BARYLORA: STATE-CONDITIONED PRODUCT BARYCENTERS FOR FEDERATED LOW-RANK ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated LoRA provides a communication-efficient way to adapt pretrained language models, but a shared adapter still requires the server to decide both what to aggregate and how to assign client contributions. Averaging low-rank factors can fail to aggregate the effective updates they encode: equivalent factorizations are non-unique, and averaging the factors separately introduces cross-client products. Meanwhile, sample-size weighting assigns each client the same relative mass at every adapted layer and therefore does not directly specify client-layer masses for a shared effective-update consensus. We introduce BaryLoRA, a state-conditioned product-space consensus framework that addresses these two mismatches jointly. BaryLoRA treats normalized effective updates as the consensus object and uses persistent optimization state to tilt a sample-size prior into client-layer contribution masses. The server then forms a weighted mean of the effective updates and returns its best rank-constrained approximation through a factorized computation that preserves the communication efficiency of low-rank adaptation. We distinguish the empirical state-to-mass design from the conditional properties of the resulting consensus. Under fixed positive state transforms and contribution masses, we establish uniqueness of the mass solution, invariance to equivalent LoRA factorizations, and optimality of the factorized rank-constrained realization. Experiments on language understanding, code generation, and mathematical reasoning tasks evaluate BaryLoRA against federated LoRA baselines.

1 INTRODUCTION

Federated learning trains a shared model while keeping client data local (McMahan et al., 2017). Statistical and system heterogeneity can make a uniformly aggregated update poorly matched to participating clients' local objectives (Li et al., 2020a; Karimireddy et al., 2020). For pretrained language models, exchanging full-model updates at every round is costly. Parameter-efficient fine-tuning addresses this cost by adapting a frozen backbone through a small trainable component; LoRA represents that component with two low-rank factors (Hu et al., 2022). In federated LoRA, clients communicate these factors for each adapted layer (Zhang et al., 2024). A server that builds a shared adapter must therefore make two coupled decisions: which object to aggregate and how much each client should contribute at each adapted layer. For LoRA, the first decision is whether to aggregate the factors or the effective updates they represent.

The factors provide a compact coordinate representation, whereas their product determines the change applied to the model. Equivalent factor pairs can represent the same effective update, and averaging the two factors separately introduces cross-client products that are absent from the submitted updates. Factor freezing, alternating optimization, and coordinate alignment address different aspects of this mismatch (Sun et al., 2024; Chen et al., 2025; Zhang et al., 2026a; Meng et al., 2026). These limitations have also motivated methods that aggregate represented updates while preserving an exact or compact federated realization. FLoRA forms an exact sum whose rank grows with participation, FlexLoRA handles heterogeneous rank budgets, and FedEx-LoRA preserves exact aggregation with a frozen-base correction (Wang et al., 2024; Bai et al., 2024; Singhal et al., 2025). Together, these methods establish effective updates as a natural aggregation object. How-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ever, a shared rank-r adapter still needs a rule for assigning client contributions at each adapted layer and a principled way to return the result at rank r.

A sample-size prior provides a natural client-weighting rule, but it assigns each client the same relative mass at every adapted layer. Layer-wise aggregation has also been studied in personalized and full-model federated optimization (Ma et al., 2022; Khan et al., 2026). Other federated LoRA methods adapt the sharing structure through local–global branches, task relevance, layer selection, or rank allocation (Yang et al., 2024; Seo et al., 2026; Bian et al., 2026; Peng et al., 2026; Zhang et al., 2026b). These approaches address different forms of personalization or heterogeneity, but they do not directly specify client–layer masses for the effective updates of a single shared adapter. Moreover, even if each client update has rank at most r, their weighted mean can have rank as high as the sum of the client ranks. The server must therefore first determine client–layer masses and then return a shared rank-r consensus for the resulting effective updates.

Taken together, these observations suggest a single design principle: federated LoRA aggregation should separate the choice of consensus object from the assignment of client contribution. BaryLoRA follows this principle by treating normalized effective updates as the consensus object. Each client maintains its optimization state across communication rounds and compresses its second-moment state into a positive layer summary. The server uses these summaries to adjust a sample-size prior and obtain client–layer masses. The summary serves as a signal of optimization activity rather than a calibrated estimate of client reliability; how this signal should affect the mass is an empirical design choice. Once the masses are fixed, the server forms the weighted mean of the effective updates and returns its best rank-at-most-r approximation in Frobenius norm.

The server realizes this consensus from the uploaded factor blocks through a QR-core computation, avoiding explicit construction of the dense weighted mean. The same persistent state also supports local adaptive optimization (Kingma & Ba, 2015; Reddi et al., 2021), while its summaries determine the aggregation masses used by the server. The resulting factorized computation returns the shared adapter for the next communication round. Figure 1 summarizes the progression from factor-space ambiguity to state-conditioned product consensus.

This construction separates the empirical design of the state-to-mass mapping from the conditional properties of the resulting consensus. For fixed positive state transforms and contribution masses, we establish uniqueness of the mass solution, invariance of the effective-update consensus to equivalent LoRA factorizations, and optimality of the factorized rank-constrained realization. Experiments on language understanding, code generation, and mathematical reasoning tasks evaluate BaryLoRA against federated LoRA baselines.

Our contributions are given as follows.

- We formulate BaryLoRA as a state-conditioned product-space consensus framework for federated LoRA. It assigns client–layer masses to effective updates and returns a shared adapter under a prescribed rank constraint.
- We develop an end-to-end realization in which clients maintain persistent optimization state, summarize it at each adapted layer, and use the summaries to determine client–layer masses. A factorized server computation then returns the corresponding rank-constrained product consensus.
- For fixed positive state transforms and contribution masses, we establish conditional properties of the resulting consensus, including uniqueness of the mass solution, invariance to equivalent LoRA factorizations, and optimality of the rank-constrained realization.

2 Related Work

Existing federated LoRA methods differ mainly in what the server aggregates and how it assigns client influence. We organize prior work around these two decisions.

2.1 From Factor Aggregation to Effective-Update Consensus

Factor-space methods coordinate the communicated factors through averaging, freezing, alternating optimization, selective sharing, orthogonalization, rotation, or alignment (Zhang et al., 2024; Sun

et al., 2024; Chen et al., 2025; Guo et al., 2025; Lee et al., 2025; Liu et al., 2026; Li et al., 2025b; Zhang et al., 2026a; Meng et al., 2026). They improve factor coordination, but retain a non-unique factor representation as the federation object. This leaves the represented update, rather than its coordinates, as the more stable object to compare across clients.

A second line instead aggregates the represented update before returning compact factors. FLoRA, FedEx-LoRA, FlexLoRA, and FRLoRA address exact aggregation, frozen-base correction, heterogeneous rank budgets, or cross-round residuals (Wang et al., 2024; Singhal et al., 2025; Bai et al., 2024; Yan et al., 2025). They separate the effective update from the factorization returned to clients. BaryLoRA follows this product-space view, but adds state-conditioned client–layer masses and a shared rank-r product consensus. The remaining question is therefore not only which product to aggregate, but how much each client should contribute at each adapted layer.

Rank and resource adaptation are related but orthogonal: HetLoRA, Ravan, Fed-PLoRA, FSLoRA, and RAFFT change ranks, heads, training subsets, or rank structure to match resource heterogeneity (Cho et al., 2024; Raje et al., 2025; Zhang et al., 2026b; Fang et al., 2026; Zhou et al., 2026; Liu et al., 2024; Li et al., 2025a). These methods change capacity or sharing structure, whereas BaryLoRA keeps a shared rank budget and reweights client–layer effective updates before the rank-constrained return.

2.2 Layer-Dependent Client Contributions

The second line of work varies client influence. Agnostic and fair learning operate at the model-objective level, while pFedLA and Fed-LAA assign client–layer weights to personalized or shared models (Mohri et al., 2019; Li et al., 2020b; Ma et al., 2022; Khan et al., 2026). These weights do not act on effective LoRA updates under a shared rank constraint.

Federated parameter-efficient personalization further changes what is shared through parameter subsets, local–global branches, task relevance, adapter fusion, layer grouping, or hierarchical adapters (Tamirisa et al., 2024; Xie et al., 2024; Yang et al., 2024; Seo et al., 2026; Zhao et al., 2026; Shen et al., 2025; Kou et al., 2026; Bian et al., 2026; Peng et al., 2026). These methods change sharing structure rather than define a state-conditioned mass over the effective updates of one shared adapter.

Taken together, product-space methods clarify what to aggregate, while layer-aware and personalized methods show why client influence may vary. BaryLoRA combines these perspectives: persistent optimization state tilts a sample-size prior into client–layer masses, which act on effective updates before the shared rank projection. The state summary is an empirical activity signal, and the consensus properties are conditional on fixed masses. Adaptive optimizers provide the local moments but do not define the server-side mass rule (Kingma & Ba, 2015; Reddi et al., 2021).

3 Methodology

3.1 Aggregation Interface: From LoRA Factors to Effective Updates

We study cross-device federated learning over communication rounds. At round t, the server selects $\mathcal{S}_t$ with $N_t = |\mathcal{S}_t|$. For an adapted layer l, the frozen backbone weight is $W_l^0 \in \mathbb{R}^{d_{\mathrm{out},l} \times d_{\mathrm{in},l}}$, while client i uploads factors $B_{i,l} \in \mathbb{R}^{d_{\mathrm{out},l} \times r}$ and $A_{i,l} \in \mathbb{R}^{r \times d_{\mathrm{in},l}}$. Their product induces the model change reconciled across clients:

$$W_{i,l} = W_l^0 + s_l B_{i,l} A_{i,l}, \qquad s_l > 0. \tag{1}$$

The layer scale s_l is shared by participating clients; with client-specific scales, the server aggregates the actual scaled updates.

This creates two choices: the consensus object and each client's contribution. BaryLoRA resolves them in sequence: clients upload effective updates and persistent state summaries, then the server maps summaries to client–layer masses, forms a rank-constrained product consensus, and returns factors.

The model-space change induced by the factors is

$$\Delta W_{i,l} = s_l B_{i,l} A_{i,l}, \qquad \mu_{i,l} := \frac{\Delta W_{i,l}}{s_l} = B_{i,l} A_{i,l}.$$

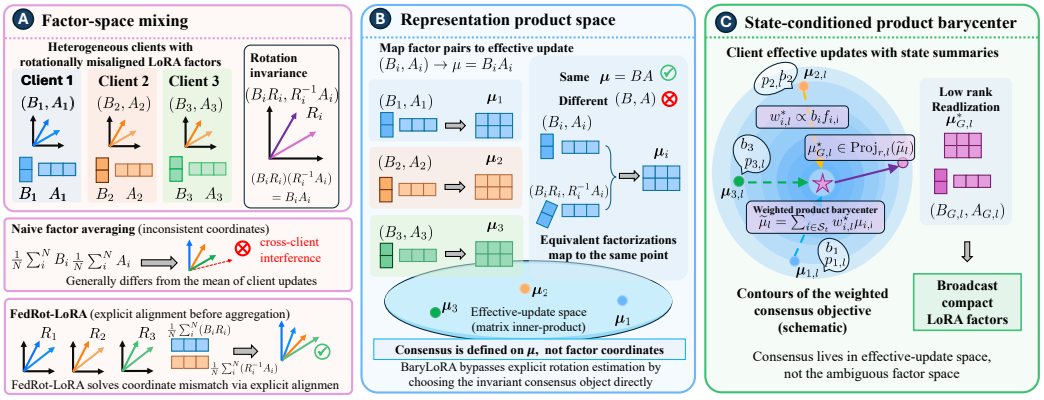


Figure 1: Overview of BaryLoRA. (A) *Factor-space mixing*: directly averaging LoRA factors depends on non-unique coordinates and can introduce cross-client interference. (B) *Representation product space*: each factor pair maps to the effective update $\mu_{i,l} = B_{i,l}A_{i,l}$, so equivalent factorizations represent the same consensus object. (C) *State-conditioned product barycenter*: persistent local state yields a client–layer summary that tilts the sample-size prior into masses $w^\star_{i,l}$; the server forms the weighted product mean, projects it to rank r, and returns compact LoRA factors.

The server aggregates model changes rather than factor coordinates, so $\mu_{i,l}$ is the product-space object used below. Its coordinates are non-unique: for every invertible $Q \in \mathbb{R}^{r\times r}$,

$$(B_{i,l}Q)(Q^{-1}A_{i,l}) = B_{i,l}A_{i,l}. \tag{2}$$

Thus factor averaging is coordinate-dependent and can introduce cross-client products absent from the submitted updates; the full expansion is in Appendix A.1. We define consensus in the Frobenius geometry of effective-update matrices.

Although every $\mu_{i,l}$ has rank at most r, a weighted mean can have rank as high as $N_t r$. The ambient consensus and the returned rank-r adapter are therefore distinct; $(B_{G,l}, A_{G,l})$ represents the latter.

Once the consensus object is fixed, client contributions must be determined from local adaptation.

3.2 Client-Side State and Local Adaptation

Task objectives. The local task loss supplies factor gradients. Classification uses local prior correction (LPC), an inverse-frequency weighted cross-entropy, whereas autoregressive generation uses causal language-model loss. Both enter the same factor-gradient interface; LPC is classification-specific. Loss forms and class weights are given in Appendix A.3.

Effective-update direction. The server aggregates products, but backpropagation exposes factor gradients. Suppress i, l and let $\mu = BA$ and $G = \nabla_\mu \mathcal{L}_i$; the chain rule gives

$$H_B := \nabla_B \mathcal{L}_i = GA^\top, \qquad H_A := \nabla_A \mathcal{L}_i = B^\top G. \tag{3}$$

These coordinate gradients are not yet the product-space direction used by the local state. With $P_B = BB^+$ and $P_A = A^+A$, the compatible minimum-Frobenius-norm direction is

$$G_\mu = P_B G + GP_A - P_B G P_A. \tag{4}$$

This direction feeds the persistent moments and is invariant under the corresponding invertible fixed-rank refactorizations in Eq. (2). The implementation uses damped recovery from H_A, H_B; approximation error and finite-damping dependence are given in Appendix A.3.

Persistent state and local update. To persist this signal across local steps and rounds, let k count cumulative steps at the layer, set $g_k = -G_{\mu,k}$, and maintain

$$m_k = \beta_1 m_{k-1} + (1-\beta_1)g_k,$$
$$v_k = \beta_2 v_{k-1} + (1-\beta_2)g_k \odot g_k.$$

Here $\odot$ is element-wise multiplication. Bias correction gives $\widehat{m}_k = m_k/(1-\beta_1^k)$ and $\widehat{v}_k = v_k/(1-\beta_2^k)$. The same persistent moments guide the local step and provide the later state summary.

For the optional zero-centered prior, define $\Omega_k = (\widehat{v}_k + \epsilon_{\text{var}})/s_{\text{var}}$ and $u_k^{\text{prior}} = -\lambda_{\text{prior}}\zeta_i\omega_k(\Omega_k \odot \mu_k)$, where $\omega_k \in [0,1]$ and $\zeta_i \geq 0$. The dense local product update is

$$\mu_{k+1}^{+} = \mu_k + \eta s_{\text{upd}} \frac{\widehat{m}_k + u_k^{\text{prior}}}{\sqrt{\widehat{v}_k} + \epsilon}. \tag{5}$$

Here $s_{\text{upd}} > 0$ is distinct from the LoRA layer scale s_l. The prior affects the local update only, not the server mass or consensus.

The dense step may leave the rank-r manifold, so the client projects back before the next step and upload:

$$\mu_{k+1} = \pi_{r,l}(\mu_{k+1}^{+}) \in \text{Proj}_{r,l}(\mu_{k+1}^{+}). \tag{6}$$

The projected matrix is refactorized for local training and communication. Truncated SVD gives the exact operator; randomized rank return is analyzed in Appendix A.4.

State summary and upload. After the local update, the client compresses the corrected second moment $\widehat{v}_{i,l}$ into a positive activity signal:

$$p_{i,l} = \left[\text{mean}\left(\frac{\widehat{v}_{i,l} + \epsilon_{\text{var}}}{s_{\text{var}}}\right)\right]^{\gamma} > 0, \qquad \gamma, s_{\text{var}}, \epsilon_{\text{var}} > 0. \tag{7}$$

It is not a calibrated variance, precision, Fisher information, or reliability estimate. Entrywise clipping keeps the upload finite (Appendix A.3), and mapping larger activity to larger server mass is an empirical choice. Each client uploads final factors and $p_{i,l}$, while dense moments persist locally across rounds.

3.3 From State Summaries to Client–Layer Masses and Product Consensus

Once clients have uploaded their factors and summaries, the server turns the summaries into a layer-wise reweighting of a sample-size prior. Let $n_i > 0$ be the examples processed by client i in the round:

$$b_i = \frac{n_i}{\sum_{j \in \mathcal{S}_t} n_j}, \qquad \mathbf{b} = (b_i)_{i \in \mathcal{S}_t} \in \Delta^{N_t - 1}.$$

This prior encodes data volume only, not learned client usefulness.

Because p is comparable only within a layer, normalize it by the participating-client mean and clip the ratio:

$$\chi_{i,l} = \text{clip}\left(\frac{p_{i,l}}{\overline{p}_l}, \frac{1}{D}, D\right), \qquad D \geq 1. \tag{8}$$

This yields a bounded, dimensionless state. Convert it into a positive correction with tunable strength:

$$f_{i,l} = (1 - \rho_l) + \rho_l \chi_{i,l}, \qquad 0 \leq \rho_l \leq 1. \tag{9}$$

The coefficient ρ_l controls the tilt and $f_{i,l} > 0$ under the stated bounds. Equal states or $\rho_l = 0$ recover the prior; invalid summaries use the fallback in Appendix A.4.

Combining the prior and correction gives the mass consumed by consensus:

$$w_{i,l}^{\star} = \frac{b_i f_{i,l}}{Z_l}, \qquad Z_l = \sum_{j \in \mathcal{S}_t} b_j f_{j,l}. \tag{10}$$

The state affects aggregation only through $w_{i,l}^{\star}$; mapping larger activity to larger mass is an empirical choice, not a reliability theorem.

For fixed positive corrections, the same rule is the unique KL-regularized deviation from the sample-size prior:

$$\min_{\mathbf{w} \in \Delta^{N_t - 1}} \left\{ \text{KL}(\mathbf{w} \| \mathbf{b}) - \sum_i w_i \log f_{i,l} \right\}. \tag{11}$$

Thus normalization and uniqueness are conditional on $f_{i,l}$, not evidence that activity is calibrated utility or reliability.

With the masses fixed, aggregate the effective updates and enforce rank r:

$$\mu^\star_{G,l} \in \mathrm{Proj}_{r,l}\left(\sum_{i\in\mathcal{S}_t} w^\star_{i,l}\mu_{i,l}\right), \tag{12}$$

Here $\mathrm{Proj}_{r,l}(X)$ is the closest set $\mathcal{M}_{r,l} = \{M \in \mathbb{R}^{d_{\mathrm{out},l}\times d_{\mathrm{in},l}} : \mathrm{rank}(M) \le r\}$ in Frobenius norm. The inner sum is the unconstrained product mean; projection makes the returned adapter feasible. Mass assignment and product consensus are sequential.

Proposition 1 (State-conditioned product consensus)**.** *Suppose $b_i > 0$ and $f_{i,l} > 0$. Then Eq. (10) uniquely solves Eq. (11). With $\widetilde{\mu}_l = \sum_i w^\star_{i,l}\mu_{i,l}$, the fixed-mass rank-$r$ barycenter is*

$$\arg\min_{\mathrm{rank}(M)\le r} \sum_i w^\star_{i,l}\|M - \mu_{i,l}\|_F^2 = \mathrm{Proj}_{r,l}(\widetilde{\mu}_l). \tag{13}$$

The solution set is invariant to product-preserving refactorizations. It is a singleton when $\mathrm{rank}(\widetilde{\mu}_l) \le r$ or when $\sigma_r(\widetilde{\mu}_l) > \sigma_{r+1}(\widetilde{\mu}_l)$.

At a positive singular-value tie, the optimal set remains defined but a particular return needs a tie rule. Proof details are in Appendix A.1. The corresponding tensor-level implementation audit is reported in Appendix A.2.

3.4 Factorized Server Realization and One-Round Procedure

QR-core rank return. The target matrix need not be materialized densely. For each adapted layer, form weighted factor blocks by concatenating $\sqrt{w^\star_{i,l}}B_{i,l}$ horizontally and stacking $\sqrt{w^\star_{i,l}}A_{i,l}$ vertically. Their product is the ambient mean:

$$\widetilde{\mu}_l = L_l R_l, \qquad L_l = Q_L T_L, \qquad R_l^\top = Q_R T_R,$$
$$\widetilde{\mu}_l = Q_L C_l Q_R^\top, \qquad C_l = T_L T_R^\top.$$

If $C_l = U_C \Sigma_C V_C^\top$, the core SVD returns

$$B_{G,l} = Q_L U_{C,r}\Sigma_{C,r}^{1/2}, \qquad A_{G,l} = \Sigma_{C,r}^{1/2} V_{C,r}^\top Q_R^\top. \tag{14}$$

The returned factors realize the same rank-r target, rather than defining a new aggregation objective.

Proposition 2 (Exact QR-core realization)**.** *For fixed positive masses, Eq. (14) belongs to $\mathrm{Proj}_{r,l}(\widetilde{\mu}_l)$. The core and ambient mean have equal optimal Frobenius residuals, with corresponding solution sets under $Z \mapsto Q_L Z Q_R^\top$.*

QR-core changes computation, not the consensus problem. A boundary tie may yield different matrices in the same optimal set; proofs and cost details are in Appendix A.5.

Optional task-head inheritance. With a separate trainable head, reuse the constructed layer masses:

$$h_i = \sum_{l=1}^{L} \alpha_l w^\star_{i,l}, \qquad \theta_G^{\mathrm{head}} = \sum_{i\in\mathcal{S}_t} h_i \theta_i^{\mathrm{head}}, \quad \boldsymbol{\alpha} \in \Delta^{L-1}. \tag{15}$$

We use $\alpha_l = 1/L$. This optional step does not alter LoRA consensus; generation tasks without a separate head omit it (Appendix A.5).

Algorithm 1 summarizes one communication round in the notation above. It separates the two coupled decisions: clients adapt in the product space and upload factors with one state summary per adapted layer, after which the server builds the client–layer masses and returns the rank-constrained consensus. Here QRCore_r is the factorized rank return of Eq. (14); the dense moments persist across rounds, and the complete finite-threshold protocol, including validation and fallback rules, is given in Appendix A1.

Algorithm 1 BaryLoRA: one communication round

Require: global factors $(B_{G,l}, A_{G,l})_l$, clients $\mathcal{S}_t$, persistent moments $(m_{i,l}, v_{i,l})$, rank r

for each client $i \in \mathcal{S}_t$ **in parallel do**
  $(B_{i,l}, A_{i,l}) \leftarrow (B_{G,l}, A_{G,l})$ for all l
  $g_k \leftarrow -\widetilde{G}_{\mu,k}$; $m_k \leftarrow \beta_1 m_{k-1} + (1-\beta_1) g_k$; $v_k \leftarrow \beta_2 v_{k-1} + (1-\beta_2) g_k \odot g_k$
  $\mu_{k+1} \leftarrow \pi_{r,l}\big(\mu_k + \eta s_{\mathrm{upd}}(\widehat{m}_k + u_k^{\mathrm{prior}})/(\sqrt{\widehat{v}_k} + \epsilon)\big)$; refactor $\mu_{k+1} = B_{i,l} A_{i,l}$
  $p_{i,l} \leftarrow [\mathrm{mean}((\widehat{v}_{i,l} + \epsilon_{\mathrm{var}})/s_{\mathrm{var}})]^{\gamma}$; upload $(B_{i,l}, A_{i,l}, p_{i,l})_l$
end for
for each adapted layer l **do**
  $\chi_{i,l} \leftarrow \mathrm{clip}(p_{i,l}/\overline{p}_l, 1/D, D)$; $f_{i,l} \leftarrow (1-\rho_l) + \rho_l \chi_{i,l}$
  $w^{\star}_{i,l} \leftarrow b_i f_{i,l} / \sum_j b_j f_{j,l}$
  $L_l \leftarrow \mathrm{hcat}_i(\sqrt{w^{\star}_{i,l}} B_{i,l})$; $R_l \leftarrow \mathrm{vcat}_i(\sqrt{w^{\star}_{i,l}} A_{i,l})$
  $(B_{G,l}, A_{G,l}) \leftarrow \mathrm{QRCore}_r(L_l, R_l)$
end for

4 Experiments

4.1 Experimental setup

For natural language understanding, we adapt RoBERTa-large (Liu et al., 2019) with LoRA (Hu et al., 2022). For generative tasks, we use Llama-3-8B (Grattafiori et al., 2024) and include the Qwen3.5-9B (Qwen Team, 2026) as a second backbone in Table 2. We evaluate classification on five GLUE tasks (Wang et al., 2018) (SST-2, QNLI, QQP, RTE, and MNLI), code synthesis on HumanEval (Chen et al., 2021) with CodeSearchNet (Husain et al., 2019) fine-tuning, and mathematical reasoning on GSM8K (Cobbe et al., 2021). We report validation accuracy for GLUE, pass@1 for HumanEval, and exact-match accuracy for GSM8K. The main comparison includes FedIT (Zhang et al., 2024), FlexLoRA (Bai et al., 2024), FFA-LoRA (Sun et al., 2024), RoLoRA (Chen et al., 2025), and FedRot-LoRA (Zhang et al., 2026a). The generative comparison also includes ShareAB, our local implementation of factor-wise sharing. The aggregation rules are summarized in Appendix B.1. The main GLUE comparison uses $N \in \{3, 10, 50\}$ clients, rank $r = 4$, and Dirichlet concentration $h = 0.5$, with additional rank and concentration sweeps on MNLI, QNLI, and QQP. The $N = 3$ client-scale runs use exact local rank return, whereas the $N = 10$ and $N = 50$ runs use randomized low-rank return. Thus, the client-count comparison also changes the rank-return backend. Each baseline retains its defined optimization rule, while the convergence curves use a matched AdamW protocol. All methods use the same task metrics and evaluation scripts. The tables use h for the Dirichlet concentration; the robustness plots label the same parameter as α. Dataset-level settings and baseline details are given in Appendix B.

4.2 Main results

Client-scale comparison. Changing the client count changes the set of client updates included in the server consensus. Table 1 evaluates the full method with rank 4 and $h = 0.5$ for $N \in \{3, 10, 50\}$. BaryLoRA has the highest mean five-task average in each row: 0.916, 0.907, and 0.905, with margins of 0.023, 0.025, and 0.032 over the strongest baseline. At $N = 3$, the largest task-level gaps are on QNLI, QQP, and RTE. The average margin grows with the displayed client counts, but this pattern cannot be attributed to client count alone because the rank-return backend also changes. A rank sweep on MNLI is reported in Appendix B.2.

Generative transfer. Table 2 extends the comparison beyond classification. On Llama-3-8B, BaryLoRA reaches 0.440 on HumanEval and 0.499 on GSM8K, with margins of 0.027 and 0.034 over the strongest baseline. On Qwen3.5-9B, it has the highest mean in both columns, with margins of 0.030 on HumanEval and 0.004 on GSM8K. The ranking therefore extends to both code generation and mathematical reasoning in the displayed settings, but the size of the advantage depends on the backbone. In particular, the Qwen GSM8K result is close to the RoLoRA baseline and should not be described as a large reasoning gain.

Table 1: Mean GLUE validation accuracy across client scales. Task entries are reported as mean ± standard deviation; Avg. is the mean across tasks.

Clients	Method	SST-2	QNLI	QQP	RTE	MNLI	Avg.
3	FedIT	0.953 ± 0.001	0.926 ± 0.002	0.771 ± 0.098	0.840 ± 0.013	0.866 ± 0.001	0.871
	FlexLoRA	0.949 ± 0.003	0.910 ± 0.007	0.832 ± 0.008	0.813 ± 0.015	0.845 ± 0.008	0.870
	FFA-LoRA	0.768 ± 0.056	0.926 ± 0.002	0.838 ± 0.002	0.830 ± 0.005	0.862 ± 0.001	0.845
	RoLoRA	0.951 ± 0.004	0.917 ± 0.011	0.841 ± 0.012	0.854 ± 0.005	0.868 ± 0.005	0.886
	FedRot-LoRA	0.954 ± 0.001	0.926 ± 0.002	0.842 ± 0.002	0.868 ± 0.014	0.876 ± 0.002	0.893
	BaryLoRA (ours)	$\mathbf{0.963 \pm 0.001}$	$\mathbf{0.948 \pm 0.007}$	$\mathbf{0.887 \pm 0.017}$	$\mathbf{0.892 \pm 0.009}$	$\mathbf{0.889 \pm 0.012}$	**0.916**
10	FedIT	0.949 ± 0.002	0.913 ± 0.004	0.849 ± 0.000	0.621 ± 0.054	0.849 ± 0.002	0.836
	FlexLoRA	0.922 ± 0.009	0.876 ± 0.018	0.801 ± 0.032	0.592 ± 0.030	0.792 ± 0.011	0.797
	FFA-LoRA	0.946 ± 0.001	0.894 ± 0.008	0.849 ± 0.002	0.600 ± 0.024	0.840 ± 0.002	0.826
	RoLoRA	0.956 ± 0.001	0.911 ± 0.015	0.863 ± 0.001	0.805 ± 0.007	0.858 ± 0.005	0.879
	FedRot-LoRA	0.958 ± 0.000	0.921 ± 0.010	0.868 ± 0.001	0.798 ± 0.005	0.864 ± 0.001	0.882
	BaryLoRA (ours)	$\mathbf{0.961 \pm 0.019}$	$\mathbf{0.942 \pm 0.008}$	$\mathbf{0.890 \pm 0.017}$	$\mathbf{0.859 \pm 0.007}$	$\mathbf{0.885 \pm 0.004}$	**0.907**
50	FedIT	0.928 ± 0.023	0.803 ± 0.005	0.814 ± 0.022	0.622 ± 0.051	0.673 ± 0.023	0.768
	FlexLoRA	0.752 ± 0.135	0.733 ± 0.121	0.739 ± 0.066	0.632 ± 0.059	0.677 ± 0.032	0.707
	FFA-LoRA	0.933 ± 0.015	0.851 ± 0.004	0.716 ± 0.018	0.658 ± 0.038	0.701 ± 0.008	0.772
	RoLoRA	0.851 ± 0.176	0.893 ± 0.042	0.847 ± 0.003	0.733 ± 0.064	0.798 ± 0.114	0.824
	FedRot-LoRA	0.947 ± 0.004	0.920 ± 0.012	0.854 ± 0.002	0.787 ± 0.004	0.859 ± 0.003	0.873
	BaryLoRA (ours)	$\mathbf{0.962 \pm 0.005}$	$\mathbf{0.944 \pm 0.002}$	$\mathbf{0.881 \pm 0.001}$	$\mathbf{0.838 \pm 0.012}$	$\mathbf{0.899 \pm 0.024}$	**0.905**

Table 2: HumanEval pass@1 and GSM8K exact-match accuracy on two backbones. Entries are mean ± standard deviation.

Method	HumanEval (Llama)	GSM8K (Llama)	HumanEval (Qwen)	GSM8K (Qwen)
BaryLoRA (ours)	$\mathbf{0.440 \pm 0.012}$	$\mathbf{0.499 \pm 0.016}$	$\mathbf{0.667 \pm 0.022}$	$\mathbf{0.892 \pm 0.005}$
FedRot-LoRA	0.404 ± 0.025	0.464 ± 0.007	0.620 ± 0.010	0.814 ± 0.013
FFA-LoRA	0.379 ± 0.019	0.465 ± 0.047	0.637 ± 0.030	0.826 ± 0.010
FlexLoRA	0.409 ± 0.025	0.418 ± 0.002	0.616 ± 0.031	0.761 ± 0.042
RoLoRA	0.371 ± 0.032	0.458 ± 0.026	0.617 ± 0.008	0.888 ± 0.028
ShareAB	0.413 ± 0.028	0.437 ± 0.040	0.621 ± 0.019	0.804 ± 0.024

4.3 Robustness and convergence

Changing the Dirichlet concentration changes the label distribution seen by each client while leaving the aggregation pipeline fixed. Figure 2 shows BaryLoRA with the highest mean accuracy in all six plotted QNLI and QQP settings. Relative to the strongest displayed baseline in each panel, the smallest and largest gaps are approximately 4.4 and 14.9 percentage points. These curves test the full method across the evaluated heterogeneity levels; they do not isolate the contribution of state-conditioned client-layer masses. The complementary MNLI sweep over three heterogeneity levels is reported in Table 3, where BaryLoRA again has the highest mean accuracy at each level.

To examine temporal behavior separately, Figure 2 tracks validation accuracy and loss over 200 rounds under the matched protocol in Appendix B. After the initial transient, BaryLoRA reaches the highest accuracy in both tasks and maintains that lead in the plotted window. At the reported round-20 checkpoint, its QNLI and QQP accuracies are 0.924 and 0.838, versus displayed-baseline ranges of 0.854 to 0.903 and 0.677 to 0.769. The corresponding losses are 0.206 and 0.350, versus 0.258 to 0.360 and 0.446 to 0.510. These curves measure the full method.

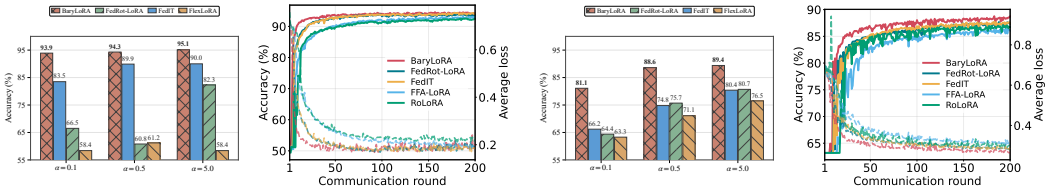
4.4 Ablation study

The ablations replace the consensus object, the client-layer mass rule, the persistent-state local update, or the local prior. Table 4 reports BaryLoRA and five component controls. FactorAvg replaces product-space consensus with factor-space averaging. BaseMass replaces state-conditioned masses with sample-size masses. FactorBase applies both changes. MonitorOnly disables the state-conditioned local update, and NoPrior sets the prior coefficient to zero. Each control cell reports validation accuracy and its difference from BaryLoRA in parentheses.

All five controls are below BaryLoRA on the four listed tasks. This pattern is consistent with a contribution from each tested ingredient, but it does not provide an additive causal ranking because the controls change different components and their effects can interact. QNLI illustrates this inter-

Table 3: Mean MNLI accuracy across data-heterogeneity levels. Entries are mean ± standard deviation.

h	FedIT	FlexLoRA	FFA-LoRA	RoLoRA	FedRot-LoRA	BaryLoRA (ours)
100	0.868 ± 0.003	0.871 ± 0.001	0.869 ± 0.000	0.767 ± 0.158	0.884 ± 0.001	$\mathbf{0.889 \pm 0.003}$
1	0.864 ± 0.015	0.839 ± 0.017	0.856 ± 0.003	0.780 ± 0.140	0.873 ± 0.001	$\mathbf{0.894 \pm 0.017}$
0.5	0.866 ± 0.001	0.845 ± 0.008	0.862 ± 0.001	0.868 ± 0.005	0.876 ± 0.002	$\mathbf{0.881 \pm 0.011}$



(a) QNLI robustness (b) QNLI convergence (c) QQP robustness (d) QQP convergence

Figure 2: Robustness to the Dirichlet concentration (a, c) and convergence over communication rounds (b, d). The robustness panels report validation accuracy; the convergence panels show validation accuracy and average loss.

Table 4: Mechanism controls for BaryLoRA on four GLUE tasks. Each control cell reports validation accuracy with the drop Δ relative to BaryLoRA in parentheses.

Dataset	Ours	FactorAvg	BaseMass	FactorBase	MonitorOnly	NoPrior
SST-2	**0.963**	0.945 (Δ 0.018)	0.941 (Δ 0.022)	0.935 (Δ 0.028)	0.934 (Δ 0.029)	0.927 (Δ 0.036)
QNLI	**0.948**	0.942 (Δ 0.006)	0.923 (Δ 0.025)	0.932 (Δ 0.016)	0.936 (Δ 0.012)	0.509 (Δ 0.439)
QQP	**0.887**	0.864 (Δ 0.023)	0.865 (Δ 0.022)	0.862 (Δ 0.025)	0.849 (Δ 0.038)	0.844 (Δ 0.043)
RTE	**0.892**	0.866 (Δ 0.026)	0.877 (Δ 0.015)	0.861 (Δ 0.031)	0.857 (Δ 0.035)	0.870 (Δ 0.022)

action: BaseMass (0.923) is lower than FactorBase (0.932), even though FactorBase also changes the consensus object. The QNLI NoPrior value of 0.509 is close to the majority-class baseline and is better read as a collapse diagnostic than as an ordinary accuracy drop.

As a direct check on the consensus object, Table 5 applies product-preserving reparameterizations $B_i' = B_i Q_i$, $A_i' = Q_i^{-1} A_i$ to recorded factor pairs. Under all three transform families, the product-space consensus changes by at most 1.23×10^{-6} in the displayed worst case, which is consistent with float32 noise. Separate factor averaging changes by median values of 0.78 and 1.04 under the orthogonal and well-conditioned transforms; the shared transform is an exception with a median change of 1.00×10^{-7}. This audit gives numerical evidence that the effective-update consensus is stable under the tested reparameterizations, whereas factor averaging is not generally invariant.

Table 5: Factorization-invariance check: median / worst relative change over the eight task-layer groups.

Transform	Product	Factor
Orthogonal	$6.28\times10^{-7}/1.18\times10^{-6}$	$7.81\times10^{-1}/1.79\times10^{1}$
Well-conditioned	$6.61\times10^{-7}/1.23\times10^{-6}$	$1.04\times10^{0}/2.68\times10^{1}$
Shared	$6.48\times10^{-7}/1.06\times10^{-6}$	$1.00\times10^{-7}/2.59\times10^{-7}$

5 Conclusion

In this work, we propose a novel BaryLoRA framework that derives the masses by tilting a sample-size prior with persistent optimization state, then returns a rank-constrained product consensus through a QR-core realization. For fixed positive state transforms, the mass optimization has a unique solution. For fixed masses, the effective-update consensus is invariant to equivalent LoRA factorizations, and the QR-core computation returns a Frobenius-optimal rank-constrained update. In the experiments, BaryLoRA has the highest mean in the displayed GLUE client-scale and heterogeneity comparisons and in both generative task columns for each backbone. In particular, the state-to-mass mapping remains an empirical design choice, and matched multi-seed controls are needed to quantify its contribution.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for the following tasks that require disclosure: implementing and debugging the method and its numerical realizations, and interpreting experimental results. We have not used generative AI tools to generate synthetic datasets, to develop theoretical models or conceptual frameworks, to formulate mathematical claims, to provide critical ingredients for or assist in the writing of proofs, to propose or refine hypotheses, to design research methodology or experiments, to assist with translation, to clean or reformat datasets, or to support qualitative and thematic data analysis, and these tasks are not applicable to this work.

Additionally, we used generative AI tools for the following tasks that carry recommended disclosure: creating or modifying scientific figures and images, drafting and editing parts of the manuscript, identifying relevant literature, summarizing or analysing existing literature, searching for information, and brainstorming.

We have reviewed all AI-assisted work. AI-assisted code was executed, tested, and verified for correctness by the authors against the reported results; all mathematical claims and proofs were derived and checked by the authors without AI assistance; AI-assisted literature summaries were verified against the original papers; and AI-generated or AI-modified figures were inspected and corrected by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

All datasets used in this work (GLUE, HumanEval, GSM8K, and CodeSearchNet) are publicly available and were used under their original licenses. The pretrained backbones (RoBERTa-large, Llama-3-8B, and Qwen3.5-9B) are publicly released checkpoints. The study does not involve human subjects, personal data, or the collection of new data, so no IRB approval was required. The work concerns federated fine-tuning, whose purpose is to keep client data local; it does not include the release of private user data or of any model trained on non-public data. We do not foresee specific harmful applications of the method, and we do not release artifacts that we believe pose a concrete misuse risk. We are not aware of any conflicts of interest that influenced the design or reporting of this work.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

Anonymized source code that reproduces the method and all reported results is available in the anonymous repository:

`https://anonymous.4open.science/r/barylora-C18F/`

REFERENCES

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 14457–14483, 2024. doi: 10.52202/079017-0461.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. FedTreeLoRA: Reconciling statistical and functional heterogeneity in federated LoRA fine-tuning, 2026.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Shuangyi Chen, Yuanxin Guo, Yue Ju, Hardik Dalal, Zhongwen Zhu, and Ashish Khisti. Robust federated finetuning of LLMs via alternating optimization of LoRA. In *Advances in Neural Information Processing Systems*, volume 38, pp. 120038–120090, 2025. doi: 10.52202/085713-4009.

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*, 2024. doi: 10.18653/v1/2024.emnlp-main.717.

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

Wenzhi Fang, Dong-Jun Han, Liangqi Yuan, Seyyedali Hosseinalipour, and Christopher G. Brinton. Federated sketching LoRA: A flexible framework for heterogeneous collaborative fine-tuning of LLMs. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning. In *International Conference on Learning Representations*, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. CodeSearchNet challenge: Evaluating the state of semantic code search, 2019. URL https://arxiv.org/abs/1909.09436.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 2020.

Anam Nawaz Khan, Atif Rizwan, Rashid Ahmad, Qazi Waqas Khan, Xueping Li, and DoHyeun Kim. Accelerating model convergence in federated learning with layer-wise adaptive weight aggregation. *Applied Soft Computing*, 201:115676, 2026. doi: 10.1016/j.asoc.2026.115676.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Mengsha Kou, Xiaoyu Xia, Ziqi Wang, Ibrahim Khalil, Runkun Luo, Jingwen Zhou, and Minhui Xue. WinFLoRA: Incentivizing client-adaptive aggregation in federated LoRA under privacy heterogeneity. In *Proceedings of the ACM Web Conference 2026*, pp. 5241–5252, 2026. doi: 10.1145/3774904.3792295.

Seanie Lee, Sangwoo Park, Dong Bok Lee, Dominik Wagner, Haebin Seong, Tobias Bocklet, Juho Lee, and Sung Ju Hwang. FedSVD: Adaptive orthogonalization for private federated learning with LoRA. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-3998.

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

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020a.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020b.

Zhixuan Li, Binqian Xu, Xiangbo Shu, Jiachao Zhang, Yazhou Yao, Guo-Sen Xie, and Jinhui Tang. Tensor-aggregated LoRA in federated fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1058–1067, 2025b.

Jiahao Liu, Yipeng Zhou, Di Wu, Miao Hu, Mohsen Guizani, and Quan Z. Sheng. FedLMT: Tackling system heterogeneity of federated learning via low-rank model training with theoretical guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Jin Liu, Yinbin Miao, Ning Xi, and Junkang Liu. Rethinking LoRA for privacy-preserving federated learning in large models. In *International Conference on Learning Representations*, 2026.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019. URL `https://arxiv.org/abs/1907.11692`.

Xiaosong Ma, Jie Zhang, Song Guo, and Wenchao Xu. Layer-wised model aggregation for personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10092–10101, 2022.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017.

Chuiyang Meng, Ming Tang, and Vincent W. S. Wong. FLoRG: Federated fine-tuning with low-rank gram matrices and procrustes alignment. In *International Conference on Learning Representations*, volume 2026, pp. 82804–82824, 2026.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4615–4625. PMLR, 2019.

Zihao Peng, Nan Zou, Jiandian Zeng, Guo Li, Ke Chen, Boyuan Li, and Tian Wang. HiLoRA: Hierarchical low-rank adaptation for personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31746–31757, 2026.

Qwen Team. Qwen3.5-9B model card, 2026. URL `https://huggingface.co/Qwen/Qwen3.5-9B`.

Arian Raje, Baris Askin, Divyansh Jhunjhunwala, and Gauri Joshi. Ravan: Multi-head low-rank adaptation for federated fine-tuning. In *Advances in Neural Information Processing Systems*, 2025.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

Minhyuk Seo, Taeheon Kim, Hankook Lee, Jonghyun Choi, and Tinne Tuytelaars. Co-LoRA: Collaborative model personalization on heterogeneous multi-modal clients. In *International Conference on Learning Representations*, 2026.

Zhanming Shen, Tianqi Xu, Hao Wang, Jian Li, and Miao Pan. pFedGPT: Hierarchically optimizing LoRA aggregation weights for personalized federated GPT models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4766–4778. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.239.

Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. FedEx-LoRA: Exact aggregation for federated and efficient fine-tuning of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1316–1336. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.67.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving LoRA in privacy-preserving federated learning. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Rishub Tamirisa, Chulin Xie, Wenxuan Bao, Andy Zhou, Ron Arel, and Aviv Shamsian. FedSelect: Personalized federated learning with customized selection of parameters for fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. 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*, pp. 353–355. Association for Computational Linguistics, 2018. doi: 10.18653/v1/W18-5446.

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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 22513–22533, 2024. doi: 10.52202/079017-0708.

Chulin Xie, De-An Huang, Wenda Chu, Daguang Xu, Chaowei Xiao, Bo Li, and Anima Anandkumar. PerAda: Parameter-efficient federated learning personalization with generalization guarantees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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 *International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=1lrrU7z8Cq`.

Yiyuan Yang, Guodong Long, Tao Shen, Jing Jiang, and Michael Blumenstein. Dual-personalizing adapter for federated foundation models. In *Advances in Neural Information Processing Systems*, 2024.

Haoran Zhang, Dongjun Kim, Seohyeon Cha, and Haris Vikalo. FedRot-LoRA: Mitigating rotational misalignment in federated LoRA. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026a.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Yufan Zhou, Guoyin Wang, and Yiran Chen. Towards building the federated GPT: Federated instruction tuning. In *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6915–6919. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10447454.

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

Ruonan Zhao, Zheng Wang, Debin Liu, Shijie Lv, and Laurence Tianruo Yang. FedARA: Resource-adaptive low-rank personalized federated learning via anchor-driven representation alignment on heterogeneous edge devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Zihan Zhou, Yang Zhou, Tianshi Che, Zeru Zhang, Jiaxiang Ren, Da Yan, Zhe Jiang, Yelong Shen, Ruoming Jin, and Jianfeng Gao. Noise-free dynamic rank-adaptation via riemannian methods in federated fine-tuning. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 478–486. PMLR, 2026.

A Theoretical and Numerical Analysis

This appendix establishes the properties used by the effective-update consensus and relates its exact definitions to the numerical operators used by the client and server. The analysis follows the order of the main method: consensus over effective updates, client-side direction recovery, state-conditioned mass, and factorized rank-constrained aggregation.

A.1 Properties of Effective-Update Consensus

Scale and representation. Under the shared layer scale in Eq. (1), normalizing $\Delta W_{i,l} = s_l B_{i,l} A_{i,l}$ by s_l preserves all relative client contributions within layer l. If the scales are client-specific, the consensus object must instead be the scaled update $\Delta W_{i,l} = s_{i,l} B_{i,l} A_{i,l}$. The following analysis then applies after replacing each $\mu_{i,l}$ with $\Delta W_{i,l}$.

The effective update is invariant to an invertible change of factor coordinates. For $Q_i \in \mathbb{R}^{r\times r}$ invertible,

$$(B_{i,l}Q_i)(Q_i^{-1}A_{i,l}) = B_{i,l}A_{i,l} = \mu_{i,l}.$$

Consequently, every expression constructed only from $\{\mu_{i,l}\}_i$ is unchanged by independent coordinate transformations of the clients. This covers the weighted mean, the rank-constrained barycenter objective, and its solution set.

Separate factor averaging produces terms that are absent from the submitted updates. For any simplex weights $\{w_i\}$,

$$\left(\sum_i w_i B_{i,l}\right)\left(\sum_i w_i A_{i,l}\right) = \sum_i w_i^2 B_{i,l}A_{i,l} + \sum_{i\neq j} w_i w_j B_{i,l}A_{j,l}. \tag{16}$$

The cross-client products can change when a single client refactorizes its submitted update.

Unique state-conditioned mass. Fix a layer and assume $b_i > 0$ and $f_{i,l} > 0$ for all participating clients. Write the objective in Eq. (11) as

$$\Phi_l(\mathbf{w}) = \mathrm{KL}(\mathbf{w}\|\mathbf{b}) - \sum_i w_i \log f_{i,l}, \qquad \mathbf{w}\in\Delta^{N_t-1},$$

where $\Delta^{N_t-1} = \{\mathbf{w} : w_i \geq 0, \sum_i w_i = 1\}$. We use the standard convention $0\log 0 = 0$. Since every b_i is positive, $\mathrm{KL}(\mathbf{w}\|\mathbf{b})$ is finite throughout the simplex. Define

$$Z_l = \sum_j b_j f_{j,l}, \qquad \overline{w}_{i,l} = \frac{b_i f_{i,l}}{Z_l}.$$

Positivity of b_i and $f_{i,l}$ gives $Z_l > 0$ and $\overline{w}_{i,l} > 0$. Moreover,

$$\sum_i \overline{w}_{i,l} = \frac{\sum_i b_i f_{i,l}}{Z_l} = 1,$$

so $\overline{\mathbf{w}}_l$ lies in the interior of the simplex.

We first rewrite the objective globally, including its boundary. For every $\mathbf{w}\in\Delta^{N_t-1}$,

$$\begin{aligned}
\Phi_l(\mathbf{w}) &= \sum_i w_i \log\frac{w_i}{b_i} - \sum_i w_i \log f_{i,l} \\
&= \sum_i w_i \log\frac{w_i}{b_i f_{i,l}} \\
&= \sum_i w_i \log\frac{w_i}{Z_l \overline{w}_{i,l}} \\
&= \sum_i w_i \log\frac{w_i}{\overline{w}_{i,l}} - \left(\sum_i w_i\right)\log Z_l \\
&= \mathrm{KL}(\mathbf{w}\|\overline{\mathbf{w}}_l) - \log Z_l.
\end{aligned}$$

The fourth line uses the simplex constraint $\sum_i w_i = 1$. To see the nonnegativity explicitly, let $S = \{i : w_i > 0\}$. Applying $-\log x \geq 1 - x$ with $x = \overline{w}_{i,l}/w_i$ gives

$$\begin{aligned}\mathrm{KL}(\mathbf{w}\|\overline{\mathbf{w}}_l) &= \sum_{i\in S} w_i \left[-\log \frac{\overline{w}_{i,l}}{w_i}\right] \\ &\geq \sum_{i\in S}(w_i - \overline{w}_{i,l}) = 1 - \sum_{i\in S}\overline{w}_{i,l} \geq 0.\end{aligned}$$

Equality in both inequalities requires $S = \{1, \ldots, N_t\}$ and $w_i = \overline{w}_{i,l}$ for every i. Therefore

$$\Phi_l(\mathbf{w}) \geq -\log Z_l = \Phi_l(\overline{\mathbf{w}}_l),$$

with equality only at $\mathbf{w} = \overline{\mathbf{w}}_l$. This already proves global optimality and uniqueness over the closed simplex, including all boundary points.

The closed form can also be recovered from the stationarity condition. The Lagrangian for an interior solution is

$$\mathcal{J}(\mathbf{w}, \lambda) = \sum_i w_i \log \frac{w_i}{b_i f_{i,l}} + \lambda\left(\sum_i w_i - 1\right).$$

Differentiating with respect to w_i gives

$$\frac{\partial \mathcal{J}}{\partial w_i} = \log \frac{w_i}{b_i f_{i,l}} + 1 + \lambda = 0,$$

and hence

$$w_i = e^{-1-\lambda} b_i f_{i,l}.$$

Enforcing $\sum_i w_i = 1$ yields $e^{-1-\lambda} = 1/Z_l$ and therefore

$$w_i = \frac{b_i f_{i,l}}{Z_l} = \overline{w}_{i,l}.$$

Finally, the Hessian on the simplex interior is

$$\nabla^2 \Phi_l(\mathbf{w}) = \mathrm{diag}\left(\frac{1}{w_1}, \ldots, \frac{1}{w_{N_t}}\right).$$

For every nonzero tangent direction $\mathbf{d}$ satisfying $\sum_i d_i = 0$, we have $\mathbf{d}^\top \nabla^2 \Phi_l(\mathbf{w})\mathbf{d} = \sum_i d_i^2/w_i > 0$. Thus the objective is strictly convex on the simplex interior. Together with the global KL identity, this establishes that Eq. (10) is the unique solution of Eq. (11). The derivation concerns normalized contribution mass for a fixed positive state transform and does not assign a probabilistic meaning to the state summary.

Rank-constrained barycenter. Fix the mass vector $\mathbf{w}_l^\star$ and define the weighted objective and its ambient mean by

$$F_l(M) = \sum_i w_{i,l}^\star \|M - \mu_{i,l}\|_F^2, \qquad \widetilde{\mu}_l = \sum_i w_{i,l}^\star \mu_{i,l}.$$

Because $\sum_i w_{i,l}^\star = 1$, expanding every squared norm gives

$$\begin{aligned}F_l(M) &= \sum_i w_{i,l}^\star \left(\|M\|_F^2 - 2\langle M, \mu_{i,l}\rangle_F + \|\mu_{i,l}\|_F^2\right) \\ &= \left(\sum_i w_{i,l}^\star\right)\|M\|_F^2 - 2\left\langle M, \sum_i w_{i,l}^\star \mu_{i,l}\right\rangle_F + \sum_i w_{i,l}^\star \|\mu_{i,l}\|_F^2 \\ &= \|M\|_F^2 - 2\langle M, \widetilde{\mu}_l\rangle_F + \sum_i w_{i,l}^\star \|\mu_{i,l}\|_F^2 \\ &= \|M - \widetilde{\mu}_l\|_F^2 + \sum_i w_{i,l}^\star \|\mu_{i,l}\|_F^2 - \|\widetilde{\mu}_l\|_F^2. \end{aligned} \tag{17}$$

The final two terms form the weighted dispersion around the mean. Indeed, expanding this dispersion gives

$$\begin{aligned} D_l &:= \sum_i w_{i,l}^\star \|\mu_{i,l} - \widetilde{\mu}_l\|_F^2 \\ &= \sum_i w_{i,l}^\star \|\mu_{i,l}\|_F^2 - 2\left\langle \sum_i w_{i,l}^\star \mu_{i,l}, \widetilde{\mu}_l \right\rangle_F + \left(\sum_i w_{i,l}^\star\right) \|\widetilde{\mu}_l\|_F^2 \\ &= \sum_i w_{i,l}^\star \|\mu_{i,l}\|_F^2 - \|\widetilde{\mu}_l\|_F^2. \end{aligned} \tag{18}$$

It also admits a pairwise form. Expanding the pairwise distances and using $\sum_i w_{i,l}^\star = 1$ gives

$$\begin{aligned} &\frac{1}{2}\sum_{i,j} w_{i,l}^\star w_{j,l}^\star \|\mu_{i,l} - \mu_{j,l}\|_F^2 \\ &\quad = \frac{1}{2}\sum_{i,j} w_{i,l}^\star w_{j,l}^\star \left(\|\mu_{i,l}\|_F^2 + \|\mu_{j,l}\|_F^2 - 2\langle \mu_{i,l}, \mu_{j,l}\rangle_F\right) \\ &\quad = \sum_i w_{i,l}^\star \|\mu_{i,l}\|_F^2 - \left\|\sum_i w_{i,l}^\star \mu_{i,l}\right\|_F^2 = D_l. \end{aligned}$$

Equivalently,

$$D_l = \frac{1}{2}\sum_{i,j} w_{i,l}^\star w_{j,l}^\star \|\mu_{i,l} - \mu_{j,l}\|_F^2.$$

Substituting Eq. (18) into Eq. (17) produces the orthogonal decomposition

$$F_l(M) = \|M - \widetilde{\mu}_l\|_F^2 + D_l. \tag{19}$$

The dispersion D_l depends on the submitted client updates and their masses, but not on the candidate M. Hence

$$\begin{aligned} \arg\min_{M \in \mathcal{M}_{r,l}} F_l(M) &= \arg\min_{M \in \mathcal{M}_{r,l}} \|M - \widetilde{\mu}_l\|_F^2 \\ &= \mathrm{Proj}_{r,l}(\widetilde{\mu}_l), \end{aligned}$$

which proves Eq. (13).

Solving the rank-constrained projection. Let $p = \min(d_{\mathrm{out},l}, d_{\mathrm{in},l})$ and write a singular-value decomposition

$$\widetilde{\mu}_l = U\,\mathrm{diag}(\sigma_1, \ldots, \sigma_p)V^\top, \qquad \sigma_1 \geq \cdots \geq \sigma_p \geq 0.$$

Lemma (singular-value trace inequality). Let $X, Y \in \mathbb{R}^{m\times n}$ have singular values $\alpha_1 \geq \cdots \geq \alpha_p \geq 0$ and $\beta_1 \geq \cdots \geq \beta_p \geq 0$, where $p = \min(m, n)$. Then

$$|\langle X, Y\rangle_F| \leq \sum_{j=1}^p \alpha_j \beta_j. \tag{20}$$

Proof. Let $s = \mathrm{rank}(Y)$ and take a compact singular-value decomposition

$$Y = \sum_{j=1}^s \beta_j u_j v_j^\top.$$

Define

$$a_j = u_j^\top X v_j, \qquad A_k = \sum_{j=1}^k a_j.$$

We first bound every partial sum A_k. Collect the first k singular vectors of Y as $U_k = [u_1, \ldots, u_k]$ and $V_k = [v_1, \ldots, v_k]$. Then

$$A_k = \mathrm{tr}(U_k^\top X V_k).$$

Let $X = P_X \Sigma_X Q_X^\top$ be a full singular-value decomposition and set $C = P_X^\top U_k$ and $D = Q_X^\top V_k$. Both C and D have orthonormal columns. If $c_t^\top$ and $d_t^\top$ denote their t-th rows, then

$$\begin{aligned} A_k &= \sum_{t=1}^{p} \alpha_t c_t^\top d_t \\ &\le \sum_{t=1}^{p} \alpha_t \frac{\|c_t\|_2^2 + \|d_t\|_2^2}{2} = \sum_{t=1}^{p} \alpha_t \gamma_t, \end{aligned}$$

where $\gamma_t = (\|c_t\|_2^2 + \|d_t\|_2^2)/2$. The row norms of a matrix with orthonormal columns are at most one, so $0 \le \gamma_t \le 1$. Moreover,

$$\sum_{t=1}^{p} \gamma_t \le \frac{\|C\|_F^2 + \|D\|_F^2}{2} = k.$$

Because the α_t are nonnegative and nonincreasing, the largest value of $\sum_t \alpha_t \gamma_t$ under these constraints is obtained by assigning unit mass to the first k indices. Hence

$$A_k \le \sum_{t=1}^{k} \alpha_t. \tag{21}$$

The Frobenius inner product can now be written in terms of these partial sums. Abel summation gives

$$\begin{aligned} \langle X, Y \rangle_F &= \sum_{j=1}^{s} \beta_j a_j \\ &= \sum_{k=1}^{s-1} (\beta_k - \beta_{k+1}) A_k + \beta_s A_s. \end{aligned}$$

All coefficients in the second line are nonnegative. Applying Eq. (21) therefore yields

$$\begin{aligned} \langle X, Y \rangle_F &\le \sum_{k=1}^{s-1} (\beta_k - \beta_{k+1}) \sum_{t=1}^{k} \alpha_t + \beta_s \sum_{t=1}^{s} \alpha_t \\ &= \sum_{j=1}^{s} \alpha_j \beta_j = \sum_{j=1}^{p} \alpha_j \beta_j, \end{aligned}$$

where the final equality pads the singular values of Y with zeros. Applying the same upper bound to $-Y$, which has the same singular values as Y, gives $-\langle X, Y \rangle_F \le \sum_j \alpha_j \beta_j$. Combining the two bounds proves Eq. (20).

Application to the rank-constrained projection. Assume $r \le p$. For any $M \in \mathcal{M}_{r,l}$, let $\tau_1 \ge \cdots \ge \tau_r \ge 0$ denote its nonzero singular values, padded with zeros when $\mathrm{rank}(M) < r$. The Frobenius distance satisfies

$$\|\widetilde{\mu}_l - M\|_F^2 = \|\widetilde{\mu}_l\|_F^2 + \|M\|_F^2 - 2\langle \widetilde{\mu}_l, M \rangle_F.$$

Applying Eq. (20) with $X = \widetilde{\mu}_l$ and $Y = M$ gives

$$\langle \widetilde{\mu}_l, M \rangle_F \le \sum_{j=1}^{r} \sigma_j \tau_j.$$

Combining the last two equations yields

$$\begin{aligned} \|\widetilde{\mu}_l - M\|_F^2 &\ge \sum_{j=1}^{p} \sigma_j^2 + \sum_{j=1}^{r} \tau_j^2 - 2 \sum_{j=1}^{r} \sigma_j \tau_j \\ &= \sum_{j=1}^{r} (\tau_j - \sigma_j)^2 + \sum_{j=r+1}^{p} \sigma_j^2 \\ &\ge \sum_{j=r+1}^{p} \sigma_j^2. \end{aligned} \tag{22}$$

The truncated matrix

$$\widetilde{\mu}_{l,r} = U_r \operatorname{diag}(\sigma_1, \ldots, \sigma_r) V_r^\top \tag{23}$$

has rank at most r, aligns its singular vectors with those of $\widetilde{\mu}_l$, and sets $\tau_j = \sigma_j$ for $j \le r$. It attains both inequalities in Eq. (22). Hence

$$\widetilde{\mu}_{l,r} \in \operatorname{Proj}_{r,l}(\widetilde{\mu}_l), \qquad \min_{M \in \mathcal{M}_{r,l}} \|\widetilde{\mu}_l - M\|_F^2 = \sum_{j=r+1}^{p} \sigma_j^2. \tag{24}$$

Combining this result with Eq. (19) gives the optimal barycenter value

$$\min_{M \in \mathcal{M}_{r,l}} F_l(M) = D_l + \sum_{j=r+1}^{p} \sigma_j^2.$$

The first term measures disagreement among client effective updates, while the second is the error required to return their mean to the LoRA rank constraint.

Proposition (existence, uniqueness, and coordinate invariance). Fix the positive masses $\{w_{i,l}^\star\}_i$ with $\sum_i w_{i,l}^\star = 1$, and suppose $1 \le r < p$. The projection set $\operatorname{Proj}_{r,l}(\widetilde{\mu}_l)$ is nonempty. Moreover, it is a singleton if and only if

$$\operatorname{rank}(\widetilde{\mu}_l) \le r \quad \text{or} \quad \sigma_r > \sigma_{r+1}. \tag{25}$$

For fixed masses, independently replacing the factors of each client by $(B_{i,l}Q_i, Q_i^{-1}A_{i,l})$, where Q_i is invertible, leaves the barycenter objective, its optimal value, and its complete solution set unchanged.

Proof of existence and the complete solution set. Eq. (23) explicitly constructs a minimizer, so the projection set is nonempty. The same conclusion also follows from the fact that $\mathcal{M}_{r,l}$ is closed and the squared-distance objective is coercive. If $\operatorname{rank}(\widetilde{\mu}_l) \le r$, the minimum projection error is zero. Only $M = \widetilde{\mu}_l$ attains zero distance, which proves uniqueness in the first case of Eq. (25).

It remains to characterize the case $\operatorname{rank}(\widetilde{\mu}_l) > r$. Let M have column space $\mathcal{S}$ and let $P_{\mathcal{S}}$ be the orthogonal projector onto that space. Since $M = P_{\mathcal{S}} M$, orthogonality gives

$$\|\widetilde{\mu}_l - M\|_F^2 = \|(I - P_{\mathcal{S}})\widetilde{\mu}_l\|_F^2 + \|P_{\mathcal{S}}\widetilde{\mu}_l - M\|_F^2. \tag{26}$$

Thus, after choosing a subspace $\mathcal{S}$ of dimension at most r, its best associated matrix is $M = P_{\mathcal{S}}\widetilde{\mu}_l$. Selecting the subspace is therefore equivalent to maximizing the captured energy

$$\begin{aligned} \|P_{\mathcal{S}}\widetilde{\mu}_l\|_F^2 &= \operatorname{tr}\big(P_{\mathcal{S}}\widetilde{\mu}_l\widetilde{\mu}_l^\top\big) \\ &= \sum_{j=1}^{p} \sigma_j^2 \|P_{\mathcal{S}} u_j\|_2^2. \end{aligned} \tag{27}$$

The coefficients $c_j = \|P_{\mathcal{S}} u_j\|_2^2$ satisfy $0 \le c_j \le 1$ and $\sum_j c_j \le \dim(\mathcal{S}) \le r$. Since the σ_j^2 are nonincreasing, the maximum captured energy is $\sum_{j=1}^{r} \sigma_j^2$.

This argument also gives the equality conditions. If $\sigma_r > \sigma_{r+1}$, every maximizing subspace must contain the first r left singular vectors and cannot contain a lower singular direction. Hence $\mathcal{S} = \operatorname{span}(U_r)$ and $M = P_{\mathcal{S}}\widetilde{\mu}_l = U_r\Sigma_r V_r^\top$ is unique.

Now suppose that $\sigma_r = \sigma_{r+1} = \tau > 0$. Let

$$a = \#\{j : \sigma_j > \tau\}, \qquad m = \#\{j : \sigma_j = \tau\}, \qquad k = r - a.$$

The truncation boundary lies strictly inside the tied block, so $1 \le k < m$. Denote the singular-vector blocks above and at the boundary by $(U_>, V_>)$ and (U_τ, V_τ), respectively. Equality in Eq. (27) requires a maximizing subspace to contain every direction in $U_>$, exclude every direction whose singular value is below τ, and select exactly k dimensions from $\operatorname{span}(U_\tau)$. Consequently, the complete solution set is

$$\operatorname{Proj}_{r,l}(\widetilde{\mu}_l) = \left\{ U_>\Sigma_> V_>^\top + \tau U_\tau P V_\tau^\top : P = P^\top = P^2,\ \operatorname{rank}(P) = k \right\}. \tag{28}$$

Indeed, every projector in this set chooses one such k-dimensional subspace and attains the residual in Eq. (24). Conversely, Eq. (26) and the equality conditions above force every minimizer to have this form. Because $1 \le k < m$, at least two distinct admissible projectors exist, proving nonuniqueness. Together with the preceding two cases, this establishes the if-and-only-if statement in Eq. (25).

Proof of coordinate invariance. For arbitrary invertible matrices Q_i, define

$$B'_{i,l} = B_{i,l}Q_i, \qquad A'_{i,l} = Q_i^{-1}A_{i,l}.$$

Then

$$\mu'_{i,l} = B'_{i,l}A'_{i,l} = B_{i,l}A_{i,l} = \mu_{i,l}.$$

For the fixed masses in the proposition, it follows term by term that

$$\widetilde{\mu}'_l = \widetilde{\mu}_l, \qquad D'_l = D_l, \qquad F'_l(M) = F_l(M) \quad \text{for every } M.$$

Therefore the optimal value and the entire projection set are invariant to independent changes of client factor coordinates.

The level of invariance depends on uniqueness. Under either condition in Eq. (25), invariance of the solution set immediately gives invariance of the unique global effective-update matrix. At a positive boundary tie, the set in Eq. (28) and its objective value remain invariant, but a particular representative is invariant only when the tie-breaking rule is defined from $\widetilde{\mu}_l$ and applied consistently. Different valid QR or SVD bases may otherwise select different members of the same optimal set; finite precision can further determine which member is returned.

Finally, this is a matrix-level statement. Even when the optimal effective update is unique, its returned factors are not: if $\mu^\star_{G,l} = B_{G,l}A_{G,l}$, then $(B_{G,l}Q, Q^{-1}A_{G,l})$ represents the same update for every invertible Q. The ideal consensus therefore guarantees invariance of the effective update rather than entrywise invariance of a particular factor pair. The claim above is conditional on the received masses; the exact effective-update state construction preserves this coordinate semantics, whereas the finite-damping qualification is analyzed in Section A.3.

A.2 Numerical Verification of Factorization Invariance

The preceding proof is a matrix-level statement. We next test whether the same invariance is preserved by the implemented product-space replay and by the QR-core rank return. For each factor record, we apply the product-preserving transformation

$$B'_i = B_iQ_i, \qquad A'_i = Q_i^{-1}A_i, \qquad Q_i \in \mathbb{R}^{r\times r} \text{ invertible},$$

so that $B'_iA'_i = B_iA_i$. We consider independent random orthogonal transforms, independent well-conditioned transforms with singular values in $[0.5, 2]$, and a shared global orthogonal transform. The metric is the relative Frobenius change $\|X' - X\|_F/(\|X\|_F + 10^{-12})$ from the untransformed aggregate, with the product-consensus implementation reusing the QR-core replay and the control averaging the B and A factors separately before multiplying them.

The product consensus changes by approximately 10^{-6} or less across all three transform families, consistent with float32 arithmetic. In contrast, independent reparameterizations change factor averaging by 0.78–1.04 at the median level and by as much as 17.9–26.8 in the worst group. The shared-global control leaves factor averaging unchanged because the same coordinate transform can be factored out of both averaged factors. The maximum relative change of each individual product B_iA_i was 1.963×10^{-7}, confirming that the transformations preserved the intended per-record products before aggregation.

This audit supports the matrix-level distinction used by BaryLoRA: product-space consensus is insensitive to the tested client-specific gauges, whereas separate factor averaging is not.

A.3 Effective-Update Direction and Damped Recovery

Minimum-norm compatible direction. For a single layer, write $P = BB^+$ and $R = A^+A$. These are the orthogonal projectors onto the column space of B and the row space of A, respectively. Let

$$H_B = \nabla_B\mathcal{L} = GA^\top, \qquad H_A = \nabla_A\mathcal{L} = B^\top G.$$

The exact direction is defined by the compatible minimum-norm problem

$$G_\mu = \arg\min_H \|H\|_F^2 \quad \text{subject to} \quad HA^\top = H_B, \qquad B^\top H = H_A. \tag{29}$$

The compatibility constraints in Eq. (29) imply

$$PH = PG, \qquad HR = GR.$$

Indeed, $B^\top H = B^\top G$ fixes the component in the column space of B, while $HA^\top = GA^\top$ fixes the action on the row space of A. Every compatible H can consequently be written as

$$H = PG + (I - P)GR + (I - P)H(I - R) = PG + GR - PGR + N,$$

where $N = (I - P)H(I - R)$. The free block N is Frobenius-orthogonal to the first three blocks. Hence

$$\|H\|_F^2 = \|PG + GR - PGR\|_F^2 + \|N\|_F^2,$$

and the unique minimum-norm solution sets $N = 0$:

$$G_\mu = PG + GR - PGR. \tag{30}$$

This proves Eq. (4). Under the transformation $(B, A) \mapsto (BQ, Q^{-1}A)$, the column space of B and row space of A are unchanged. The projectors P and R, and therefore G_μ, are independent of the chosen factor coordinates.

Damped computation from factor gradients. The implementation does not require an explicit unconstrained ambient gradient G. For damping coefficients $\delta_B, \delta_A > 0$, define

$$\widetilde{P}_B = B(B^\top B + \delta_B I)^{-1} B^\top,$$
$$\widetilde{P}_A = A^\top (AA^\top + \delta_A I)^{-1} A,$$

and compute directly from the factor gradients

$$X_A = B(B^\top B + \delta_B I)^{-1} H_A,$$
$$X_B = H_B(AA^\top + \delta_A I)^{-1} A,$$
$$\widetilde{G}_\mu = X_A + X_B - \widetilde{P}_B X_B.$$

When $H_A = B^\top G$ and $H_B = GA^\top$, these terms satisfy $X_A = \widetilde{P}_B G$ and $X_B = G\widetilde{P}_A$. Thus

$$\widetilde{G}_\mu = \widetilde{P}_B G + G\widetilde{P}_A - \widetilde{P}_B G \widetilde{P}_A,$$

which is the damped counterpart of Eq. (30).

Damping error and numerical meaning. The exact direction G_μ supplies the effective-update semantics used by the local moments, whereas the implementation uses $\widetilde{G}_\mu$ to avoid unstable Gram pseudoinverses. The following bound connects these two objects. If σ_B^+ and σ_A^+ denote the smallest nonzero singular values of B and A, respectively, define

$$e_B = \frac{\delta_B}{(\sigma_B^+)^2 + \delta_B}, \qquad e_A = \frac{\delta_A}{(\sigma_A^+)^2 + \delta_A},$$

with the corresponding term set to zero when the factor is the zero matrix. The singular-value decompositions of A and B give

$$\|\widetilde{P}_B - P\|_2 = e_B, \qquad \|\widetilde{P}_A - R\|_2 = e_A.$$

Let $E_B = \widetilde{P}_B - P$ and $E_A = \widetilde{P}_A - R$. Rewriting the exact and damped directions as

$$G_\mu = G - (I - P)G(I - R), \qquad \widetilde{G}_\mu = G - (I - \widetilde{P}_B)G(I - \widetilde{P}_A)$$

gives the exact error decomposition

$$\widetilde{G}_\mu - G_\mu = E_B G(I - R) + (I - \widetilde{P}_B)GE_A.$$

Both damped projectors are positive semidefinite contractions. Moreover, $E_B = E_B P$ and $E_A = RE_A$ because damping changes the nonzero spectral coefficients without changing the corresponding column and row spaces. Consequently,

$$\|\widetilde{G}_\mu - G_\mu\|_F \le e_B\|PG\|_F + e_A\|GR\|_F \le (e_B + e_A)\|G\|_F. \tag{31}$$

The first inequality is sharper than bounding the three terms of the damped formula separately and isolates the components identifiable from the two factor gradients.

The same result has a factor-gradient form that does not require materializing the ambient gradient. The compatibility identities give

$$PG = B(B^\top B)^+ H_A, \qquad GR = H_B(AA^\top)^+ A.$$

Substitution into Eq. (31) yields

$$\begin{aligned}\|\widetilde{G}_\mu - G_\mu\|_F &\le e_B\|B(B^\top B)^+ H_A\|_F + e_A\|H_B(AA^\top)^+ A\|_F \\ &\le \frac{e_B}{\sigma_B^+}\|H_A\|_F + \frac{e_A}{\sigma_A^+}\|H_B\|_F. \end{aligned} \tag{32}$$

The corresponding summand is defined as zero when its factor is the zero matrix. All quantities in the last line are available from the transmitted factor representation or the local backward pass, so this form can be used as a layer-wise numerical diagnostic.

Eqs. (31) and (32) make the conditioning tradeoff explicit. Damping stabilizes the Gram solves but attenuates directions associated with small nonzero singular values. In particular, for any $0 < \eta_B < 1$, choosing

$$\delta_B \le \frac{\eta_B}{1-\eta_B}(\sigma_B^+)^2$$

ensures $e_B \le \eta_B$, with an analogous relation for A. Hence $\widetilde{G}_\mu \to G_\mu$ in Frobenius norm as $\delta_A, \delta_B \downarrow 0$.

The error bound also quantifies the loss of exact factor-coordinate invariance at finite damping. Consider two equivalent factorizations of the same μ and let their damped directions be $\widetilde{G}_\mu^{(1)}$ and $\widetilde{G}_\mu^{(2)}$. Their exact target G_μ is shared, while their factor spectra and damping errors may differ. The triangle inequality gives

$$\|\widetilde{G}_\mu^{(1)} - \widetilde{G}_\mu^{(2)}\|_F \le \epsilon_{\text{damp}}^{(1)} + \epsilon_{\text{damp}}^{(2)}, \qquad \epsilon_{\text{damp}}^{(s)} = (e_B^{(s)} + e_A^{(s)})\|G\|_F.$$

Thus finite isotropic damping is generally not invariant to an arbitrary non-orthogonal change of factor coordinates, but its deviation from the shared invariant direction is controlled by the damping-to-spectrum ratios. The balanced SVD return in Eq. (35) removes arbitrary reciprocal rescaling of the two factors before the next step, and exact coordinate invariance is recovered in the zero-damping limit.

This analysis controls the first numerical link from factor gradients to the effective-update moments and the state summary $p_{i,l}$. It does not include stochastic-gradient noise, moment-recursion perturbations, randomized rank truncation, or the subsequent server aggregation error; those are distinct sources from the damping approximation analyzed here.

Persistent moments and state scope. The raw moments $(m_{i,l}, v_{i,l})$ and the cumulative local step are stored for each client–layer pair across communication rounds. Bias-corrected states, the prior direction, and μ^+ are temporary quantities computed within a local update. The server receives the recompressed factors and the scalar $p_{i,l}$; it does not receive the dense moments. This separation keeps the state-conditioned aggregation path one-way: local optimization state changes the server mass, while the server does not overwrite the client moments.

Local supervision. For classification, the local objective is LPC. On a minibatch $\mathcal{B}$ with observed label y_j, it is

$$\mathcal{L}_i^{\text{LPC}} = -\frac{\sum_{(x_j,y_j)\in\mathcal{B}} q_{i,y_j} \log p_{\theta_i}(y_j \mid x_j)}{\sum_{(x_j,y_j)\in\mathcal{B}} q_{i,y_j}}. \tag{33}$$

Let $N_{i,c}$ be the number of client-i examples in class c, $N_i = \sum_c N_{i,c}$, and $\mathcal{P}_i = \{c : N_{i,c} > 0\}$. Its normalized inverse-frequency weights are

$$\widetilde{q}_{i,c} = \frac{N_i}{N_{i,c}}, \qquad q_{i,c} = \frac{\widetilde{q}_{i,c}}{|\mathcal{P}_i|^{-1}\sum_{a\in\mathcal{P}_i}\widetilde{q}_{i,a}}, \quad c \in \mathcal{P}_i. \tag{34}$$

These weights determine the local loss in Eq. (33). For autoregressive generation, the model's language-model loss is cross-entropy over supervised token positions $\mathcal{T}_i$:

$$\mathcal{L}_i^{\mathrm{LM}} = -\frac{1}{|\mathcal{T}_i|} \sum_{(j,t)\in\mathcal{T}_i} \log p_{\theta_i}(y_{j,t} \mid x_j, y_{j,<t}).$$

LPC class weights are not applied to this generative objective.

Finite-threshold realization of local transforms. Eqs. (34) and (7) display the uncapped inverse-frequency and second-moment transforms. The implemented versions use finite bounds:

$$\begin{aligned}
\widetilde{q}_{i,c}^{\mathrm{impl}} &= \frac{1}{\max(N_{i,c}/N_i, \tau)}, && c \in \mathcal{P}_i, \\
\Omega_k^{\mathrm{impl}} &= \mathrm{clip}\left(\frac{\widehat{v}_k + \epsilon_{\mathrm{var}}}{s_{\mathrm{var}}}, 0, \Omega_{\max}\right), \\
z_{i,l}^{\mathrm{impl}} &= \mathrm{clip}\left(\frac{\widehat{v}_{i,l} + \epsilon_{\mathrm{var}}}{s_{\mathrm{var}}}, p_{\min}, p_{\max}\right), \\
p_{i,l}^{\mathrm{impl}} &= \left(\mathrm{mean}(z_{i,l}^{\mathrm{impl}})\right)^{\gamma}.
\end{aligned}$$

where $\tau > 0$, $\Omega_{\max} > 0$, and $0 < p_{\min} \le p_{\max}$. The normalization in Eq. (34) is applied to $\widetilde{q}_{i,c}^{\mathrm{impl}}$; the prior direction uses Ω_k^{impl}; and the transmitted summary is $p_{i,l}^{\mathrm{impl}}$. These bounds limit extreme local coefficients and preserve a finite positive state summary. They do not enter the consensus derivation, which requires only $p_{i,l} > 0$ and the bounded relative-state transform in Eq. (8).

A.4 Rank Truncation and Mass Stability

Exact and randomized rank return. Set $X = \mu_{k+1}^+$ and let

$$X_r = U_r \Sigma_r V_r^\top \in \mathrm{Proj}_{r,l}(X)$$

be an exact truncated SVD. Its balanced factor representation is

$$B_{k+1} = U_r \Sigma_r^{1/2}, \qquad A_{k+1} = \Sigma_r^{1/2} V_r^\top, \tag{35}$$

and its unavoidable rank-r compression error is

$$E_r^2(X) := \|X - X_r\|_F^2 = \sum_{j=r+1}^{p} \sigma_j^2(X), \qquad p = \min(d_{\mathrm{out},l}, d_{\mathrm{in},l}).$$

This is the mathematical return in Eq. (6).

For a large effective-update matrix, the implementation instead uses a randomized subspace iteration. Let $Q \in \mathbb{R}^{d_{\mathrm{out},l} \times q}$ have orthonormal columns spanning the sampled subspace, where $q \ge r$ is the sampled width and ν is the number of power iterations. Define the compressed matrix $C_Q = Q^\top X$ and let $[C_Q]_r$ be one of its exact rank-r truncated SVDs. The randomized return is

$$\widehat{\mu}_{k+1} = \widehat{\pi}_{r,q,\nu}(\mu_{k+1}^+) = Q[C_Q]_r \in \mathcal{M}_{r,l}.$$

The reference configuration uses $q = 24$ and $\nu = 4$. In exact arithmetic, a power-iteration range finder with a test matrix $\Omega \in \mathbb{R}^{d_{\mathrm{in},l} \times q}$ is driven by

$$(XX^\top)^\nu X\Omega = U\Sigma^{2\nu+1}V^\top\Omega,$$

so increasing ν amplifies the separation between leading and trailing singular directions before the sampled basis is formed.

Optimality inside the sampled subspace. Let $P_Q = QQ^\top$. Every matrix whose columns lie in $\mathrm{range}(Q)$ can be written as QZ. Orthogonality gives

$$\|X - QZ\|_F^2 = \|(I - P_Q)X\|_F^2 + \|C_Q - Z\|_F^2.$$

The first term is independent of Z. Applying the exact rank-r projection result to C_Q therefore proves

$$Q[C_Q]_r \in \arg \min_{\substack{\mathrm{rank}(M)\le r \\ \mathrm{col}(M)\subseteq \mathrm{range}(Q)}} \|X - M\|_F^2. \tag{36}$$

If $\widehat{\sigma}_1 \ge \cdots \ge \widehat{\sigma}_q \ge 0$ are the singular values of C_Q, padded with zeros when necessary, then

$$\begin{aligned}\|X - \widehat{\pi}_{r,q,\nu}(X)\|_F^2 &= \|(I - P_Q)X\|_F^2 + \sum_{j=r+1}^{q} \widehat{\sigma}_j^2 \\ &= \|X\|_F^2 - \sum_{j=1}^{r} \widehat{\sigma}_j^2.\end{aligned}$$

The first line separates energy missed by the sampled subspace from the rank truncation performed inside that subspace.

Error relative to the exact truncated SVD. Global optimality of X_r gives the lower bound

$$\|X - \widehat{\pi}_{r,q,\nu}(X)\|_F^2 - E_r^2(X) \ge 0.$$

For the converse direction, $P_Q X_r$ has rank at most r and columns in $\mathrm{range}(Q)$, so it is feasible in Eq. (36). Hence

$$\|X - \widehat{\pi}_{r,q,\nu}(X)\|_F^2 \le \|X - P_Q X_r\|_F^2.$$

Write $T_r = X - X_r$. The leading and trailing right singular subspaces are orthogonal, which implies $\langle T_r, (I - P_Q)X_r\rangle_F = 0$. Therefore

$$\|X - P_Q X_r\|_F^2 = E_r^2(X) + \|(I - P_Q)X_r\|_F^2.$$

Combining the last three relations yields the deterministic per-run bound

$$\boxed{0 \le \|X - \widehat{\pi}_{r,q,\nu}(X)\|_F^2 - E_r^2(X) \le \|(I - P_Q)X_r\|_F^2.} \tag{37}$$

Thus its excess objective is controlled by the leading rank-r energy missed by the sampled subspace. In particular, if $\mathrm{range}(Q)$ contains a leading left singular subspace of X, the upper bound vanishes and the randomized return belongs to $\mathrm{Proj}_{r,l}(X)$.

Two diagnostics distinguish total compression from randomized suboptimality. The total normalized residual is

$$\varepsilon_{\mathrm{proj}}(X) = \frac{\|X - \widehat{\pi}_{r,q,\nu}(X)\|_F}{\max(\|X\|_F, \epsilon)}$$

and the normalized excess error is

$$\varepsilon_{\mathrm{excess}}(X) = \frac{\sqrt{[\|X - \widehat{\pi}_{r,q,\nu}(X)\|_F^2 - E_r^2(X)]_+}}{\max(\|X\|_F, \epsilon)} \le \frac{\|(I - P_Q)X_r\|_F}{\max(\|X\|_F, \epsilon)},$$

where $[a]_+ = \max(a, 0)$ absorbs floating-point roundoff. The first diagnostic is available from the randomized return itself. The second requires the exact spectral tail and is therefore intended for small-scale numerical audits rather than the production path.

Eq. (37) compares objective values rather than the distance to one selected X_r. This distinction is necessary at a positive boundary tie, where Eq. (28) gives several exact rank-r projections that can be far apart while attaining the same residual. A probability bound in terms of q and ν additionally requires the sampling distribution and the reorthogonalization schedule; the deterministic bound above holds for every sampled subspace and keeps the exact method definition separate from its randomized realization.

Missing-state fallback. Let $a_{i,l} \in \{0,1\}$ indicate whether client i supplies a finite, positive state summary for layer l. Invalid values are replaced by a fixed $p_{\mathrm{def}} > 0$, and the state interpolation is reduced according to the valid-client fraction:

$$p'_{i,l} = a_{i,l}p_{i,l} + (1 - a_{i,l})p_{\mathrm{def}}, \qquad \rho_l = \rho_0 \frac{1}{N_t} \sum_{i \in \mathcal{S}_t} a_{i,l}, \tag{38}$$

where $\rho_0 \in [0,1]$. Substituting $p'_{i,l}$ into Eq. (8) preserves positivity. If no valid state is available, then $\rho_l = 0$, $f_{i,l} = 1$, and the aggregation reduces exactly to $w^\star_{i,l} = b_i$. Thus the safeguard keeps the state-conditioned mass map well defined under incomplete metadata and recovers the sample-size prior whenever a layer provides no usable state.

Bounded distortion of the base mass. The clipping in Eq. (8) gives $\chi_{i,l} \in [1/D, D]$. For a fixed ρ_l, define

$$f_l^- = 1 - \rho_l + \frac{\rho_l}{D}, \qquad f_l^+ = 1 - \rho_l + \rho_l D.$$

Then $0 < f_l^- \le f_{i,l} \le f_l^+$ and $f_l^- \le \sum_j b_j f_{j,l} \le f_l^+$. Combining these inequalities with Eq. (10) gives

$$\frac{f_l^-}{f_l^+} \le \frac{w^\star_{i,l}}{b_i} \le \frac{f_l^+}{f_l^-}.$$

The bound applies to the final normalized mass. In particular, clipping $\chi_{i,l}$ to $[1/D, D]$ does not imply $w^\star_{i,l}/b_i \in [1/D, D]$; for $\rho_l = 1$, the general bound is $[1/D^2, D^2]$. Thus D limits the worst-case leverage of layer-wise state, while ρ_l interpolates between the unmodified base mass and the full bounded tilt.

Sensitivity to the state transform. The preceding bound controls deviation from the base mass. We now bound how state perturbations propagate to the normalized mass, with the base mass, interpolation coefficient, and valid-client mask fixed. For two positive transforms $\mathbf{f}_l$ and $\widehat{\mathbf{f}}_l$, let $\delta_{i,l} = \log \widehat{f}_{i,l} - \log f_{i,l}$ and connect them by

$$\theta_{i,l}(t) = \log b_i + \log f_{i,l} + t\delta_{i,l}, \qquad w_{i,l}(t) = \frac{e^{\theta_{i,l}(t)}}{\sum_j e^{\theta_{j,l}(t)}}, \quad t \in [0,1].$$

The softmax Jacobian gives

$$\frac{\mathrm{d}w_{i,l}(t)}{\mathrm{d}t} = w_{i,l}(t)\left(\delta_{i,l} - \sum_j w_{j,l}(t)\delta_{j,l}\right), \qquad \sum_j \left|\frac{\partial w_{i,l}(t)}{\partial \theta_{j,l}(t)}\right| = 2w_{i,l}(t)(1 - w_{i,l}(t)) \le \frac{1}{2}.$$

Integrating along the path yields the finite-perturbation bound

$$\left|\widehat{w}^\star_{i,l} - w^\star_{i,l}\right| \le \frac{1}{2}\left\|\log \widehat{\mathbf{f}}_l - \log \mathbf{f}_l\right\|_\infty. \tag{39}$$

The interpolation in Eq. (9) further controls this response. Let $g_{\rho_l}(x) = \log(1 - \rho_l + \rho_l e^x)$ and define $\kappa_l = \rho_l D/(1 - \rho_l + \rho_l D)$. For $x = \log\chi \in [-\log D, \log D]$,

$$0 \le g'_{\rho_l}(x) = \frac{\rho_l e^x}{1 - \rho_l + \rho_l e^x} \le \kappa_l \le 1.$$

The mean-value theorem and Eq. (39) yield

$$\left|\widehat{w}^\star_{i,l} - w^\star_{i,l}\right| \le \frac{\kappa_l}{2}\left\|\log \widehat{\boldsymbol{\chi}}_l - \log \boldsymbol{\chi}_l\right\|_\infty.$$

It remains to relate this result to the positive state summaries. Set $\eta_{i,l} = \log \widehat{p}_{i,l} - \log p_{i,l}$, $\epsilon_l = \|\boldsymbol{\eta}_l\|_\infty$, and $q_{j,l} = p_{j,l}/\sum_k p_{k,l}$. Then

$$\frac{\widehat{\overline{p}}_l}{\overline{p}_l} = \sum_j q_{j,l} e^{\eta_{j,l}} \in [e^{-\epsilon_l}, e^{\epsilon_l}], \qquad \left\|\log \widehat{\boldsymbol{\chi}}_l - \log \boldsymbol{\chi}_l\right\|_\infty \le 2\epsilon_l.$$

The second inequality follows because relative normalization contributes at most another ϵ_l in log scale and clipping is non-expansive. Combining the bounds gives

$$\boxed{\left|\widehat{w}^{\star}_{i,l} - w^{\star}_{i,l}\right| \le \kappa_l \left\|\log \widehat{\mathbf{p}}_l - \log \mathbf{p}_l\right\|_\infty}.$$

Thus state-summary perturbations are not amplified by the state-to-mass map; $\rho_l = 0$ removes their influence. A change of the valid-client mask instead invokes the discrete fallback in Eq. (38) and lies outside this continuous bound.

A.5 QR-Core Equivalence and Computational Cost

Exact factorized projection. Let $d_o = d_{\mathrm{out},l}$, $d_i = d_{\mathrm{in},l}$, and $k = N_t r$. The weighted blocks are

$$L_l = \left[\sqrt{w^{\star}_{1,l}} B_{1,l} \quad \cdots \quad \sqrt{w^{\star}_{N_t,l}} B_{N_t,l}\right], \qquad R_l = \begin{bmatrix} \sqrt{w^{\star}_{1,l}} A_{1,l} \\ \vdots \\ \sqrt{w^{\star}_{N_t,l}} A_{N_t,l} \end{bmatrix}.$$

They satisfy $L_l \in \mathbb{R}^{d_o \times k}$ and $R_l \in \mathbb{R}^{k \times d_i}$. Block multiplication gives

$$L_l R_l = \sum_{i \in \mathcal{S}_t} w^{\star}_{i,l} B_{i,l} A_{i,l} = \widetilde{\mu}_l.$$

Thus the factorized representation is an identity for the ambient mean, not an approximation. Take thin QR factorizations

$$L_l = Q_L T_L, \qquad R_l^\top = Q_R T_R,$$

where $Q_L \in \mathbb{R}^{d_o \times q_L}$, $Q_R \in \mathbb{R}^{d_i \times q_R}$, $q_L = \min(d_o, k)$, $q_R = \min(d_i, k)$, and $Q_L^\top Q_L = I$, $Q_R^\top Q_R = I$. Defining $C_l = T_L T_R^\top \in \mathbb{R}^{q_L \times q_R}$ gives

$$\widetilde{\mu}_l = Q_L C_l Q_R^\top, \qquad C_l = T_L T_R^\top. \tag{40}$$

We first show that the ambient projection can be restricted to these QR subspaces without changing its solution. Let $P_L = Q_L Q_L^\top$ and $P_R = Q_R Q_R^\top$. For any $X \in \mathcal{M}_{r,l}$, define its two-sided subspace projection $X_\parallel = P_L X P_R$. Then $\mathrm{rank}(X_\parallel) \le r$, and $\widetilde{\mu}_l = P_L \widetilde{\mu}_l P_R$. Moreover, $\widetilde{\mu}_l - X_\parallel$ is Frobenius-orthogonal to $X - X_\parallel$, so

$$\|\widetilde{\mu}_l - X\|_F^2 = \|\widetilde{\mu}_l - X_\parallel\|_F^2 + \|X - X_\parallel\|_F^2.$$

Any component outside the QR subspaces therefore increases the objective and cannot occur in an optimum.

Every matrix in the retained subspaces has the form $X_\parallel = Q_L Z Q_R^\top$, where $Z = Q_L^\top X_\parallel Q_R$. Orthogonality of the QR bases preserves both rank and Frobenius norm:

$$\mathrm{rank}(Q_L Z Q_R^\top) = \mathrm{rank}(Z), \qquad \|Q_L Z Q_R^\top\|_F = \|Z\|_F. \tag{41}$$

Combining Eqs. (40)–(41) yields the solution-set equivalence

$$\mathrm{Proj}_{r,l}(\widetilde{\mu}_l) = \left\{ Q_L Z Q_R^\top : Z \in \arg\min_{\mathrm{rank}(Y) \le r} \|C_l - Y\|_F^2 \right\}. \tag{42}$$

This identity reduces the ambient rank-constrained problem to the core without assuming that its minimizer is unique.

Let $C_l = U_C \Sigma_C V_C^\top$ be a compact SVD with singular values $\sigma_1 \ge \cdots \ge \sigma_s > 0$. Because $Q_L U_C$ and $Q_R V_C$ also have orthonormal columns,

$$\widetilde{\mu}_l = (Q_L U_C) \Sigma_C (Q_R V_C)^\top$$

is a compact SVD of the ambient mean. The truncated core $C_{l,r} = U_{C,r} \Sigma_{C,r} V_{C,r}^\top$ therefore satisfies

$$\min_{\mathrm{rank}(Y) \le r} \|C_l - Y\|_F^2 = \|C_l - C_{l,r}\|_F^2 = \sum_{j > r} \sigma_j^2.$$

Mapping this minimizer through Eq. (42) gives

$$\mu^\star_{G,l} = Q_L U_{C,r} \Sigma_{C,r} V_{C,r}^\top Q_R^\top \in \mathrm{Proj}_{r,l}(\widetilde{\mu}_l).$$

The balanced split in Eq. (14) obeys

$$B_{G,l} A_{G,l} = Q_L U_{C,r} \Sigma_{C,r}^{1/2} \Sigma_{C,r}^{1/2} V_{C,r}^\top Q_R^\top = \mu^\star_{G,l}.$$

If the retained boundary singular value is tied, each optimal core truncation maps to the corresponding member of the ambient solution set; if fewer than r singular values are positive, zero padding preserves the product. Hence QR-core returns a member of the same optimal projection set as a dense truncated SVD in exact arithmetic. When the projection is unique, both methods return the same effective-update matrix. The computation changes without changing the aggregation objective.

Task-head inheritance. Let $h_i = \sum_l \alpha_l w^\star_{i,l}$, where $\boldsymbol{\alpha} \in \Delta^{L-1}$. Reordering the sums gives

$$\sum_l \alpha_l \sum_i w^\star_{i,l} \|\theta - \theta_i^{\mathrm{head}}\|_2^2 = \sum_i h_i \|\theta - \theta_i^{\mathrm{head}}\|_2^2.$$

Because $\sum_i h_i = 1$, differentiating the right-hand side yields the unique minimizer $\theta = \sum_i h_i \theta_i^{\mathrm{head}}$, which is Eq. (15). Head inheritance consequently propagates the already-computed layer masses and introduces no additional state model.

Communication and client storage. For one adapted layer, transmitting the LoRA factors requires

$$O(r(d_{\mathrm{out},l} + d_{\mathrm{in},l}))$$

scalars. State-conditioned aggregation adds one scalar $p_{i,l}$. The current client realization explicitly maintains the effective update and its two dense moments, giving $O(d_{\mathrm{out},l} d_{\mathrm{in},l})$ auxiliary storage per adapted layer. Randomized subspace iteration with sampled width q uses matrix products with μ^+ and $(\mu^+)^\top$; with a materialized dense update, these products cost $O((\nu+1) d_{\mathrm{out},l} d_{\mathrm{in},l} q)$ up to constant factors.

Server computation. Write $d_o = d_{\mathrm{out},l}$, $d_i = d_{\mathrm{in},l}$, and $k = N_t r$. The server forms $L_l \in \mathbb{R}^{d_o \times k}$ and $R_l^\top \in \mathbb{R}^{d_i \times k}$ directly from the uploaded weighted factors. Their thin QR dimensions are $q_L = \min(d_o, k)$ and $q_R = \min(d_i, k)$, so that $Q_L \in \mathbb{R}^{d_o \times q_L}$, $T_L \in \mathbb{R}^{q_L \times k}$, $Q_R \in \mathbb{R}^{d_i \times q_R}$, and $T_R \in \mathbb{R}^{q_R \times k}$.

A Householder thin QR of an $m \times k$ matrix costs $O(mk \min(m,k))$. The two factorizations therefore require $O(d_o k q_L + d_i k q_R)$ operations. The core multiplication $C_l = T_L T_R^\top \in \mathbb{R}^{q_L \times q_R}$ costs $O(q_L q_R k)$, and a full SVD of this core costs $O(q_L q_R \min(q_L, q_R))$. Finally, constructing the returned factors in Eq. (14) requires $O(d_o q_L r + d_i q_R r)$ operations to multiply the retained singular vectors by Q_L and Q_R. The complete per-layer server cost is thus

$$\begin{aligned} &O\big(d_o k q_L + d_i k q_R + q_L q_R k \\ &\quad + q_L q_R \min(q_L, q_R) + d_o q_L r + d_i q_R r\big). \end{aligned} \tag{43}$$

Since $r \le k$, the lifting terms are no larger than the corresponding QR terms.

In the intended low-rank regime $k \le \min(d_o, d_i)$, we have $q_L = q_R = k$. Eq. (43) then reduces to

$$\text{time} = O\left((d_o + d_i)k^2 + k^3\right), \qquad \text{storage} = O\left((d_o + d_i)k + k^2\right).$$

where the two QR factorizations contribute $O((d_o + d_i)k^2)$ and core formation plus core SVD contributes $O(k^3)$. The storage bound covers the factor blocks and QR bases together with the triangular and core matrices. In-place QR can reuse the block storage but does not change this asymptotic bound. The returned adapter has rank at most $\min(r, \mathrm{rank}(C_l))$.

For comparison, explicitly forming $\widetilde{\mu}_l = L_l R_l \in \mathbb{R}^{d_o \times d_i}$ costs $O(d_o d_i k)$ operations and $O(d_o d_i)$ storage before any rank truncation. An exact dense SVD then costs $O(d_o d_i \min(d_o, d_i))$. QR-core removes these mandatory $d_o d_i$ terms when $k \ll \min(d_o, d_i)$ by performing the spectral decomposition on a core of size at most $k \times k$. Its advantage decreases as $k = N_t r$ approaches the ambient dimensions, so the general bound in Eq. (43) remains the appropriate statement outside the low-rank regime.

The algebraic solver never needs to materialize the dense mean. A diagnostic implementation may still form client products or $\widetilde{\mu}_l$ to measure projection error; that additional $O(d_o d_i)$ memory belongs to the diagnostic path rather than the QR-core solve.

Exact, approximate, and empirical components. The origin $\mu = \Delta W/s = BA$, the mass normalization, the barycenter decomposition, factorization invariance, and QR-core equivalence are exact under the stated assumptions. Damped Gram recovery and randomized local rank truncation are numerical approximations to their corresponding exact operators. Mapping the second-moment summary to a larger contribution mass is an empirical design choice; none of the preceding identities assigns that summary the meaning of statistical precision or calibrated client quality.

Numerical round protocol. Algorithm A1 records the finite-threshold client and server path corresponding to the round description in the main text. Its local dense moments persist across rounds for each client–layer pair. The server receives factors and one summary per layer, rather than the moments themselves.

Algorithm A1 Numerical client update and server aggregation

Require: Participating clients $\mathcal{S}_t$, global factors, persistent $(m_{i,l}, v_{i,l}, k_{i,l})$, damping and truncation settings

for $i \in \mathcal{S}_t$ **in parallel do**
  $(B_{i,l}, A_{i,l}) \leftarrow (B_{G,l}, A_{G,l})$ for each l
  for local minibatches and adapted layers **do**
    Evaluate $\mathcal{L}_i^{\mathrm{LPC}}$ or $\mathcal{L}_i^{\mathrm{LM}}$; obtain (H_A, H_B)
    $X_A \leftarrow B(B^\top B + \delta_B I)^{-1} H_A$
    $X_B \leftarrow H_B(AA^\top + \delta_A I)^{-1} A$
    $\widetilde{G}_\mu \leftarrow X_A + X_B - \widetilde{P}_B X_B$; set $g_k \leftarrow -\widetilde{G}_\mu$
    $k \leftarrow k+1$; $m_k \leftarrow \beta_1 m_{k-1} + (1-\beta_1) g_k$
    $v_k \leftarrow \beta_2 v_{k-1} + (1-\beta_2) g_k \odot g_k$
    $\widehat{m}_k \leftarrow m_k/(1-\beta_1^k)$; $\widehat{v}_k \leftarrow v_k/(1-\beta_2^k)$
    $\Omega_k^{\mathrm{impl}} \leftarrow \mathrm{clip}((\widehat{v}_k + \epsilon_{\mathrm{var}})/s_{\mathrm{var}}, 0, \Omega_{\max})$
    $u_k^{\mathrm{prior}} \leftarrow -\lambda_{\mathrm{prior}} \zeta_i \omega_k (\Omega_k^{\mathrm{impl}} \odot \mu_k)$
    $\mu_{k+1}^{+} \leftarrow \mu_k + \eta s_{\mathrm{upd}} (\widehat{m}_k + u_k^{\mathrm{prior}})/(\sqrt{\widehat{v}_k} + \epsilon)$
    $\mu_{k+1} \leftarrow \pi_{r,l}(\mu_{k+1}^{+})$ or $\widehat{\pi}_{r,q,\nu}(\mu_{k+1}^{+})$
    Refactor $\mu_{k+1} = U_r \Sigma_r V_r^\top$ as $B_{i,l} \leftarrow U_r \Sigma_r^{1/2}$, $A_{i,l} \leftarrow \Sigma_r^{1/2} V_r^\top$
  end for
  For each l, set $z_{i,l}^{\mathrm{impl}} \leftarrow \mathrm{clip}((\widehat{v}_{i,l} + \epsilon_{\mathrm{var}})/s_{\mathrm{var}}, p_{\min}, p_{\max})$
  $p_{i,l}^{\mathrm{impl}} \leftarrow [\mathrm{mean}(z_{i,l}^{\mathrm{impl}})]^\gamma$
  Upload $(B_{i,l}, A_{i,l}, p_{i,l}^{\mathrm{impl}})_l$ and any task head
end for
for each adapted layer l **do**
  Set $a_{i,l} \leftarrow \mathbf{1}\{p_{i,l} \text{ valid}\}$ and $p'_{i,l} \leftarrow a_{i,l} p_{i,l} + (1-a_{i,l}) p_{\mathrm{def}}$
  $\rho_l \leftarrow \rho_0 N_t^{-1} \sum_i a_{i,l}$; $\overline{p}_l \leftarrow N_t^{-1} \sum_i p'_{i,l}$
  $\chi_{i,l} \leftarrow \mathrm{clip}(p'_{i,l}/\overline{p}_l, 1/D, D)$; $f_{i,l} \leftarrow 1 - \rho_l + \rho_l \chi_{i,l}$
  $w_{i,l}^\star \leftarrow b_i f_{i,l} / \sum_j b_j f_{j,l}$
  $L_l \leftarrow \mathrm{hcat}_i(\sqrt{w_{i,l}^\star} B_{i,l})$; $R_l \leftarrow \mathrm{vcat}_i(\sqrt{w_{i,l}^\star} A_{i,l})$
  $L_l = Q_L T_L$, $R_l^\top = Q_R T_R$; $C_l \leftarrow T_L T_R^\top$
  Compute $C_l = U_C \Sigma_C V_C^\top$
  $B_{G,l} \leftarrow Q_L U_{C,r} \Sigma_{C,r}^{1/2}$; $A_{G,l} \leftarrow \Sigma_{C,r}^{1/2} V_{C,r}^\top Q_R^\top$
end for
if a separate task head is aggregated **then**
  $h_i \leftarrow L^{-1} \sum_l w_{i,l}^\star$; $\theta_G^{\mathrm{head}} \leftarrow \sum_i h_i \theta_i^{\mathrm{head}}$
end if

Here hcat and vcat denote horizontal and vertical concatenation of the weighted factor blocks, respectively.

B DETAILED EXPERIMENTAL SETTINGS

Tables 6 and 7 give the BaryLoRA configurations used for the two task families. In these tables, B is the batch size, S is the number of local steps, R is the number of communication rounds, and L is the maximum sequence length. The GLUE LoRA tuple $(r, \alpha_{\mathrm{L}}, p)$ denotes rank, scaling, and dropout; the generative table reports (r, α_{L}) and uses B/S for batch size and local steps.

Table 6: BaryLoRA training settings across GLUE datasets.

Dataset	N	Dirichlet h	LoRA $(r, \alpha_{\mathrm{L}}, p)$	B	S	R	LR	L
SST-2	$3, 10, 50$	0.5	$(4, 8, 0.05)$	128	10	250	5×10^{-4}	128
QNLI	$3, 10, 50$	0.5	$(4, 8, 0.05)$	128	10	250	5×10^{-4}	128
QQP	$3, 10, 50$	0.5	$(4, 8, 0.05)$	128	10	250	5×10^{-4}	128
RTE	$3, 10, 50$	0.5	$(4, 8, 0.05)$	128	10	250	5×10^{-4}	128
MNLI	$3, 10, 50$	0.5	$(4, 8, 0.05)$	128	10	250	5×10^{-4}	128

In the client-scale comparison, the three-client runs use exact local rank return, whereas the ten- and fifty-client runs use randomized low-rank return. The two paths share the listed training parameters but are different implementations, so client scale and rank-return backend are not separated in that comparison. The MNLI rank comparison also reports $r \in \{2, 4, 8, 16, 24\}$, and its heterogeneity comparison reports $h \in \{0.5, 1, 100\}$. The QNLI and QQP heterogeneity comparison reports $h \in \{0.1, 0.5, 5.0\}$.

Table 7: BaryLoRA training settings for generative datasets.

Evaluation	Backbone	Training data	Client split	N	R	LoRA (r, α_{L})	B/S	LR	L
HumanEval	Llama-3-8B	CSN	meta, 5%	6	201	$(8, 16)$	$1/30$	5×10^{-3}	650
HumanEval	Qwen3.5-9B	CSN	meta, 5%	6	201	$(8, 16)$	$1/30$	5×10^{-3}	650
GSM8K	Llama-3-8B	GSM8K	IID, 100%	3	201	$(8, 16)$	$1/30$	5×10^{-3}	650
GSM8K	Qwen3.5-9B	GSM8K	IID, 100%	3	201	$(8, 16)$	$1/30$	5×10^{-3}	650

The generative configurations use a $0.998/0.001/0.001$ train/validation/test split of each fine-tuning set and cap the sequence length at 1000. HumanEval uses pass@1; GSM8K uses exact-match accuracy. The Qwen3.5-9B CodeSearchNet run targets `q_proj` and `v_proj`; its GSM8K run also targets `in_proj_qkv`.

Convergence protocol. Figure 2 compares QNLI and QQP over communication rounds 1 through 200 with $N = 3$, Dirichlet concentration $h = 0.5$, LoRA rank $r = 4$, and AdamW with learning rate 5×10^{-4} and seed 42. The plotted accuracy and average-loss values are raw per-round observations without smoothing.

B.1 BASELINE AGGREGATION RULES

FedIT (Zhang et al., 2024) trains both LoRA factors and averages them factor-wise. SHAREAB is our local implementation of the same factor-sharing rule and is reported separately in the generative comparison. **FlexLoRA** (Bai et al., 2024) targets heterogeneous tasks and client resources by aggregating full-rank updates and redistributing them to the local factors. **FFA-LoRA** (Sun et al., 2024) freezes the A factor and shares only B. **RoLoRA** (Chen et al., 2025) alternates the optimization of the two factors. **FedRot-LoRA** (Zhang et al., 2026a) aligns the local factors with an orthogonal rotation before aggregation. Each baseline retains its own optimization rule and is evaluated with the shared task metrics and evaluation scripts.

B.2 MNLI RANK SENSITIVITY

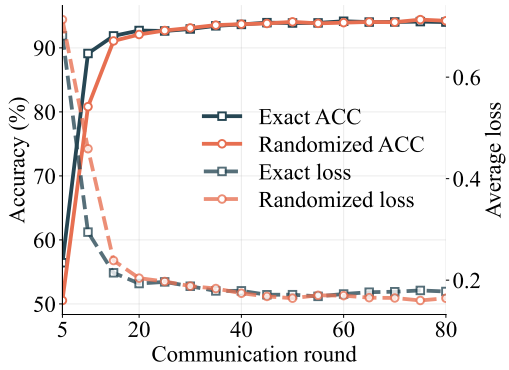
Table 8 reports MNLI accuracy at ranks 2, 4, 8, 16, and 24. BaryLoRA has the highest mean in every column, ranging from 0.883 to 0.894, and remains ahead of FedRot-LoRA by 0.006 to 0.022 in the displayed settings. At ranks 2 and 4, it already reaches 0.890 and 0.889, respectively. The results therefore show competitive accuracy at small rank, although the accuracy is not monotonic in rank.

Table 8: MNLI accuracy under different LoRA ranks ($r \in \{2, 4, 8, 16, 24\}$). BaryLoRA has the highest mean accuracy in the displayed columns.

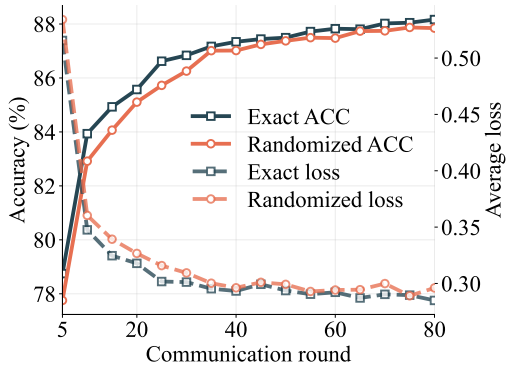
Rank	Methods	ACC±std@MNLI
2	FedIT	0.844 ± 0.051
	FlexLoRA	0.811 ± 0.108
	FFA-LoRA	0.837 ± 0.008
	RoLoRA	0.860 ± 0.002
	FedRot-LoRA	0.877 ± 0.005
	BaryLoRA (ours)	$\mathbf{0.890 \pm 0.003}$
4	FedIT	0.866 ± 0.001
	FlexLoRA	0.845 ± 0.008
	FFA-LoRA	0.862 ± 0.001
	RoLoRA	0.868 ± 0.005
	FedRot-LoRA	0.876 ± 0.002
	BaryLoRA (ours)	$\mathbf{0.889 \pm 0.012}$
8	FedIT	0.861 ± 0.001
	FlexLoRA	0.836 ± 0.021
	FFA-LoRA	0.864 ± 0.001
	RoLoRA	0.869 ± 0.004
	FedRot-LoRA	0.877 ± 0.002
	BaryLoRA (ours)	$\mathbf{0.883 \pm 0.006}$
16	FedIT	0.855 ± 0.003
	FlexLoRA	0.852 ± 0.010
	FFA-LoRA	0.834 ± 0.009
	RoLoRA	0.722 ± 0.215
	FedRot-LoRA	0.865 ± 0.002
	BaryLoRA (ours)	$\mathbf{0.886 \pm 0.013}$
24	FedIT	0.857 ± 0.001
	FlexLoRA	0.864 ± 0.009
	FFA-LoRA	0.844 ± 0.001
	RoLoRA	0.746 ± 0.274
	FedRot-LoRA	0.872 ± 0.001
	BaryLoRA (ours)	$\mathbf{0.894 \pm 0.007}$

B.3 Exact and randomized rank-return diagnostics

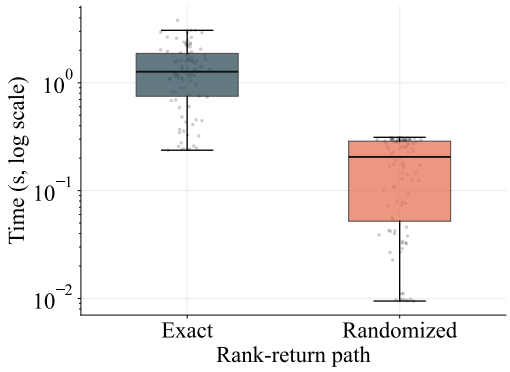
This subsection reports the exact and randomized rank-return comparison.



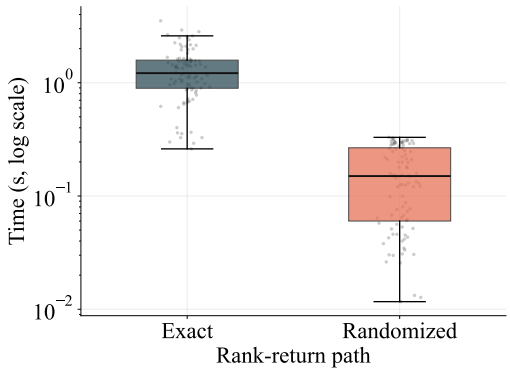
(a) QNLI accuracy and loss over communication rounds.



(b) QQP accuracy and loss over communication rounds.

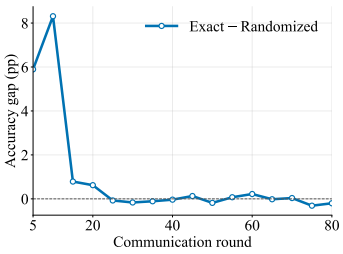


(c) QNLI recompression time.

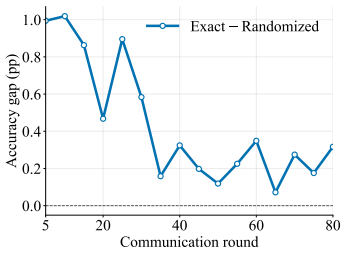


(d) QQP recompression time.

Figure 3: Accuracy, loss, and recompression time for exact and randomized rank return. Both methods use the same rank-constrained consensus; only the return backend changes.

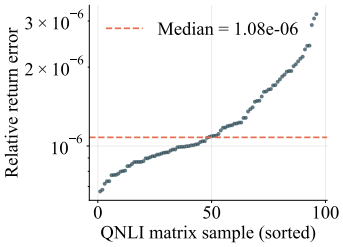


(a) QNLI

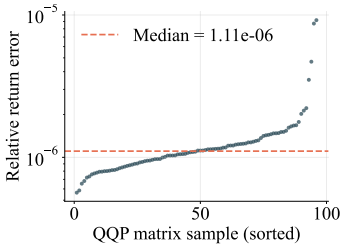


(b) QQP

Figure 4: Accuracy difference (exact minus randomized rank return) at each matched validation checkpoint.



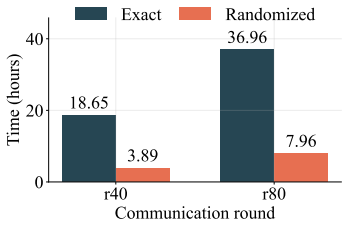
(a) QNLI



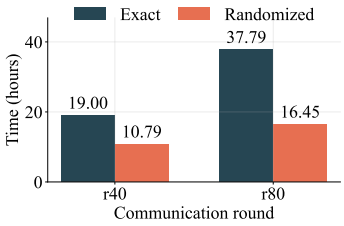
(b) QQP

Figure 5: Relative reconstruction error of randomized rank return on the same server-side matrix.

The accuracy curves show an early gap between the two rank-return implementations. The gap becomes small after the first few checkpoints, but the later curves are not identical at every point. The reconstruction panels report median relative errors of 1.08×10^{-6} on QNLI and 1.11×10^{-6} on QQP. The time panels show lower measured time for the randomized backend at the displayed rounds. Together, these diagnostics support randomized rank return as an approximate, lower-cost realization of the same rank-constrained consensus on these two tasks.



(a) QNLI



(b) QQP

Figure 6: End-to-end time for exact and randomized rank return.

C SUPPLEMENTARY P0 AUDITS

This appendix records supplementary evidence that supports the scope of the main claims. We organize the empirical evidence into two groups. The first tests whether product-space consensus is insensitive to equivalent factorizations and is reported in Appendix A.2. The second examines whether the state-conditioned rule produces non-degenerate, layer-specific client masses. The final subsection is an analytic communication-cost audit rather than a training experiment; it counts the tensors attached to the client upload and distinguishes them from client-local state and server-side computation.

C.1 STATE-CONDITIONED CLIENT–LAYER DIAGNOSTICS

Layer-resolved mass variation. For each adapted module, we record the realized client masses $w^{\star}_{i,l}$, the normalized state summary $\chi_{i,l}$, and the reallocation factor $f_{i,l}$. Table 9 summarizes the resulting client–layer records. The mean client spread is the average range of $w^{\star}_{i,l}$ across the three clients. The cross-layer standard deviation is computed within each client–round and then averaged; its maximum is included to show the observed range.

Table 9: Layer-mass diagnostics over the instrumented MNLI run.

Run	Rounds	Layers	Client-layer records	Client spread	Cross-layer std	$\mathrm{std}(\chi)$	$\overline{f}$
MNLI instrumented	99	48	14256	0.176	0.0570 / 0.1972	0.283	1.003

The average mass remains $1/3$ because masses are normalized across clients, but the nonzero spreads show that the normalized state summary changes the relative client contribution at a given layer. The nonzero cross-layer standard deviations further show that the assignment is not a single client-level weight copied to every adapted module. A client-shared-state control would have zero cross-layer variation by construction.

C.2 COMMUNICATION-COST AUDIT

For an adapted module of shape $d_{\mathrm{out}} \times d_{\mathrm{in}}$ and rank r, a client uploads $r(d_{\mathrm{out}} + d_{\mathrm{in}})$ factor scalars. BaryLoRA adds one positive state summary $p_{i,l}$ per adapted module. Thus, with L adapted modules and fp32 payloads, its additional client-to-server cost is L scalars, or $4L$ bytes per round. Table 10 compares this payload with FedIT and records the corresponding factor payloads for the evaluated backbones.

Table 10: Analytic client-to-server communication audit. Bytes count raw fp32 tensor payloads; the ratios are relative to FedIT.

Backbone	Method	Factors	Summaries	Total bytes	Ratio
RoBERTa-large	BaryLoRA	393216	48	5779656	1.000033
	FFA-LoRA	196608	0	4993032	0.863926
	RoLoRA	196608	0	4993032	0.863926

For RoBERTa-large with $r = 4$ and $L = 48$, BaryLoRA adds 192 bytes, which is 0.012% of the factor payload. The dense effective-update moments used to construct $p_{i,l}$ remain on the client and are not part of the transmitted payload. QR-core is likewise a server-side operation. FLoRA, FlexLoRA, and FedEx-LoRA keep the client factor payload at the FedIT level under the homogeneous-rank setting, although their server-side realization or broadcast differs.

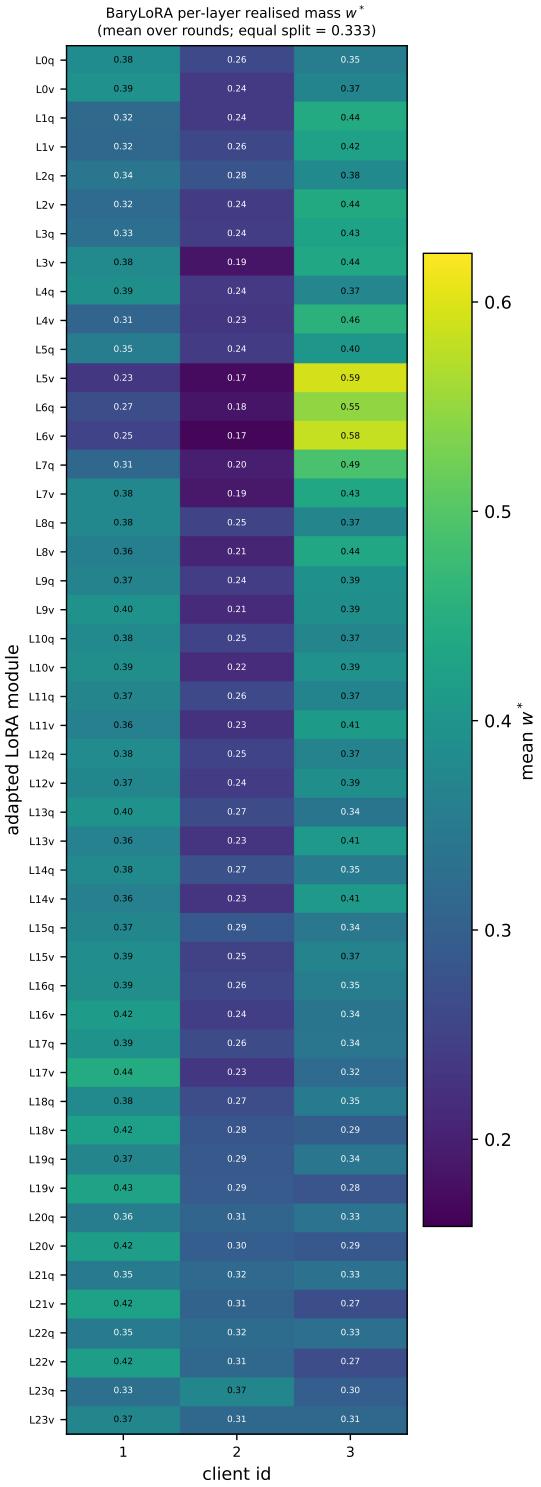


Figure 7: MNLI client–layer heatmap of the realized mass $w^{\star}_{i,l}$, averaged over recorded rounds. Each row is one adapted LoRA module and each column is one client; the equal-share reference is $1/3$. The non-constant rows and column patterns are consistent with a layer-resolved state-conditioned mass assignment.