserted into `q_proj` and `v_proj` of every transformer block, giving 64 adapted layers. The Dolly-15K dataset is partitioned across N=8 clients via a Dirichlet split with concentration 0.5 over instruction categories; all clients participate in every round. Each round runs one local epoch of AdamW (lr 2×10^{-4},

batch size 8, sequence length 512) followed by FlexLoRA [1] server aggregation (compose $\Delta W_i = sB_iA_i$, average with data-size weights, rank-$r_{\max}$ truncated SVD, prefix dispatch). We deliberately use a naive aggregator as the training backbone so the measured quantities reflect the behavior of existing methods, not of FedAbA. Three rank regimes are studied: homogeneous r_i=8; mild $(4, 4, 8, 8, 8, 8, 16, 16)$; strong $(4, 4, 8, 8, 16, 16, 32, 32)$. All runs use the same random seed so that differences across regimes are attributable to rank configuration alone. Training runs on 4×A100 40 GB with FSDP, taking ≈ 6 minutes per round.

Measurements. Both subplots use the consistency score from Eq. (1). At each round t and each adapted layer ℓ, we save every active client's uploaded right factor $A^t_{i,\ell}$ and the server's previous-round global right factor $A^{t-1}_{g,\ell}$; we then compute $Q^t_{i,\ell}$ as the thin-QR Q-factor of $(A^t_{i,\ell})^\top$, $Q^{t-1}_{\mathrm{ref},i,\ell} = (A^{t-1}_{g,\ell})^\top[:, 1 : r_i]$, and $c^t_{i,\ell} = \frac{1}{r_i}\sum_k \sigma_k\big((Q^t_{i,\ell})^\top Q^{t-1}_{\mathrm{ref},i,\ell}\big)$.

For Observation 1 (subplot (a)), we record $c^{10}_{i,\ell}$ at round t=10 for every client i and every adapted layer ℓ, giving $N \times 64 = 512$ values per regime, which are aggregated into the violin in subplot (a). An intermediate round is chosen because A^t_i is not yet informative before round ~ 5 and FlexLoRA has partially converged after round ~ 30; the qualitative trend is stable for any $t \in [8, 20]$.

For Observation 2 (subplot (b)), we run the strong-heterogeneity regime for T=50 rounds and, for every round $t \in [5, 30]$ and every layer ℓ, plot $(p_i, c^t_{i,\ell})$ as a single point, where $p_i = n_i/\sum_j n_j$ is the (round-independent) data-size weight. The reported correlation ρ is the Pearson correlation between p_i and $c^t_{i,\ell}$ over all such points. All QR and SVD operations are done in double precision on CPU after detaching from the training graph.

C Additional Related Works

D Related Work

D.1 LoRA with Adaptive Rank

LoRA [16, 3] reduces the trainable parameters of large pretrained models [20, 25] by representing each weight update as a low-rank product. Since a fixed rank may be suboptimal across layers, tasks, or clients, several works study adaptive-rank LoRA [17, 18, 10, 23, 5, 39]. AdaLoRA [40] allocates rank budgets according to parameter importance under a global parameter constraint, while subsequent LoRA variants explore routing, pruning, or decomposing LoRA components to improve parameter efficiency [21, 23]. These methods mainly focus on centralized fine-tuning or single-model adaptation, where rank allocation is optimized within one training process.

D.2 Federated Learning with LoRA

Federated fine-tuning with LoRA [2, 16, 41, 38, 15, 22, 4] has been widely studied to reduce the communication and computation cost of adapting large language models. A direct application of FedAvg to LoRA factors averages the two factors separately, but this does not generally match the average of the composed updates B_iA_i, leading to aggregation discordance [34]. To avoid this issue, FFA-LoRA freezes one LoRA factor and aggregates the other, while FedSA-LoRA [14] analyzes the asymmetric roles of the LoRA factors and shares only the factor associated with general knowledge. LoRA-A^2 [19] alternates which factor is trained and combines this with adaptive rank selection. FRLoRA [37] instead accumulates residual low-rank updates into the model weights and reinitializes local LoRA factors in the principal singular space to mitigate constrained parameter space and client drift. RFedLR [11] studies robust federated PEFT under noisy labels, combining sensitivity-aware local tuning with adaptive federated LoRA aggregation.

Another line of work targets heterogeneous client ranks [6, 12, 35]. HetLoRA [6] supports different client ranks through zero-padding and truncation. FLoRA [34] stacks client LoRA factors so that the product of the stacked factors exactly represents the sum of local updates, naturally supporting heterogeneous ranks but increasing the effective global rank. FlexLoRA [1] and ECLoRA [26] compose each local update into a full-size matrix, aggregate these matrix updates, and redistribute rank-specific factors by SVD, thereby leveraging heterogeneous client resources. Fed-PLoRA [42]

further reformulates LoRA as multiple parallel one-rank adapters to reduce initialization and aggregation noise under heterogeneous client ranks. These methods address rank mismatch and aggregation discordance, but they do not explicitly reconcile the client-specific local bases against a shared reference before aggregation. FedAbA is complementary: it first reconstructs each local update under a shared global reference basis, weights updates by both data size and basis consistency, and then performs an exact coefficient-space SVD refresh with gauge fixing for stable rank-prefix dispatch.

E Theoretical Proofs

This appendix provides complete statements and proofs for the theoretical results in Section 3, together with two formal results deferred from the main text: the exact-SVD-refresh, gauge-invariance, and span-invariance properties (Section E.3, referenced from Section 3.1) and the variational characterization of the consistency-aware weighting rule (Section E.6, referenced from Remark 3.6). We restate each proposition, theorem, and assumption before its proof so that the appendix is self-contained.

Notation. After the thin QR decomposition in Eq. (2), the local matrix update of client i at round t can be written as

$$\Delta W_i^t = s\,\bar{B}_i^t (Q_i^t)^\top, \qquad (Q_i^t)^\top Q_i^t = I_{r_i}, \tag{13}$$

where $\bar{B}_i^t = B_i^t C_i^t$. The reference prefix for client i is

$$Q_{\mathrm{ref},i}^{t-1} = Q_g^{t-1}[:, 1:r_i]. \tag{14}$$

The orthogonal Procrustes alignment is

$$T_i^t = \arg\min_{T \in \mathcal{O}(r_i)} \|Q_i^t T - Q_{\mathrm{ref},i}^{t-1}\|_F^2. \tag{15}$$

The basis-consistent local matrix update is

$$\Delta\widetilde{W}_i^t = s\,\bar{B}_i^t T_i^t (Q_{\mathrm{ref},i}^{t-1})^\top. \tag{16}$$

Writing

$$(Q_i^t)^\top Q_{\mathrm{ref},i}^{t-1} = U_i^t \Sigma_i^t (V_i^t)^\top, \tag{17}$$

the consistency score is

$$c_i^t = \frac{1}{r_i}\,\mathrm{tr}(\Sigma_i^t) = \frac{1}{r_i}\sum_{j=1}^{r_i} \sigma_{i,j}^t. \tag{18}$$

Unless otherwise stated, we assume that Q_i^t and $Q_{\mathrm{ref},i}^{t-1}$ have orthonormal columns. We also assume

$$\sum_{j \in \mathcal{S}^t} p_j^t\, c_j^t > 0$$

whenever the consistency-aware weight

$$\alpha_i^t = \frac{p_i^t\, c_i^t}{\sum_{j \in \mathcal{S}^t} p_j^t\, c_j^t}$$

is used.

E.1 Factor Rotation Invariance (Proposition 3.1)

Proposition E.1 (Factor rotation invariance; restatement of Proposition 3.1)**.** *For any orthogonal matrix $T \in \mathcal{O}(r_i)$,*

$$s\,\bar{B}_i^t T (Q_i^t T)^\top = s\,\bar{B}_i^t (Q_i^t)^\top = \Delta W_i^t. \tag{19}$$

Proof. Since $T \in \mathcal{O}(r_i)$, we have $TT^\top = I_{r_i}$. Therefore,

$$\begin{aligned} s\,\bar{B}_i^t T (Q_i^t T)^\top &= s\,\bar{B}_i^t T\, T^\top (Q_i^t)^\top && (20)\\ &= s\,\bar{B}_i^t (TT^\top)(Q_i^t)^\top && (21)\\ &= s\,\bar{B}_i^t (Q_i^t)^\top && (22)\\ &= \Delta W_i^t, && (23) \end{aligned}$$

where the last equality follows from Eq. (13). □

Proposition E.1 shows that rotating both the coefficient matrix and the local basis is only a reparameterization. It changes the LoRA factors but leaves the composed matrix update unchanged. Therefore, merely rotating local factors before aggregation cannot change the server-side aggregate; FedAbA must explicitly replace the local basis with the reference prefix, as in Eq. (16).

E.2 Shared-Basis Representation (Proposition 3.2)

Lemma E.2 (Nested prefix consistency)**.** *For any two active clients $i, j \in \mathcal{S}^t$ with $r_i \le r_j$,*

$$Q_{\mathrm{ref},i}^{t-1} = Q_{\mathrm{ref},j}^{t-1}[:, 1 : r_i]. \tag{24}$$

Proof. By definition,

$$Q_{\mathrm{ref},i}^{t-1} = Q_g^{t-1}[:, 1 : r_i], \qquad Q_{\mathrm{ref},j}^{t-1} = Q_g^{t-1}[:, 1 : r_j].$$

Since $r_i \le r_j$, the first r_i columns of $Q_{\mathrm{ref},j}^{t-1}$ are exactly the first r_i columns of Q_g^{t-1}, which are $Q_{\mathrm{ref},i}^{t-1}$. □

Proposition E.3 (Shared-basis representation; restatement of Proposition 3.2)**.** *Define the zero-padded coefficient matrix*

$$\widetilde{C}_i^t = \left[\bar{B}_i^t T_i^t,\ 0_{m\times(r_{\max}-r_i)}\right] \in \mathbb{R}^{m\times r_{\max}}. \tag{25}$$

Then

$$\Delta\widetilde{W}_i^t = s\,\widetilde{C}_i^t (Q_g^{t-1})^\top, \qquad \Delta\bar{W}_g^t = s\left(\sum_{i\in\mathcal{S}^t} \alpha_i^t \widetilde{C}_i^t\right)(Q_g^{t-1})^\top. \tag{26}$$

Consequently,

$$\mathrm{rank}(\Delta\bar{W}_g^t) \le r_{\max}. \tag{27}$$

Proof. Since

$$Q_{\mathrm{ref},i}^{t-1} = Q_g^{t-1}[:, 1 : r_i],$$

we can write

$$(Q_{\mathrm{ref},i}^{t-1})^\top = \left[I_{r_i}\ \ 0_{r_i\times(r_{\max}-r_i)}\right](Q_g^{t-1})^\top. \tag{28}$$

Substituting Eq. (28) into Eq. (16),

$$\begin{aligned}
\Delta\widetilde{W}_i^t &= s\,\bar{B}_i^t T_i^t (Q_{\mathrm{ref},i}^{t-1})^\top && (29)\\
&= s\,\bar{B}_i^t T_i^t \left[I_{r_i}\ \ 0_{r_i\times(r_{\max}-r_i)}\right](Q_g^{t-1})^\top && (30)\\
&= s\left[\bar{B}_i^t T_i^t,\ 0_{m\times(r_{\max}-r_i)}\right](Q_g^{t-1})^\top && (31)\\
&= s\,\widetilde{C}_i^t (Q_g^{t-1})^\top. && (32)
\end{aligned}$$

Aggregating over active clients gives

$$\begin{aligned}
\Delta\bar{W}_g^t &= \sum_{i\in\mathcal{S}^t} \alpha_i^t \Delta\widetilde{W}_i^t && (33)\\
&= s\left(\sum_{i\in\mathcal{S}^t} \alpha_i^t \widetilde{C}_i^t\right)(Q_g^{t-1})^\top. && (34)
\end{aligned}$$

Let

$$C_g^t = \sum_{i\in\mathcal{S}^t} \alpha_i^t \widetilde{C}_i^t \in \mathbb{R}^{m\times r_{\max}}.$$

Then

$$\Delta\bar{W}_g^t = s C_g^t (Q_g^{t-1})^\top,$$

so $\mathrm{rank}(\Delta\bar{W}_g^t) \le r_{\max}$. □

Proposition E.3 is a per-round statement: at round t, all reconstructed client updates are expressed in the global reference basis Q_g^{t-1}. After aggregation, the server may refactorize $\Delta\bar{W}_g^t$ and obtain a refreshed basis Q_g^t, which becomes the global reference basis for the next round.

E.3 Exact SVD Refresh, Gauge Invariance, and Span Invariance

This subsection provides the formal statements deferred from Section 3.1. It establishes that the SVD refresh in Eq. (10) and the gauge fixing in Eq. (11) together preserve the aggregated update exactly and keep the global basis within its warm-started span, justifying the "bias-free" claim used in Section 3.3.

Proposition E.4 (Exact SVD refresh)**.** *Let*

$$C_g^t = \sum_{i \in \mathcal{S}^t} \alpha_i^t \widetilde{C}_i^t \in \mathbb{R}^{m \times r_{\max}},$$

and let its SVD be

$$C_g^t = U_c^t \Sigma_c^t (V_c^t)^\top. \tag{35}$$

If

$$\widehat{A}_g^t = (Q_g^{t-1} V_c^t)^\top, \qquad \widehat{B}_g^t = U_c^t \Sigma_c^t, \tag{36}$$

then

$$s\, \widehat{B}_g^t \widehat{A}_g^t = \Delta \bar{W}_g^t. \tag{37}$$

Proof. By Proposition E.3,

$$\Delta \bar{W}_g^t = s C_g^t (Q_g^{t-1})^\top.$$

Substituting the SVD of C_g^t,

$$\begin{aligned} \Delta \bar{W}_g^t &= s U_c^t \Sigma_c^t (V_c^t)^\top (Q_g^{t-1})^\top && (38)\\ &= s U_c^t \Sigma_c^t (Q_g^{t-1} V_c^t)^\top && (39)\\ &= s\, \widehat{B}_g^t \widehat{A}_g^t. && (40) \end{aligned}$$

The SVD refresh therefore exactly refactorizes the reconstructed aggregate, with no additional projection error. □

Definition E.5 (Block-wise SVD gauge fixing)**.** Let $\widehat{Q}_g^t = (\widehat{A}_g^t)^\top$ be the provisional right basis after SVD refresh. The SVD is sign-ambiguous and, inside nearly degenerate singular-value blocks, order-ambiguous. Given a relative threshold $\tau_{\mathrm{svd}} > 0$, we form blocks of indices whose singular values are nearly equal. Within each block, we match provisional directions in $\widehat{Q}_g^t$ to directions in Q_g^{t-1} by maximum absolute cosine similarity and flip signs so that matched cosines are nonnegative. This produces a block-wise signed permutation matrix P_g^t.

Proposition E.6 (Gauge-fixing invariance)**.** *Let $(\widehat{A}_g^t, \widehat{B}_g^t)$ be provisional global factors satisfying $s\, \widehat{B}_g^t \widehat{A}_g^t = \Delta \bar{W}_g^t$. Let P_g^t be any orthogonal matrix, including the block-wise signed permutation from Definition E.5, and define*

$$A_g^t = P_g^t \widehat{A}_g^t, \qquad B_g^t = \widehat{B}_g^t (P_g^t)^\top. \tag{41}$$

Then

$$s B_g^t A_g^t = s \widehat{B}_g^t \widehat{A}_g^t = \Delta \bar{W}_g^t. \tag{42}$$

Proof. Since P_g^t is orthogonal, $(P_g^t)^\top P_g^t = I$. Therefore,

$$\begin{aligned} s B_g^t A_g^t &= s \widehat{B}_g^t (P_g^t)^\top P_g^t \widehat{A}_g^t && (43)\\ &= s \widehat{B}_g^t \widehat{A}_g^t && (44)\\ &= \Delta \bar{W}_g^t. && (45) \end{aligned}$$

□

The gauge-fixing step therefore changes only the factor representation, not the composed global matrix update.

Proposition E.7 (Span invariance of the default FedAbA refresh; formal version of Remark 2.1)**.** *Let*

$$\widehat{A}_g^t = (Q_g^{t-1} V_c^t)^\top$$

be the provisional right factor after coefficient-space SVD refresh, and let

$$A_g^t = P_g^t \widehat{A}_g^t$$

be the gauge-fixed right factor, where both V_c^t and P_g^t are orthogonal. Then

$$Q_g^t := (A_g^t)^\top = Q_g^{t-1} O_g^t$$

for some $O_g^t \in \mathcal{O}(r_{\max})$. Consequently,

$$\operatorname{span}(Q_g^t) = \operatorname{span}(Q_g^{t-1}).$$

Proof. By definition,

$$Q_g^t = (A_g^t)^\top = (P_g^t \widehat{A}_g^t)^\top = (\widehat{A}_g^t)^\top (P_g^t)^\top.$$

Since $(\widehat{A}_g^t)^\top = Q_g^{t-1} V_c^t$,

$$Q_g^t = Q_g^{t-1} V_c^t (P_g^t)^\top.$$

Let $O_g^t := V_c^t (P_g^t)^\top$. Because both V_c^t and P_g^t are orthogonal, O_g^t is also orthogonal. Hence $Q_g^t = Q_g^{t-1} O_g^t$, which implies $\operatorname{span}(Q_g^t) = \operatorname{span}(Q_g^{t-1})$. □

This proposition shows that the default FedAbA refresh changes the global basis only within the span of the previous global basis. It does not introduce new basis directions after warm-start.

E.4 Reconstruction Residual (Theorem 3.3)

Theorem E.8 (Reconstruction residual is controlled by basis consistency; restatement of Theorem 3.3)**.** *The minimum basis alignment residual satisfies*

$$\min_{T \in \mathcal{O}(r_i)} \|Q_i^t T - Q_{\mathrm{ref},i}^{t-1}\|_F^2 = 2r_i(1 - c_i^t), \tag{46}$$

and the basis-consistent local matrix update satisfies

$$\|\Delta W_i^t - \Delta \widetilde{W}_i^t\|_F \le s \|\bar{B}_i^t\|_2 \sqrt{2r_i(1 - c_i^t)}. \tag{47}$$

Proof. We prove the two claims separately.

Step 1: exact alignment residual. For any $T \in \mathcal{O}(r_i)$,

$$\|Q_i^t T - Q_{\mathrm{ref},i}^{t-1}\|_F^2 = \operatorname{tr}\left[(Q_i^t T - Q_{\mathrm{ref},i}^{t-1})^\top (Q_i^t T - Q_{\mathrm{ref},i}^{t-1})\right] \tag{48}$$

$$= \operatorname{tr}\left(T^\top (Q_i^t)^\top Q_i^t T\right) + \operatorname{tr}\left((Q_{\mathrm{ref},i}^{t-1})^\top Q_{\mathrm{ref},i}^{t-1}\right) \tag{49}$$

$$- 2 \operatorname{tr}\left(T^\top (Q_i^t)^\top Q_{\mathrm{ref},i}^{t-1}\right). \tag{50}$$

Since both Q_i^t and $Q_{\mathrm{ref},i}^{t-1}$ have orthonormal columns, $(Q_i^t)^\top Q_i^t = I_{r_i}$ and $(Q_{\mathrm{ref},i}^{t-1})^\top Q_{\mathrm{ref},i}^{t-1} = I_{r_i}$. Thus,

$$\|Q_i^t T - Q_{\mathrm{ref},i}^{t-1}\|_F^2 = 2r_i - 2\operatorname{tr}(T^\top M_i^t), \qquad M_i^t := (Q_i^t)^\top Q_{\mathrm{ref},i}^{t-1}. \tag{51}$$

Let $M_i^t = U_i^t \Sigma_i^t (V_i^t)^\top$. By von Neumann's trace inequality,

$$\operatorname{tr}(T^\top M_i^t) \le \operatorname{tr}(\Sigma_i^t)$$

for every orthogonal T, with equality attained by $T_i^t = U_i^t (V_i^t)^\top$. Therefore,

$$\min_{T \in \mathcal{O}(r_i)} \|Q_i^t T - Q_{\mathrm{ref},i}^{t-1}\|_F^2 = 2r_i - 2\operatorname{tr}(\Sigma_i^t) = 2r_i(1 - c_i^t). \tag{52}$$

Step 2: matrix reconstruction residual. For any $T \in \mathcal{O}(r_i)$, define $\Delta W_i^t(T) := s\bar{B}_i^t T(Q_{\mathrm{ref},i}^{t-1})^\top$. Since $TT^\top = I_{r_i}$,

$$\Delta W_i^t = s\bar{B}_i^t(Q_i^t)^\top = s\bar{B}_i^t TT^\top(Q_i^t)^\top = s\bar{B}_i^t T(Q_i^t T)^\top. \tag{53}$$

Hence

$$\Delta W_i^t - \Delta W_i^t(T) = s\bar{B}_i^t T(Q_i^t T)^\top - s\bar{B}_i^t T(Q_{\mathrm{ref},i}^{t-1})^\top \tag{54}$$

$$= s\bar{B}_i^t T(Q_i^t T - Q_{\mathrm{ref},i}^{t-1})^\top. \tag{55}$$

Taking Frobenius norms and using $\|XY\|_F \le \|X\|_2\|Y\|_F$,

$$\|\Delta W_i^t - \Delta W_i^t(T)\|_F \le s\|\bar{B}_i^t T\|_2 \|Q_i^t T - Q_{\mathrm{ref},i}^{t-1}\|_F \tag{56}$$

$$= s\|\bar{B}_i^t\|_2 \|Q_i^t T - Q_{\mathrm{ref},i}^{t-1}\|_F, \tag{57}$$

where the last equality uses the orthogonality of T. Setting $T = T_i^t$ and using Step 1,

$$\|\Delta W_i^t - \Delta\widetilde{W}_i^t\|_F \le s\|\bar{B}_i^t\|_2\sqrt{2r_i(1-c_i^t)}.$$

□

E.5 Consistency-Aware Weighting (Assumption 3.4 and Theorem 3.5)

Assumption E.9 (Uniform coefficient-radius bound; restatement of Assumption 3.4)**.** There exists $\kappa_t > 0$ such that

$$s\|\bar{B}_i^t\|_2\sqrt{r_i} \le \kappa_t \tag{58}$$

for every $i \in \mathcal{S}^t$.

Theorem E.10 (Consistency-aware weighting tightens a reconstruction bound; restatement of Theorem 3.5)**.** *Let*

$$b_i^t := \kappa_t\sqrt{2(1-c_i^t)}. \tag{59}$$

Under Assumption E.9,

$$\|\Delta W_i^t - \Delta\widetilde{W}_i^t\|_F \le b_i^t \tag{60}$$

for every $i \in \mathcal{S}^t$. If $\sum_{j\in\mathcal{S}^t} p_j^t\, c_j^t > 0$, then

$$\sum_{i\in\mathcal{S}^t} \alpha_i^t\, b_i^t \le \sum_{i\in\mathcal{S}^t} p_i^t\, b_i^t. \tag{61}$$

Proof. By Theorem E.8,

$$\|\Delta W_i^t - \Delta\widetilde{W}_i^t\|_F \le s\|\bar{B}_i^t\|_2\sqrt{2r_i(1-c_i^t)}.$$

Using Assumption E.9,

$$s\|\bar{B}_i^t\|_2\sqrt{2r_i(1-c_i^t)} = \sqrt{2(1-c_i^t)}\cdot s\|\bar{B}_i^t\|_2\sqrt{r_i} \le \kappa_t\sqrt{2(1-c_i^t)} = b_i^t. \tag{62}$$

This proves Eq. (60).

It remains to prove the weighted inequality. Define $b(c) = \kappa_t\sqrt{2(1-c)}$, which is non-increasing on $[0,1]$. Let I and J be two independent random indices drawn from the distribution $\{p_i\}_{i\in\mathcal{S}^t}$. Since $c \mapsto c$ is non-decreasing and $c \mapsto b(c)$ is non-increasing,

$$(c_I^t - c_J^t)(b_I^t - b_J^t) \le 0.$$

Taking expectation,

$$\mathbb{E}[(c_I^t - c_J^t)(b_I^t - b_J^t)] \le 0. \tag{63}$$

Expanding and using independence of I, J,

$$\mathbb{E}[(c_I^t - c_J^t)(b_I^t - b_J^t)] = 2\mathbb{E}[c_I^t b_I^t] - 2\mathbb{E}[c_I^t]\mathbb{E}[b_I^t]. \tag{64}$$

Thus, $\mathbb{E}[c_I^t b_I^t] \le \mathbb{E}[c_I^t]\mathbb{E}[b_I^t]$, i.e.,

$$\sum_{i\in\mathcal{S}^t} p_i\, c_i^t\, b_i^t \le \left(\sum_{i\in\mathcal{S}^t} p_i^t\, c_i^t\right)\left(\sum_{i\in\mathcal{S}^t} p_i^t\, b_i^t\right). \tag{65}$$

Since $\sum_i p_i^t\, c_i^t > 0$, dividing by this positive denominator gives

$$\frac{\sum_i p_i^t\, c_i^t\, b_i^t}{\sum_i p_i^t\, c_i^t} \le \sum_i p_i^t\, b_i^t,$$

i.e., $\sum_i \alpha_i^t b_i^t \le \sum_i p_i^t\, b_i^t$. □

Remark on equality. If all consistency scores are equal on the support of p, then $\alpha_i^t = p_i^t$ and Eq. (61) holds with equality. If all $p_i^t > 0$ and the consistency scores are not all equal, then the inequality is strict.

E.6 Variational Characterization of the Weighting Rule (Remark 3.6)

This subsection formalizes Remark 3.6 from the main text, showing that the consistency-aware weighting rule α_i^t admits a variational characterization as a KL-regularized log-consistency maximizer.

Proposition E.11 (Variational characterization of α_i^t)**.** *Assume $p_i^t > 0$ and $c_i^t > 0$ for every $i \in \mathcal{S}^t$. Then*

$$\alpha_i^t = \frac{p_i^t\, c_i^t}{\sum_{j\in\mathcal{S}^t} p_j^t\, c_j^t} \tag{66}$$

is the unique maximizer of

$$\max_{\alpha\in\Delta_{|\mathcal{S}^t|}} \left\{\sum_{i\in\mathcal{S}^t} \alpha_i \log c_i^t - \mathrm{KL}(\alpha\|p)\right\}, \tag{67}$$

where $\mathrm{KL}(\alpha\|p) = \sum_i \alpha_i \log \frac{\alpha_i}{p_i}$.

Proof. The Lagrangian is

$$\mathcal{L}(\alpha,\lambda) = \sum_i \alpha_i \log c_i^t - \sum_i \alpha_i \log \frac{\alpha_i}{p_i} + \lambda\left(\sum_i \alpha_i - 1\right).$$

Differentiating with respect to α_i,

$$\frac{\partial\mathcal{L}}{\partial\alpha_i} = \log c_i^t - \left(\log\frac{\alpha_i}{p_i} + 1\right) + \lambda.$$

Setting the derivative to zero, $\log\frac{\alpha_i}{p_i} = \log c_i^t + \lambda - 1$, so $\alpha_i = e^{\lambda-1} p_i^t\, c_i^t$. The simplex constraint gives $e^{\lambda-1} = 1/\sum_j p_j\, c_j^t$, hence

$$\alpha_i = \frac{p_i^t\, c_i^t}{\sum_j p_j^t\, c_j^t}.$$

The objective in Eq. (67) is a linear function of α minus $\mathrm{KL}(\alpha\|p)$. Since $\mathrm{KL}(\alpha\|p)$ is strictly convex on the relative interior of the simplex when $p_i^t > 0$, the objective is strictly concave, and the stationary point is the unique maximizer. □

E.7 Idealized Optimization Interpretation (Assumption 3.7 and Theorem 3.8)

Assumption E.12 (Idealized smooth stochastic descent model; restatement of Assumption 3.7)**.** The objective F is L-smooth, and the raw aggregate $G^t = \sum_{i\in\mathcal{S}^t} \alpha_i^t \Delta W_i^t$ satisfies

$$\mathbb{E}[G^t \mid W^t] = \nabla F(W^t), \qquad \mathbb{E}[\|G^t - \nabla F(W^t)\|_F^2 \mid W^t] \le \sigma^2. \tag{68}$$

Theorem E.13 (Convergence with reconstruction bias; restatement of Theorem 3.8)**.** *Define*

$$\widetilde{G}^t = \sum_{i \in \mathcal{S}^t} \alpha_i^t \Delta \widetilde{W}_i^t, \qquad e^t = \widetilde{G}^t - G^t. \tag{69}$$

Consider the idealized server update

$$W^{t+1} = W^t - \eta \widetilde{G}^t \tag{70}$$

with $\eta \leq 1/(4L)$. Under Assumption E.12,

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(W^t)\|_F^2 \leq \frac{4(F(W^0) - F^\star)}{\eta T} + 4L\eta\sigma^2 + \frac{2 + 4L\eta}{T}\sum_{t=0}^{T-1} \mathbb{E}\|e^t\|_F^2. \tag{71}$$

Moreover,

$$\|e^t\|_F \leq \sum_{i \in \mathcal{S}^t} \alpha_i^t \, s \|\bar{B}_i^t\|_2 \sqrt{2r_i(1 - c_i^t)}. \tag{72}$$

Proof. We first bound the reconstruction perturbation. By definition,

$$e^t = \sum_{i \in \mathcal{S}^t} \alpha_i^t (\Delta \widetilde{W}_i^t - \Delta W_i^t). \tag{73}$$

Using the triangle inequality and $\alpha_i^t \geq 0$,

$$\|e^t\|_F \leq \sum_{i \in \mathcal{S}^t} \alpha_i^t \|\Delta \widetilde{W}_i^t - \Delta W_i^t\|_F.$$

By Theorem E.8,

$$\|\Delta \widetilde{W}_i^t - \Delta W_i^t\|_F \leq s\|\bar{B}_i^t\|_2 \sqrt{2r_i(1 - c_i^t)},$$

which proves Eq. (72).

We now prove the convergence bound. Since F is L-smooth,

$$F(W^{t+1}) \leq F(W^t) + \langle \nabla F(W^t), W^{t+1} - W^t \rangle + \frac{L}{2}\|W^{t+1} - W^t\|_F^2 \tag{74}$$

$$= F(W^t) - \eta \langle \nabla F(W^t), \widetilde{G}^t \rangle + \frac{L\eta^2}{2}\|\widetilde{G}^t\|_F^2. \tag{75}$$

Using $\widetilde{G}^t = G^t + e^t$,

$$F(W^{t+1}) \leq F(W^t) - \eta \langle \nabla F(W^t), G^t \rangle - \eta \langle \nabla F(W^t), e^t \rangle \tag{76}$$

$$+ \frac{L\eta^2}{2}\|G^t + e^t\|_F^2. \tag{77}$$

Using $\|G^t + e^t\|_F^2 \leq 2\|G^t\|_F^2 + 2\|e^t\|_F^2$,

$$F(W^{t+1}) \leq F(W^t) - \eta \langle \nabla F(W^t), G^t \rangle - \eta \langle \nabla F(W^t), e^t \rangle \tag{78}$$

$$+ L\eta^2\|G^t\|_F^2 + L\eta^2\|e^t\|_F^2. \tag{79}$$

Taking conditional expectation given W^t and using Assumption E.12,

$$\mathbb{E}[\langle \nabla F(W^t), G^t \rangle \mid W^t] = \|\nabla F(W^t)\|_F^2,$$

and

$$\mathbb{E}[\|G^t\|_F^2 \mid W^t] = \|\nabla F(W^t)\|_F^2 + \mathbb{E}[\|G^t - \nabla F(W^t)\|_F^2 \mid W^t] \tag{80}$$

$$\leq \|\nabla F(W^t)\|_F^2 + \sigma^2. \tag{81}$$

Thus,

$$\mathbb{E}[F(W^{t+1}) \mid W^t] \leq F(W^t) - \eta\|\nabla F(W^t)\|_F^2 - \eta\mathbb{E}[\langle \nabla F(W^t), e^t \rangle \mid W^t] \tag{82}$$

$$+ L\eta^2\|\nabla F(W^t)\|_F^2 + L\eta^2\sigma^2 + L\eta^2\mathbb{E}[\|e^t\|_F^2 \mid W^t]. \tag{83}$$

By Young's inequality,

$$-\langle \nabla F(W^t), e^t \rangle \leq \frac{1}{2}\|\nabla F(W^t)\|_F^2 + \frac{1}{2}\|e^t\|_F^2.$$

Substituting,

$$\mathbb{E}[F(W^{t+1}) \mid W^t] \leq F(W^t) - \eta\left(\frac{1}{2} - L\eta\right)\|\nabla F(W^t)\|_F^2 \tag{84}$$

$$+ L\eta^2\sigma^2 + \left(\frac{\eta}{2} + L\eta^2\right)\mathbb{E}[\|e^t\|_F^2 \mid W^t]. \tag{85}$$

If $\eta \leq 1/(4L)$, then $\frac{1}{2} - L\eta \geq \frac{1}{4}$, so

$$\frac{\eta}{4}\|\nabla F(W^t)\|_F^2 \leq F(W^t) - \mathbb{E}[F(W^{t+1}) \mid W^t] + L\eta^2\sigma^2 + \left(\frac{\eta}{2} + L\eta^2\right)\mathbb{E}[\|e^t\|_F^2 \mid W^t].$$

Taking full expectation and summing over $t = 0, \ldots, T-1$,

$$\frac{\eta}{4}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(W^t)\|_F^2 \leq F(W^0) - \mathbb{E}[F(W^T)] + TL\eta^2\sigma^2 \tag{86}$$

$$+ \left(\frac{\eta}{2} + L\eta^2\right)\sum_{t=0}^{T-1}\mathbb{E}\|e^t\|_F^2. \tag{87}$$

Using $F(W^T) \geq F^\star$ and dividing by $\eta T/4$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(W^t)\|_F^2 \leq \frac{4(F(W^0) - F^\star)}{\eta T} + 4L\eta\sigma^2 + \frac{2 + 4L\eta}{T}\sum_{t=0}^{T-1}\mathbb{E}\|e^t\|_F^2.$$

□

Why there is no projection bias. By Proposition E.3, $\widetilde{G}^t = sC_g^t(Q_g^{t-1})^\top$ for some $C_g^t \in \mathbb{R}^{m \times r_{\max}}$, so $\mathrm{rank}(\widetilde{G}^t) \leq r_{\max}$. Combined with Propositions E.4 and E.6, the rank-$r_{\max}$ SVD refresh and gauge fixing exactly represent $\widetilde{G}^t$. They change only the factor representation and the next-round global reference basis; they do not add a separate low-rank projection error.

F Additional Experimental Results

F.1 Results across Independent Runs

We compare FedAbA and Fed-PLoRA over three independent matched runs under the settings of the main evaluation. The repeated-run study covers MRPC F1, RTE accuracy, MMLU after Dolly-15K tuning, MT-Bench after Alpaca tuning, and Natural Instructions ROUGE-L and ROUGE-1. For metrics evaluated at every communication round, the final score of each run is averaged over the last five rounds. Table 7 reports the mean and sample standard deviation across the three run-level scores. Thus, the reported standard deviation measures variation across independent runs, rather than variation across the five evaluation rounds within a run.

Table 7: Repeated-run comparison: mean ± sample standard deviation over three independent matched runs. Dolly-15K and Alpaca denote training datasets, evaluated with MMLU and MT-Bench, respectively.

Method	MRPC F1	RTE Acc.	Dolly-15K MMLU	Alpaca MT-Bench	NI ROUGE-L	NI ROUGE-1
Fed-PLoRA	89.42 ± 0.45	81.63 ± 1.10	47.17 ± 0.58	4.21 ± 0.12	57.98 ± 0.46	65.66 ± 0.48
FedAbA	$\mathbf{90.28} \pm 0.64$	$\mathbf{83.28} \pm 0.82$	$\mathbf{47.87} \pm 0.40$	$\mathbf{4.31} \pm 0.14$	$\mathbf{58.37} \pm 0.30$	$\mathbf{66.05} \pm 0.63$

Table 8: Natural Instructions results over all clients and the fixed lowest-consistency 20% subset.

Method	All clients		Lowest-consistency 20%	
	ROUGE-L	ROUGE-1	ROUGE-L	ROUGE-1
Fed-PLoRA	58.1	65.8	54.8	62.4
FedAbA	**58.4**	**66.1**	**55.1**	**62.6**

F.2 Performance on Low-Consistency Clients

We examine performance on clients whose uploaded bases are less aligned with the reference in the Natural Instructions task-as-client setting. We rank clients by their average consistency in the default FedAbA run and select the lowest 20%. The same fixed subset is used to evaluate both FedAbA and Fed-PLoRA. Table 8 reports ROUGE-L and ROUGE-1 over all clients and over this subset.

Compared with Fed-PLoRA, FedAbA improves all-client ROUGE-L and ROUGE-1 by 0.3 points each. On the fixed low-consistency subset, the corresponding improvements are 0.3 and 0.2 points. The improvement over Fed-PLoRA therefore also appears in the average performance of this subset. This comparison evaluates the complete methods and does not isolate the effect of residual removal or establish improvement for every individual client.

F.3 Additional Backbone Evaluation

We evaluate additional backbones while retaining the Dolly-15K high-rank-heterogeneity protocol and the MMLU evaluation used in the main experiment. Qwen3-4B tests a model from another architecture family, while LLaMA-2-13B tests a larger model in the LLaMA-2 family. Table 9 includes the original LLaMA-2-7B results for reference and compares FedAbA with FlexLoRA, ECLoRA, Fed-PLoRA, and a FedAbA variant without basis-consistent reconstruction.

Table 9: Backbone comparison after federated fine-tuning on Dolly-15K, evaluated with MMLU.

Method	LLaMA-2-7B	Qwen3-4B	LLaMA-2-13B
FlexLoRA	45.9	57.0	51.7
ECLoRA	46.1	57.2	51.9
Fed-PLoRA	47.2	58.3	53.0
FedAbA w/o basis-consistent reconstruction	46.9	58.0	52.7
FedAbA	**47.8**	**59.0**	**53.7**

G Broader Impact

FedAbA aims to improve federated LLM fine-tuning under heterogeneous client capacities by stabilizing rank-prefix dispatch through basis-consistent aggregation, which may benefit privacy-preserving adaptation in domains where data ownership or regulation limits data sharing, particularly for resource-constrained clients that can only afford low-rank adapters.

H Limitations

Despite its effectiveness, FedAbA has several limitations. First, the default aggregation operates within the span of the warm-started global basis: the SVD refresh and gauge fixing evolve the global basis only by an orthogonal transform of the warm-start basis, which stabilizes rank-prefix semantics across rounds but cannot introduce genuinely new basis directions if the warm-start span is unrepresentative. Second, FedAbA aggregates only the basis-consistent reconstruction of each client's local matrix update and discards the corresponding residual; when the consistency score is small for low-rank clients in early rounds, this trades a non-negligible residual signal for cross-round prefix stability, and a principled mechanism for selectively recovering it without reintroducing inconsistency is left to future work.

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 main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: We detail the limitations of the framework in the appendix.

   Guidelines:

   - The answer [N/A] 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: The complete assumptions, lemmas, and proofs are provided in the appendix.

Guidelines:

- The answer [N/A] 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: The paper provides comprehensive information necessary to reproduce the main experimental results, ensuring transparency and replicability of the findings.

   Guidelines:

   - The answer [N/A] 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: Our code is available at https://anonymous.4open.science/r/FedAbA-6774

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all the training and test details in the appendix.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: We provide the information on the computer resources in appendix.

   Guidelines:

   - The answer [N/A] 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: The research conducted in this paper strictly adheres to the NeurIPS Code of Ethics. We ensured that all data used in our study was obtained with proper consent, and we adhered to principles of fairness, transparency, and accountability. The research does not involve any form of discrimination or harm to individuals or groups, and we made sure to conduct the study with full respect for privacy and confidentiality.

   Guidelines:

   - The answer [N/A] 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: [Yes]

    Justification: The paper discusses the potential positive societal impacts of the work performed, highlighting the benefits and advancements it can bring to the field and society at large in the appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper presents no such risks.

    Guidelines:

    - The answer [N/A] 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: We have appropriately credited and cited the creators or original owners of all assets, including code, data, and papers, used in our work. Additionally, the licenses and terms of use for these assets have been explicitly stated and fully respected.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not introduce any new assets.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This paper does not involve crowdsourcing or research involving human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] 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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.