FEDCODI: DIRECTIONAL COHERENCE FOR PERSONALIZED FEDERATED LOW-RANK ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated low-rank adaptation enables efficient fine-tuning of pretrained language models on decentralized data. Under heterogeneous data, however, client updates can point in conflicting directions and partially cancel when averaged. We characterize this *directional cancellation* through pairwise directional coherence and observe it in client adapters during federated training. Mitigating cancellation must also preserve client-specific adaptations, rather than simply enforce uniform directions. To address this tension, we propose **FedCoDi**, a personalized federated fine-tuning framework that coordinates how clients incorporate global directions and form subsequent local updates. Building on a magnitude–direction decomposition, FedCoDi retains trainable weight magnitudes locally and shares both low-rank factors. A row-wise coherence gate blends broadcast and previous local factors, retaining more local state where their directions disagree. During local training, a complementary regularizer penalizes angular disagreement between the local low-rank product and the unblended global product, which serves as a shared reference across clients. Both mechanisms operate on clients, while the server retains standard factor-wise averaging. Experiments with RoBERTa-large and Llama-3-8B on four GLUE tasks, GSM8K, and Dolly-15K show consistent gains over the evaluated federated LoRA baselines, achieving substantial improvements over the strongest compared baseline in each setting.

1 INTRODUCTION

Federated learning (FL) enables clients to collaboratively train models without centralizing their training data McMahan et al. (2017). Low-rank adaptation (LoRA) makes this collaboration more practical for pretrained language models by reducing the number of trainable and communicated parameters Hu et al. (2022); Zhang et al. (2024); Babakniya et al. (2023). However, heterogeneous client data can require different model adaptations Zhao et al. (2018). Personalized federated fine-tuning must therefore share useful knowledge across clients without discarding adaptations needed by individual clients.

Existing federated LoRA methods address heterogeneity by modifying parameter sharing or refining aggregation. FFA-LoRA Sun et al. (2024) freezes one low-rank factor, whereas FedSA-LoRA Guo et al. (2025) keeps one factor local. LoRA-FAIR Bian et al. (2025) refines aggregation and client initialization, while FedEx-LoRA Singhal et al. (2025) corrects the mismatch between factor-wise and update-level aggregation. These approaches address important aspects of what is shared and how client updates are combined. However, even an accurately aggregated adapter need not be equally suitable for every client. Correcting aggregation is therefore distinct from deciding which global directions a personalized model should incorporate and how its subsequent local training should be guided.

We examine this problem through the directional coherence of low-rank updates. Local training can produce updates whose directions disagree across clients. When these update products are averaged, opposing components can partially cancel, reducing the aggregate norm relative to fully coherent updates with the same magnitudes. We characterize this *directional cancellation* through pairwise cosine similarity (Theorem 1) and explicitly distinguish it from the mismatch introduced by separately averaging the low-rank factors (Proposition 1, Lemma 1). Yet directional differences can

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.

also reflect client-specific adaptations. Preserving every local deviation can weaken collaboration, whereas enforcing uniform agreement can restrict personalization.

This tension calls for different treatment at two stages of a federated round. When incorporating the global adapter, a client should selectively combine shared information with its existing local state rather than replace that state indiscriminately. During local training, the client also needs a shared directional reference to discourage divergence that can increase conflict in the next aggregation. The resulting design objective is not to eliminate all directional differences, but to balance selective global transfer with coherence during local optimization.

We propose **FedCoDi**, a personalized federated fine-tuning framework that implements these complementary operations. Building on DoRA's magnitude–direction decomposition Liu et al. (2024), we retain weight magnitudes locally and share both low-rank factors. *Decoupled Directional Coherence Alignment (DDCA)* blends the broadcast factors with the client's previous local factors using row-wise cosine coherence, backed by formal selection guarantees (Proposition 2). Rows with lower coherence retain more of the local state, while rows with higher coherence receive more of the global state. During local training, *Directional Coherence Regularization (DCR)* penalizes angular disagreement between the local low-rank product and the product of the unblended broadcast factors, with a relative strength invariant to task loss scale (Proposition 3). Using the unblended product retains a shared reference rather than a different personalized anchor for each client. DDCA thus controls how global information enters the local model, while DCR guides the update subsequently contributed to collaboration. Both mechanisms run on clients; the server continues to average the two factors separately, and local magnitudes are not transmitted. Our main contributions are summarized as follows:

- We characterize directional cancellation in federated low-rank adaptation in closed form (Theorem 1), verify the prediction empirically on probed client adapters, and quantify its dependence on pairwise coherence, distinguishing this structural degradation from factor-wise aggregation mismatch (Proposition 1, Lemma 1).
- We propose **FedCoDi**, which combines row-wise coherence gating for selective global transfer (DDCA, Proposition 2) with product-level directional regularization during local training (DCR, Proposition 3).
- Experiments on GLUE, GSM8K, and Dolly-15K show consistent gains over the evaluated federated PEFT baselines. Ablation and efficiency studies further examine component contributions and the communication rounds needed to reach target accuracy.

2 Related Work

Federated Low-Rank Adaptation. FedIT Zhang et al. (2024) applies federated averaging to both LoRA factors, while SLoRA Babakniya et al. (2023) improves adaptation through data-driven initialization. Other methods modify factor training and sharing. FFA-LoRA Sun et al. (2024) freezes A, FedSA-LoRA Guo et al. (2025) shares A while retaining B locally, and RoLoRA Chen et al. (2025) alternates optimization of the two factors across rounds. Heterogeneous-rank methods Bai et al. (2024); Cho et al. (2024) accommodate differences in client resources. Aggregation refinement is another line of work. LoRA-FAIR Bian et al. (2025) refines aggregation and client initialization, while FedEx-LoRA Singhal et al. (2025) compensates for the error introduced by factor-wise aggregation. Our focus is *directional cancellation*, which can occur even when update products are averaged exactly. FedCoDi manages directional coherence on clients while keeping both factors trainable, rather than modifying the server's factor-wise averaging rule.

Personalization and Client Drift. Personalized FL combines shared knowledge with client-specific adaptation. FedRep Collins et al. (2021) and FedRoD Chen & Chao (2022) combine shared representations with local prediction heads. FedProx Li et al. (2020) and Ditto Li et al. (2021) penalize deviations from global parameters. SCAFFOLD Karimireddy et al. (2020) uses control variates to address client drift, while FedDC Gao et al. (2022) decouples and corrects local drift. In multi-task learning, PCGrad Yu et al. (2020) addresses conflicting task gradients through projection.

FedCoDi adopts DoRA's magnitude–direction decomposition Liu et al. (2024) and retains trainable magnitudes locally while sharing both low-rank factors. Personalization also arises from row-wise

mixing of global and previous local factors. In contrast to parameter-distance penalties and task-gradient projection, DCR penalizes angular disagreement between the local low-rank product and the unblended global product during local training. This provides a shared directional reference alongside the personalized initialization produced by DDCA.

3 Directional Cancellation: A Motivating Analysis

We first define the low-rank parameterization, then quantify directional cancellation and examine it in client adapters. Finally, we relate the analyzed update average to factor-wise aggregation.

3.1 Low-Rank and Decoupled Parameterizations

Consider a pretrained weight matrix $W_0 \in \mathbb{R}^{d_{out}\times d_{in}}$. Standard LoRA Hu et al. (2022) parameterizes the weight update as a low-rank decomposition: $W = W_0 + \Delta W = W_0 + BA$, where $B \in \mathbb{R}^{d_{out}\times r}$ and $A \in \mathbb{R}^{r\times d_{in}}$ with rank $r \ll \min(d_{out}, d_{in})$.

To separate weight magnitude from direction, we adopt the magnitude–direction decomposition of DoRA Liu et al. (2024). For the j-th output neuron, the weight row is

$$W_{j,:} = m_j \frac{W_{0,j,:} + B_{j,:}A}{\|W_{0,j,:} + B_{j,:}A\|_2}, \tag{1}$$

where m_j is the corresponding entry of the trainable magnitude vector $M \in \mathbb{R}^{d_{out}}$, W_0 remains frozen, and $B_{j,:}A$ is the low-rank term that adapts the direction.

Remark 1 (Scope of the analysis)**.** *The following analysis concerns the unnormalized low-rank products BA and their averages, so it also applies to standard LoRA. In the decoupled parameterization these products are not generally equal to the full weight changes $W - W_0$, nor are they round-to-round parameter increments. We do not assume that their pairwise coherence is preserved after adding W_0 and normalizing.*

3.2 Quantitative Analysis of Directional Cancellation

Consider one aggregation round with N participating clients. Each client i retains its trainable magnitude vector $M^{(i)}$ locally and uploads the low-rank factors $A^{(i)}$ and $B^{(i)}$. The server averages the factors separately:

$$A^g = \sum_{i=1}^{N} w_i A^{(i)}, \quad B^g = \sum_{i=1}^{N} w_i B^{(i)}, \qquad w_i \geq 0,\ \textstyle\sum_i w_i = 1. \tag{2}$$

Under heterogeneous data, clients can learn low-rank updates whose directions disagree. We first isolate their cancellation in the average of update products; its relation to B^gA^g is addressed in Section 3.3.

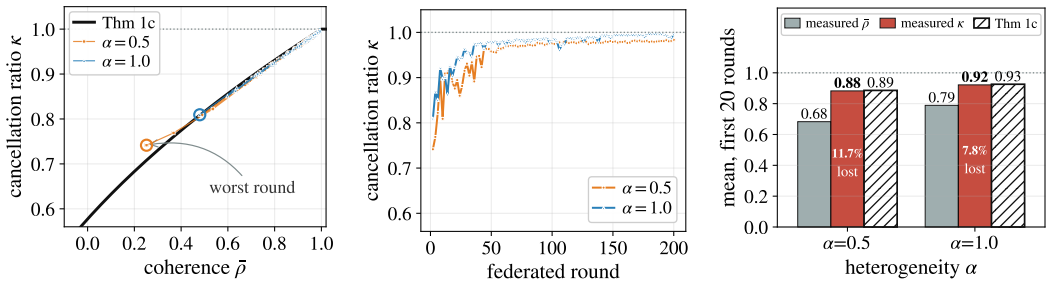
Theorem 1 (Directional Cancellation)**.** *Let $N \geq 2$ clients have weights $w_i \geq 0$ with $\sum_i w_i = 1$. For a fixed output neuron j, let $u_j^{(i)} = B_{j,:}^{(i)}A^{(i)} \in \mathbb{R}^{d_{in}}$ be the directional update of client i and $u_j^g := \sum_{i=1}^N w_i\, u_j^{(i)}$ their weighted average. Let $\rho_{ik}^{(j)} = \cos(u_j^{(i)}, u_j^{(k)})$ for nonzero rows and $\bar\rho^{(j)} = \binom{N}{2}^{-1}\sum_{i<k}\rho_{ik}^{(j)}$ the mean coherence. Then:*

(a) Closed-form expansion. *The squared aggregated norm decomposes as:*

$$\|u_j^g\|_2^2 = \sum_{i=1}^{N} w_i^2\|u_j^{(i)}\|_2^2 + 2\sum_{i<k} w_i w_k \rho_{ik}^{(j)}\|u_j^{(i)}\|_2\|u_j^{(k)}\|_2. \tag{3}$$

(b) Norm deficit relative to aligned updates. *Since $\rho_{ik}^{(j)} \leq 1$:*

$$\begin{aligned}\|u_j^g\|_2^2 &= \left(\sum_{i=1}^{N} w_i\|u_j^{(i)}\|_2\right)^2 - 2\sum_{i<k} w_i w_k\big(1-\rho_{ik}^{(j)}\big)\|u_j^{(i)}\|_2\|u_j^{(k)}\|_2 \\ &\leq \left(\sum_{i=1}^{N} w_i\|u_j^{(i)}\|_2\right)^2.\end{aligned} \tag{4}$$



(a) (b) (c)

Figure 1: **(a)** Measured cancellation ratios versus mean pairwise coherence. **(b)** Cancellation ratios over federated rounds. **(c)** A persistent early-round (first 20 round) loss at both heterogeneity levels.

Equality holds if and only if all positively weighted nonzero update rows have the same direction. Thus, for fixed update norms and aggregation weights, disagreement among these updates reduces the aggregate norm relative to the fully aligned case.

(c) Equal-weight equal-norm formula. *When* $w_i = 1/N$ *and* $\|u_j^{(i)}\|_2 = c > 0$*:*

$$\|u_j^g\|_2 = c\sqrt{\frac{1 + (N-1)\bar{\rho}^{(j)}}{N}}. \tag{5}$$

In this special case, mean pairwise coherence determines the retained fraction of the update norm. Complete cancellation occurs when $\bar{\rho}^{(j)} = -1/(N-1)$.

The proof, together with a supplementary lower bound under equal update norms, is provided in Appendix A.2. The norm identities also extend to zero update rows through the inner-product expansion, without evaluating their undefined cosines.

Remark 2. *The theorem characterizes a geometric norm deficit, not a task-loss or accuracy guarantee. Directional disagreement can also reflect client-specific adaptations, so a larger aggregate norm is not, by itself, a sufficient objective for personalized learning.*

Empirical observation. We probe the factor-averaging baseline on three-client MNLI. For each output row, we measure pairwise coherence and the cancellation ratio $\kappa_j = \frac{\left\|\sum_i w_i u_j^{(i)}\right\|_2}{\sum_i w_i \|u_j^{(i)}\|_2}$, for nonzero denominators. Under the equal-weight, equal-norm assumptions of Theorem 1(c), κ_j follows the stated coherence curve, which we use only as a reference outside that setting. The measured ratios closely follow this curve (correlation 0.997, MAE 0.004; Fig. 1a), cancellation is strongest early in training (Fig. 1b), and the first-20-round mean norm deficit is larger for $\alpha = 0.5$ than for $\alpha = 1.0$ (11.7% vs. 7.8%; Fig. 1c). This motivates selective global–local blending on the incoming path (DDCA) and directional regularization on the outgoing path (DCR), rather than uniform agreement. Probe settings and interpretive caveats are given in Appendix A.1.

3.3 From Analyzed Average to Aggregated Product

Theorem 1 concerns the *average of update products* $u_j^g = \sum_i w_i u_j^{(i)}$, whereas the server averages A and B separately (Eq. 2). The resulting product $B^g A^g$ is also distinct from a client's personalized product after DDCA blending. We first quantify the difference between update averaging and factor-wise aggregation, then relate factor deviations to product deviations—that is, we tie the analysis to the object that is actually formed on clients.

Proposition 1 (Aggregation Mismatch Decomposition)**.** *Let* $u_j^g = \sum_{i=1}^N w_i\, u_j^{(i)}$ *and let* $u_j^{\text{agg}} = B_{j,:}^g A^g$ *denote the product formed from the server-averaged factors, before client-side blending. Then*

$$u_j^{\text{agg}} = u_j^g - \mathcal{E}_j, \qquad \mathcal{E}_j := \sum_{i=1}^N w_i \left(B_{j,:}^{(i)} - B_{j,:}^g\right)\left(A^{(i)} - A^g\right), \tag{6}$$

and the mismatch term satisfies $\|\mathcal{E}_j\|_2 \le \sum_{i=1}^{N} w_i \left\|B_{j,:}^{(i)} - B_{j,:}^{g}\right\|_2 \left\|A^{(i)} - A^{g}\right\|_F$. *If both factor deviations are uniformly* $O(\epsilon)$ *across clients, then* $\|\mathcal{E}_j\|_2 = O(\epsilon^2)$. *In particular,*

$$\left|\|u_j^{\text{agg}}\|_2 - \|u_j^{g}\|_2\right| \le \|\mathcal{E}_j\|_2. \tag{7}$$

The proof is provided in Appendix A.3. This decomposition separates directional cancellation in the average of products from factor-wise aggregation mismatch. In regimes with small factor deviations the aggregated product thus tracks the analyzed average u_j^g up to a second-order term, so the cancellation quantified in Theorem 1 is the dominant effect rather than an artifact of factor-wise averaging. The second-order bound is conditional on small factor deviations: it does not establish that FedCoDi always satisfies this condition, nor that the relative mismatch is small when the averaged product is near zero.

Lemma 1 (Component-to-Product Deviation Bound)**.** *Let* $B^{(i)}, A^{(i)}$ *and* B^g, A^g *denote the local and global directional matrices for a given layer. Define the component-level deviations* $\Delta_B = \|B^{(i)} - B^g\|_F$ *and* $\Delta_A = \|A^{(i)} - A^g\|_F$, *and let* $\|B^g\|_F = \beta$, $\|A^{(i)}\|_F = \alpha$. *Then the product-level deviation satisfies:*

$$\|B^{(i)}A^{(i)} - B^g A^g\|_F \le \Delta_B \cdot \alpha + \beta \cdot \Delta_A. \tag{8}$$

Consequently, if $\Delta_B \le \epsilon_B$ *and* $\Delta_A \le \epsilon_A$, *then the product-level deviation is bounded by* $\epsilon_B\alpha + \beta\epsilon_A$, *which is* $O(\epsilon)$ *when* $\epsilon_B, \epsilon_A = O(\epsilon)$ *and the factor norms remain bounded.*

Remark 3. *Lemma 1 controls an* absolute *deviation. Converting it into a product-level* coherence *statement requires a non-degeneracy condition: if* $\|B^g A^g\|_F \ge c_0 > 0$ *and the product deviation satisfies* $\delta = \epsilon_B\alpha + \beta\epsilon_A < c_0$, *then* $\cos(\mathrm{vec}(B^{(i)}A^{(i)}), \mathrm{vec}(B^gA^g)) \ge (c_0 - \delta)/(c_0 + \delta)$, *which approaches* 1 *as* $\delta \to 0$ *(derivation in Appendix A.4). The bound does not apply to a zero reference. This distinction motivates complementing factor-wise initialization with product-level directional regularization during local training: DDCA selectively mixes the incoming factors, while DCR directly penalizes angular disagreement of the local product with the global reference.*

4 Our Method: FedCoDi

We use the decoupled parameterization introduced in Section 3. Within one federated round, a client receives the broadcast factors, forms a personalized initialization with DDCA, trains locally under DCR, and uploads only the low-rank factors; the magnitude vector $M^{(i)}$ remains trainable and local. The server continues to average the two factors separately. Figure 2 provides an overview.

4.1 Decoupled Directional Coherence Alignment (DDCA)

At the start of a round, let $A^{(i)}$ and $B^{(i)}$ denote the client's retained local factors, before incorporating the broadcast factors A^g and B^g. Direct replacement would discard this local factor state, and the broadcast factors need not be equally suitable for every neuron. **Decoupled Directional Coherence Alignment (DDCA)** therefore uses row-wise directional coherence to determine the relative contributions of the global and local factors.

For the output projection $B \in \mathbb{R}^{d_{out} \times r}$, whose rows index output neurons, we measure the directional coherence between the global and local representations by cosine similarity:

$$\mathrm{sim}(B_{j,:}^g, B_{j,:}^{(i)}) = \frac{B_{j,:}^g \cdot B_{j,:}^{(i)}}{\|B_{j,:}^g\|_2 \|B_{j,:}^{(i)}\|_2 + \epsilon}, \tag{9}$$

where $\epsilon > 0$ is a small constant (e.g., 10^{-8}) for numerical stability. We map this similarity to an alignment coefficient $G_j^B \in [0, 1]$ via a parameterized sigmoid function:

$$G_j^B = \sigma\left(\tau \cdot \left(\mathrm{sim}(B_{j,:}^g, B_{j,:}^{(i)}) - \theta\right)\right), \tag{10}$$

where $\tau > 0$ controls the sharpness and $\theta \in [-1, 1]$ is the coherence threshold. For finite τ we have $G_j^B < 1/2$ below the threshold and $G_j^B > 1/2$ above it; the limits $G_j^B \to 0$ and $G_j^B \to 1$ are approached as τ grows. The blended matrix B_{aligned} is constructed row-wise:

$$B_{\text{aligned},j,:} = G_j^B B_{j,:}^g + (1 - G_j^B) B_{j,:}^{(i)}. \tag{11}$$

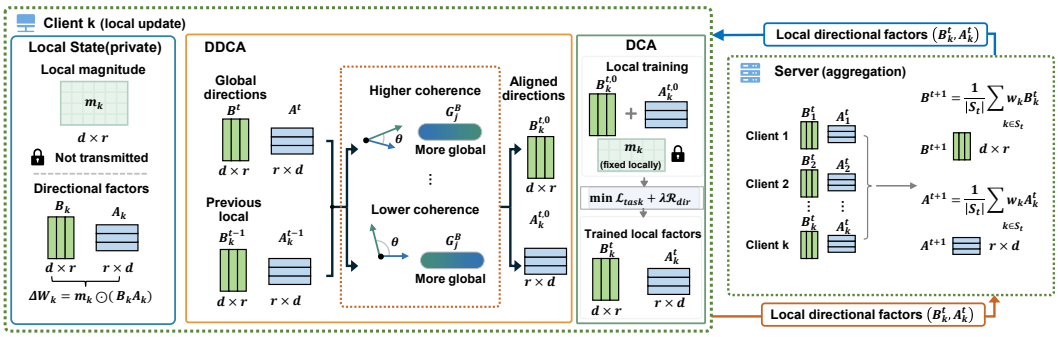


Figure 2: The FedCoDi framework. On the *incoming* path, DDCA blends the global directions with the previous local ones via row-wise coherence gating; on the *outgoing* path, DCR softly penalizes angular deviation from the cached unblended reference $\mathcal{W}^g$; the magnitudes M remain local.

Thus lower-coherence rows retain a larger contribution from the local state, whereas higher-coherence rows receive a larger global contribution.

The same row-wise gating is applied to the input down-projection $A \in \mathbb{R}^{r \times d_{in}}$ with rows indexed by k: writing $G_k^A = \sigma\big(\tau \cdot (\mathrm{sim}(A_{k,:}^g, A_{k,:}^{(i)}) - \theta)\big)$, we obtain $A_{\text{aligned},k,:} = G_k^A A_{k,:}^g + (1 - G_k^A) A_{k,:}^{(i)}$. Here the rows of A index low-rank components rather than output neurons. The client loads B_{aligned} and A_{aligned} as its local factors and retains its current $M^{(i)}$.

The following proposition characterizes the mixing error under an assumption linking coherence to the relative quality of the global and local rows.

Proposition 2 (Conditional DDCA Selection Bound)**.** *Let $B_{j,:}^{(i)}$ and $B_{j,:}^g$ denote the local and global directional vectors for neuron j, and let $B_{j,:}^*$ denote an idealized per-row reference: the minimizer of client i's local loss for row j with all other parameters held fixed, which serves only as an analysis anchor and need not be jointly attainable across rows. Define $\rho_j = \mathrm{sim}(B_{j,:}^g, B_{j,:}^{(i)})$, $G_j = \sigma(\tau(\rho_j - \theta))$, and $e_j^g = \|B_{j,:}^g - B_{j,:}^*\|_2$, $e_j^l = \|B_{j,:}^{(i)} - B_{j,:}^*\|_2$.*

(a) Pointwise bound. *For any neuron j:*

$$\|B_{\text{aligned},j,:} - B_{j,:}^*\|_2 \le G_j\, e_j^g + (1 - G_j)\, e_j^l. \tag{12}$$

(b) Comparison with fixed-weight bounds. *Assume the* coherence-quality monotonicity *condition: $\rho_j > \theta \Rightarrow e_j^g \le e_j^l$ and $\rho_j < \theta \Rightarrow e_j^g \ge e_j^l$. Let $\mathcal{H} = \{j : \rho_j > \theta\}$, $\mathcal{L} = \{j : \rho_j < \theta\}$, and suppose $m = \min_j |\rho_j - \theta| > 0$. If there exists $j_H \in \mathcal{H}$ with $e_{j_H}^g < e_{j_H}^l$ and $j_L \in \mathcal{L}$ with $e_{j_L}^l < e_{j_L}^g$, then there exists a finite τ_0 such that for all $\tau \ge \tau_0$ and every fixed interpolation weight $\alpha \in [0, 1]$:*

$$\sum_j \big(G_j e_j^g + (1 - G_j) e_j^l\big)^2 \le \sum_j \big(\alpha\, e_j^g + (1 - \alpha) e_j^l\big)^2. \tag{13}$$

(c) Limit behavior. *As $\tau \to \infty$, $G_j \to \mathbb{I}[\rho_j > \theta]$ and the bound becomes $\sum_j \min\big((e_j^g)^2, (e_j^l)^2\big)$, corresponding to selecting the lower-error global or local row independently for each j. This limit requires the monotonicity and margin conditions, but not the two-sided strictness used in (b).*

Scope. *The comparison in (b) concerns parameter-error bounds, not realized task losses: it certifies a tighter error bound than any fixed interpolation, rather than a uniformly smaller realized error. It also does not establish that a finite threshold is attained by the default τ.*

The proof is provided in Appendix A.5.

Remark 4. *The result is* conditional *on the coherence–quality monotonicity assumption, which is a modeling assumption rather than a consequence of cosine similarity. The intended intuition is that global–local agreement can indicate reusable shared structure, while disagreement may indicate a local adaptation worth retaining. Coherence is measured against the global direction B^g, whose*

quality can itself be reduced by cancellation (Theorem 1); the assumption should therefore be read as a working hypothesis about the relative *ordering of e_j^g and e_j^l. Its justification and empirical verifiability are discussed in Appendix A.8.*

4.2 Directional Coherence Regularization (DCR)

DDCA provides a personalized initialization, but subsequent local optimization can again move client updates in different directions. DCR complements this initialization with a soft penalty on directional disagreement during local training.

Let $\mathcal{W}^g = B^g A^g$ be the global reference product, formed from the unblended factors received from the server. It is cached at the start of the round and held fixed during local optimization, while $B^{(i)}A^{(i)}$ changes with training. For nonzero products, DCR measures the directional drift of the local product using the cosine distance:

$$\mathcal{L}_{\text{DCR}} = 1 - \cos\left(\text{vec}(B^{(i)}A^{(i)}), \text{vec}(\mathcal{W}^g)\right), \tag{14}$$

where $\cos(\cdot,\cdot)$ is the cosine similarity between vectorized matrices. DCR deliberately acts on the vectorized product as a whole, in contrast to DDCA's row-wise gating: at the broadcast, row-level coherence is measurable and rows carry distinct specialization, so fine-grained selection pays off; during local training, a single global drift signal is preferable, as it constrains the trajectory without overriding row-level task learning.

DCR regularizes local training toward the raw broadcast consensus $\mathcal{W}^g$ rather than the personalized blend $B_{\text{aligned}}A_{\text{aligned}}$ produced by DDCA: the unblended product is shared by all clients that receive the same broadcast in a round, whereas the blend depends on each client's retained factors. Regularizing toward the blend would anchor clients at different personalized starting points, whereas $\mathcal{W}^g$ provides a single shared reference. A detailed rationale is provided in Appendix A.7.

Since $\mathcal{L}_{\text{DCR}}$ is scale-free (bounded in $[0, 2]$), naively adding it to $\mathcal{L}_{\text{task}}$ leads to inconsistent regularization across tasks with different loss magnitudes; we instead apply DCR as a *relative modulation factor*:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} \cdot (1 + \lambda \cdot \mathcal{L}_{\text{DCR}}), \tag{15}$$

where λ controls the regularization strength. The following proposition makes this scale invariance precise.

Proposition 3 (Loss-Ratio Scale Invariance of Multiplicative DCR)**.** *For $\mathcal{L}_{task} > 0$, the* loss ratio *(regularization contribution relative to task loss) under the additive form $\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda\mathcal{L}_{DCR}$ is $R_{\text{add}} = \lambda\mathcal{L}_{DCR}/\mathcal{L}_{task}$, which scales with $1/\mathcal{L}_{task}$ and thus varies across tasks with different loss magnitudes. Under the multiplicative form $\mathcal{L}_{total} = \mathcal{L}_{task} \cdot (1 + \lambda\mathcal{L}_{DCR})$, the loss ratio is $R_{\text{mul}} = \lambda\mathcal{L}_{DCR}$: holding λ and $\mathcal{L}_{DCR}$ fixed, rescaling the task loss leaves this ratio unchanged.*

The proof is provided in Appendix A.6.

Remark 5. *The invariance concerns ratios of loss values under task-loss rescaling. It does not imply identical regularization gradients across different tasks, whose task gradients and directional disagreement can differ, nor does it guarantee that λ never requires tuning.*

Local optimization and communication. Clients optimize $A^{(i)}$, $B^{(i)}$, and $M^{(i)}$, together with the task head when present, using $\mathcal{L}_{\text{total}}$. The directional penalty contains no direct magnitude-matching term, while $M^{(i)}$ continues to adapt through the local task loss. After training, clients upload only $A^{(i)}$ and $B^{(i)}$ for the factor-wise averaging in Eq. 2; their magnitudes and the retained local factor states needed for the next DDCA step remain on the clients. The complete procedure is summarized in Algorithm 1 (Appendix B).

5 Experiments

5.1 Experimental Setup

Baselines and adapter configuration. We compare with FedIT Zhang et al. (2024), FFA-LoRA Sun et al. (2024), FedSA-LoRA Guo et al. (2025), LoRA-FAIR Bian et al. (2025), FedEx-LoRA Singhal

Table 1: Accuracy (%) on GLUE under non-IID (Dirichlet $\alpha = 0.5$) partitions across three client scales (3 and 10 clients with full participation; 50 clients with 10% sampling per round).

Clients	Method	MNLI	QNLI	QQP	SST-2	Avg.
3	FedIT	88.76 ± 0.08	92.37 ± 0.20	85.83 ± 0.07	95.06 ± 0.14	90.51
	FFA-LoRA	88.52 ± 0.21	91.63 ± 0.33	86.11 ± 0.36	94.97 ± 0.17	90.31
	FedSA-LoRA	88.84 ± 0.07	$\underline{92.78} \pm 0.32$	84.54 ± 0.21	$\underline{95.17} \pm 0.13$	90.33
	LoRA-FAIR	$\underline{88.95} \pm 0.10$	92.57 ± 0.19	87.28 ± 0.14	95.13 ± 0.05	$\underline{90.98}$
	FedEx-LoRA	88.63 ± 0.11	92.74 ± 0.20	$\underline{87.43} \pm 0.06$	95.08 ± 0.08	90.97
	RoLoRA	88.56 ± 0.05	92.10 ± 0.36	87.08 ± 0.05	94.83 ± 0.08	90.64
	FedCoDi	$\mathbf{90.21} \pm 0.14$	$\mathbf{93.30} \pm 0.34$	$\mathbf{87.79} \pm 0.18$	$\mathbf{95.41} \pm 0.17$	**91.68**
10	FedIT	87.03 ± 0.29	91.06 ± 0.39	86.23 ± 0.04	96.57 ± 0.06	90.22
	FFA-LoRA	86.47 ± 0.36	90.55 ± 1.26	83.80 ± 0.12	95.30 ± 0.13	89.03
	FedSA-LoRA	$\underline{87.43} \pm 0.21$	$\underline{91.88} \pm 0.36$	82.44 ± 0.08	$\underline{96.64} \pm 0.06$	89.60
	LoRA-FAIR	87.17 ± 0.32	91.13 ± 0.43	86.86 ± 0.17	96.51 ± 0.50	$\underline{90.42}$
	FedEx-LoRA	87.34 ± 0.18	91.22 ± 0.49	$\underline{86.97} \pm 0.13$	95.36 ± 0.11	90.22
	RoLoRA	87.08 ± 0.34	90.86 ± 0.26	85.72 ± 0.10	96.27 ± 0.08	89.98
	FedCoDi	$\mathbf{87.97} \pm 0.23$	$\mathbf{92.75} \pm 0.61$	$\mathbf{88.19} \pm 0.15$	$\mathbf{97.12} \pm 0.24$	**91.51**
50	FedIT	87.26 ± 0.33	89.97 ± 0.53	83.46 ± 0.34	94.97 ± 0.67	88.92
	FFA-LoRA	84.31 ± 0.46	88.96 ± 0.52	81.62 ± 2.04	94.17 ± 0.61	87.27
	FedSA-LoRA	$\underline{88.06} \pm 0.23$	91.06 ± 0.36	80.17 ± 0.21	$\underline{95.70} \pm 0.46$	88.75
	LoRA-FAIR	87.22 ± 0.38	89.82 ± 0.62	84.33 ± 0.19	94.40 ± 0.47	88.94
	FedEx-LoRA	87.46 ± 0.21	$\underline{91.18} \pm 0.42$	$\underline{84.73} \pm 0.33$	94.88 ± 0.56	$\underline{89.56}$
	RoLoRA	86.28 ± 0.27	89.31 ± 0.63	84.56 ± 0.14	94.05 ± 0.69	88.55
	FedCoDi	$\mathbf{88.60} \pm 0.26$	$\mathbf{91.72} \pm 0.27$	$\mathbf{86.48} \pm 0.41$	$\mathbf{96.21} \pm 0.34$	**90.75**

et al. (2025), and RoLoRA Chen et al. (2025). All methods use rank r=8, and adapter dropout 0.05 at the query and value projections of all Transformer layers, with Gaussian initialization for A and zero initialization for B. We set $\tau = 10$, $\theta = 0.3$, and $\lambda = 0.05$. For each setting, the learning rate is selected by validation performance from $\{5 \times 10^{-4}, 1 \times 10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 2 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-1}\}$; all reported hyperparameters are the best validation-selected configurations from this grid search.

GLUE. We evaluate RoBERTa-large Liu et al. (2019) on four tasks from GLUE Wang et al. (2019): MNLI, QNLI, QQP, and SST-2. Data are partitioned across clients using a Dirichlet distribution with $\alpha = 0.5$. We consider 3 and 10 clients with full participation, and 50 clients with 10% sampled per round. All GLUE experiments run for 1000 federated rounds.

GSM8K. We evaluate Meta-Llama-3-8B Dubey et al. (2024) on GSM8K Cobbe et al. (2021) with three clients under an IID partition and a 10% holdout for model selection. All compared methods use 100 rounds with full participation, 30 local steps per round, batch size 1, and learning rate 5×10^{-3}. Evaluation uses 8-shot chain-of-thought prompting, greedy decoding, and exact-match accuracy. The checkpoint is selected by validation loss and evaluated once on the test set.

Dolly-15K. We also evaluate Llama-3-8B on Dolly-15K Conover et al. (2023) for instruction following. We use a stratified $90/10$ split and partition the training portion across three clients using a Dirichlet distribution with $\alpha = 0.5$ over the eight task categories. The adapter and training settings match those used for GSM8K. We evaluate the final-round checkpoint on the held-out split using greedy decoding with at most 256 new tokens and ROUGE-L F1 Lin (2004).

5.2 Main Results

GLUE performance. FedCoDi achieves the highest mean accuracy in all 12 task–client configurations in Table 1. Its average accuracies are 91.68%, 91.51%, and 90.75% with 3, 10, and 50 clients, respectively. The corresponding improvements over the strongest baseline at each client scale are 0.70, 1.09, and 1.19 percentage points. In the 50-client setting, for example, QQP accuracy increases from 84.73% with FedEx-LoRA to 86.48% with FedCoDi. These results show benefits at all evaluated client scales, including partial participation.

Reasoning and instruction following. Table 2 reports the Llama-3-8B results. On GSM8K, FedCoDi achieves 50.11% exact-match accuracy, exceeding FFA-LoRA (46.93%) by 3.18 percentage points and FedIT (46.10%) by 4.01 points. On Dolly-15K, FedCoDi obtains a ROUGE-L F1 score of 0.3162, compared with 0.3105 for FedIT, 0.3047 for FedSA-LoRA, and 0.2906 for FFA-LoRA.

Table 2: Results on GSM8K and Dolly-15K (Dirichlet α=0.5) with Llama-3-8B and 3 clients, using the IID GSM8K partition.

Method	GSM8K (acc)	Dolly (ROUGE-L)
FedIT	46.10	<u>0.3105</u>
FFA-LoRA	<u>46.93</u>	0.2906
FedSA-LoRA	46.47	0.3047
FedCoDi	**50.11**	**0.3162**

Table 3: Ablation on MNLI (3 clients, Dirichlet α=0.5), toggling DDCA and DCR on top of the personalized decoupled backbone.

Variant	DDCA	DCR	Acc. (%)
Shared magnitude	×	×	88.3
Local magnitude	×	×	88.9
+ DCR	×	✓	89.6
+ DDCA	✓	×	<u>89.9</u>
+ DDCA + DCR (**FedCoDi**)	✓	✓	**90.2**

Table 4: System efficiency comparison: trainable parameters, per-round communicated parameters (client→server), time per round, and rounds to reach the target on MNLI and QQP.

	# Trainable Param.	# Per-round Commu. Param.	Time per Round (s)		Rounds to Target	
			MNLI	QQP	MNLI	QQP
FedIT	1.83M	0.78M	33.24	30.09	138	179
FFA-LoRA	1.44M	0.39M	30.13	28.31	187	195
FedSA-LoRA	1.83M	0.39M	32.61	29.71	108	168
LoRA-FAIR	1.83M	0.78M	35.32	31.26	145	183
FedEx-LoRA	1.83M	0.78M	41.74	36.53	132	174
RoLoRA	1.44M	0.39M	31.58	28.67	141	189
FedCoDi	1.88M	0.78M	35.57	30.92	**68**	**132**

The improvement over the strongest compared baseline is 0.0057. These results extend the observed performance gains to a decoder-only backbone under both the IID GSM8K partition and the non-IID Dolly-15K partition.

5.3 Ablation Study

Table 3 examines magnitude retention, DDCA, and DCR on MNLI. All variants use the same decoupled parameterization. Without DDCA or DCR, retaining the magnitude vector locally improves accuracy from 88.3% to 88.9% compared with aggregating it. Starting from this local-magnitude variant, DCR alone raises accuracy to 89.6%, while DDCA alone reaches 89.9%. Combining both mechanisms yields 90.2%, an improvement of 1.3 percentage points over the local-magnitude variant. FedCoDi also exceeds DDCA alone by 0.3 points and DCR alone by 0.6 points. This ablation supports using both mechanisms in addition to local magnitude retention in the evaluated setting.

5.4 System Efficiency

Table 4 reports parameter counts, client-to-server uplink payload, per-round time, and rounds to a common target on MNLI and QQP with three clients under Dirichlet $\alpha = 0.5$. The target is 95% of the strongest baseline accuracy in Table 1. FedCoDi reaches it in 68 rounds on MNLI and 132 rounds on QQP, the fewest among the compared methods. It communicates 0.78 million parameters per round, matching FedIT, and therefore reduces cumulative uplink volume relative to FedIT by 50.7% and 26.3%, respectively; these counts follow the table's client-to-server scope and exclude downlink communication. Its per-round time is 7.0% higher on MNLI and 2.8% higher on QQP than FedIT, but the lower round count yields the lowest estimated time to the target among the compared methods. The communication advantages depend on each baseline's per-round payload.

6 Conclusion

We characterize *directional cancellation* in federated low-rank adaptation under non-IID data and relate it to the aggregated product up to a conditional second-order mismatch (Proposition 1, Lemma 1). Building on this analysis, **FedCoDi** combines row-wise global-local factor mixing (DDCA) with product-level directional regularization during local training (DCR). Experiments on GLUE, GSM8K, and Dolly-15K show consistent gains over the evaluated federated PEFT baselines.

ETHICS STATEMENT

This work does not involve human subjects or the release of new datasets. FedCoDi enables personalized fine-tuning of foundation models without centralizing private data, lowering the barrier for resource-constrained participants; at the same time, federated systems remain susceptible to poisoning attacks and performance disparities across heterogeneous clients, and models fine-tuned on private data could be misused, so deployments should incorporate Byzantine-robust aggregation Blanchard et al. (2017), anomaly detection, and appropriate safeguards.

REPRODUCIBILITY STATEMENT

Complete proofs of all theoretical results are provided in Appendix A, and the coherence-quality monotonicity assumption underlying Proposition 2 is explicitly stated and discussed in Appendix A.8. The experimental setup—datasets, data partitioning protocols, adapter configurations, training budgets, hyperparameters, and evaluation protocols—is fully described in Section 5, and all methods share identical configurations and budgets to ensure fair comparison.

AI USE STATEMENT

In this work, we used generative AI tools for assistance with language polishing of the manuscript. We have not used generative AI tools for generating research ideas, experimental results, or core technical claims. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Sara Babakniya, Ahmed Roushdy Elkordy, Yahya H Ezzeldin, Qingfeng Liu, Kee-Bong Song, Mostafa El-Khamy, and Salman Avestimehr. SLoRA: Federated parameter efficient fine-tuning of language models. In *NeurIPS 2023 Workshop on Federated Learning in the Age of Foundation Models*, 2023.

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. FlexLoRA: Federated fine-tuning of large language models under heterogeneous tasks and client resources. In *Advances in Neural Information Processing Systems*, 2024.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. LoRA-FAIR: Federated LoRA fine-tuning with aggregation and initialization refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

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

Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. In *International Conference on Learning Representations*, 2022.

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*, 2025.

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 (EMNLP)*, 2024.

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. *arXiv preprint arXiv:2110.14168*, 2021.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pp. 2089–2099. PMLR, 2021.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free Dolly: Introducing the world's first truly open instruction-tuned LLM. https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. FedDC: Federated learning with non-IID data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10112–10121, 2022.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. FedSA-LoRA: 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.

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

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*, pp. 429–450, 2020.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning*, pp. 32100–32121, 2024.

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. *arXiv preprint arXiv:1907.11692*, 2019.

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

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, 2025.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving LoRA in privacy-preserving federated learning. In *International Conference on Learning Representations*, 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 *International Conference on Learning Representations*, 2019.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, 2020.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6915–6919, 2024. doi: 10.1109/ICASSP48485.2024.10447454.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-IID data. *arXiv preprint arXiv:1806.00582*, 2018.

A Detailed Proofs and Discussion

This appendix provides the proofs of the main-text results, a supplementary lower bound for directional cancellation, and discussion of the analysis assumptions and the DCR reference.

A.1 Caveats on the Cancellation Verification

Probe configuration. We evaluate the factor-averaging baseline on MNLI with three clients under Dirichlet $\alpha \in \{0.5, 1.0\}$, using rank 8, learning rate 2×10^{-2}, 50 local steps, and a fixed seed. For each probed output row, we reconstruct $u_j^{(i)} = B_{j,:}^{(i)} A^{(i)}$.

Assumptions behind the reference curve. The expression in Theorem 1(c) assumes both equal aggregation weights and equal, nonzero update norms. For general weights and nonzero update rows, the exact ratio instead satisfies

$$\kappa_j^2 = \frac{\sum_i w_i^2 \|u_j^{(i)}\|_2^2 + 2 \sum_{i<k} w_i w_k \rho_{ik}^{(j)} \|u_j^{(i)}\|_2 \|u_j^{(k)}\|_2}{\left(\sum_i w_i \|u_j^{(i)}\|_2 \right)^2}, \tag{16}$$

provided the denominator is positive. Thus mean pairwise coherence alone does not generally determine the measured ratio. The equal-weight, equal-norm curve in Figure 1 is a reference for the reported probes, not an identity for arbitrary client updates. Differences in update norms do not imply a fixed sign of the discrepancy from this curve. The reported MAE of 0.004 describes the measurements in this experiment rather than a general error bound.

Interpretation of the cancellation ratio. The ratio measures the aggregate norm relative to the weighted sum of client update norms. A lower value does not, by itself, imply higher task loss or lower accuracy. Directional differences may also encode client-specific adaptations. Accordingly, FedCoDi uses selective factor mixing on the incoming path (Section 4.1) and a soft directional penalty on the outgoing path (Section 4.2), rather than maximizing this ratio or enforcing identical client directions.

A.2 Proof of Theorem 1

Part (a): Closed-form expansion. For a fixed output row j, expand the squared norm:

$$\begin{aligned} \|u_j^g\|_2^2 &= \left\langle \sum_i w_i u_j^{(i)}, \sum_k w_k u_j^{(k)} \right\rangle \\ &= \sum_i w_i^2 \|u_j^{(i)}\|_2^2 + 2 \sum_{i<k} w_i w_k \langle u_j^{(i)}, u_j^{(k)} \rangle. \end{aligned} \tag{17}$$

Substituting $\langle u_j^{(i)}, u_j^{(k)} \rangle = \rho_{ik}^{(j)} \|u_j^{(i)}\|_2 \|u_j^{(k)}\|_2$ gives the claimed expression. The inner-product identity also holds for zero update rows, without evaluating their undefined cosines.

Part (b): Norm deficit relative to aligned updates. Adding and subtracting the cross terms with coherence equal to one yields

$$\|u_j^g\|_2^2 = \left(\sum_i w_i \|u_j^{(i)}\|_2\right)^2 - 2\sum_{i<k} w_i w_k \big(1 - \rho_{ik}^{(j)}\big)\|u_j^{(i)}\|_2 \|u_j^{(k)}\|_2. \tag{18}$$

Since $\rho_{ik}^{(j)} \in [-1, 1]$, every term in the subtracted sum is nonnegative, which proves the upper bound. Equality holds precisely when all positively weighted nonzero update rows have the same direction. This comparison fixes the client update norms and aggregation weights; the reference is the fully aligned configuration.

Part (c): Equal-weight, equal-norm formula. For $N \geq 2$, $w_i = 1/N$, and $\|u_j^{(i)}\|_2 = c > 0$, part (a) gives

$$\|u_j^g\|_2^2 = \frac{c^2}{N^2}\left(N + 2\sum_{i<k} \rho_{ik}^{(j)}\right) = \frac{c^2}{N}\left(1 + (N-1)\bar{\rho}^{(j)}\right). \tag{19}$$

Taking square roots proves the formula. Nonnegativity implies $\bar{\rho}^{(j)} \geq -1/(N-1)$, with equality if and only if $\sum_i u_j^{(i)} = 0$. A regular-simplex configuration is one way to attain equality when the ambient dimension permits, but it is not the only zero-sum configuration.

Supplementary lower bound under equal norms. Let $\rho_{\min}^{(j)} = \min_{i<k} \rho_{ik}^{(j)}$ and $\|u_j^{(i)}\|_2 = c > 0$. Since $w_i w_k \geq 0$, part (a) implies

$$\begin{aligned}
\|u_j^g\|_2^2 &= c^2 \sum_i w_i^2 + 2c^2 \sum_{i<k} w_i w_k \rho_{ik}^{(j)} \\
&\geq c^2 \left(\sum_i w_i^2 + 2\rho_{\min}^{(j)} \sum_{i<k} w_i w_k\right) \\
&= c^2 \left((1 - \rho_{\min}^{(j)}) \sum_i w_i^2 + \rho_{\min}^{(j)}\right),
\end{aligned} \tag{20}$$

where the last step uses $\sum_i w_i = 1$. Equality holds when every positively weighted pair attains $\rho_{\min}^{(j)}$. For equal weights, the right-hand side becomes $\frac{c^2}{N}(1 + (N-1)\rho_{\min}^{(j)})$. For fixed weights, the bound is nondecreasing in $\rho_{\min}^{(j)}$, since its derivative is $c^2(1 - \sum_i w_i^2) \geq 0$. A lower minimum coherence therefore weakens this guarantee on the retained norm; it does not establish a corresponding reduction in the actual norm. If the right-hand side is negative, the trivial lower bound of zero is stronger. In particular, a zero lower bound does not imply that complete cancellation occurs: complete cancellation in the equal-weight, equal-norm setting is characterized by the mean-coherence condition in part (c), not by the minimum pairwise coherence alone.

A.3 Proof of Proposition 1

Using the definitions of the averaged factors and $\sum_i w_i = 1$:

$$\begin{aligned}
\mathcal{E}_j &= \sum_i w_i B_{j,:}^{(i)} A^{(i)} - \Big(\sum_i w_i B_{j,:}^{(i)}\Big) A^g - B_{j,:}^g \Big(\sum_i w_i A^{(i)}\Big) + B_{j,:}^g A^g \\
&= u_j^g - B_{j,:}^g A^g = u_j^g - u_j^{\text{agg}}.
\end{aligned} \tag{21}$$

Rearranging gives the identity. The triangle inequality and submultiplicativity give

$$\|\mathcal{E}_j\|_2 \leq \sum_i w_i \big\|B_{j,:}^{(i)} - B_{j,:}^g\big\|_2 \big\|A^{(i)} - A^g\big\|_F. \tag{22}$$

If these deviations are at most $C_B\epsilon$ and $C_A\epsilon$ for all clients, with constants independent of ϵ, then $\|\mathcal{E}_j\|_2 \leq C_B C_A \epsilon^2$. The reverse triangle inequality further yields

$$\big|\|u_j^{\text{agg}}\|_2 - \|u_j^g\|_2\big| \leq \|u_j^{\text{agg}} - u_j^g\|_2 = \|\mathcal{E}_j\|_2.$$

This is an absolute perturbation bound conditional on the stated factor deviations. It does not establish that those deviations are small during training or that the relative error is small near a zero averaged update.

A.4 Proof of Lemma 1 and the Coherence Perturbation Bound

The decomposition

$$B^{(i)}A^{(i)} - B^gA^g = (B^{(i)} - B^g)A^{(i)} + B^g(A^{(i)} - A^g)$$

and the triangle inequality imply

$$\begin{aligned}\|B^{(i)}A^{(i)} - B^gA^g\|_F &\le \|B^{(i)} - B^g\|_F \, \|A^{(i)}\|_F + \|B^g\|_F \, \|A^{(i)} - A^g\|_F \\ &= \Delta_B\alpha + \beta\Delta_A. \end{aligned} \tag{23}$$

For the angular bound, let $P = \mathrm{vec}(B^gA^g)$ and $Q = \mathrm{vec}(B^{(i)}A^{(i)})$, with $\|P\|_2 \ge c_0 > 0$ and $\|Q - P\|_2 \le \delta < c_0$. Then $\|Q\|_2 \ge \|P\|_2 - \delta > 0$, and

$$\langle Q, P\rangle = \|P\|_2^2 + \langle Q - P, P\rangle \ge \|P\|_2(\|P\|_2 - \delta) > 0.$$

Together with $\|Q\|_2 \le \|P\|_2 + \delta$, this gives

$$\cos(Q, P) \ge \frac{\|P\|_2 - \delta}{\|P\|_2 + \delta} \ge \frac{c_0 - \delta}{c_0 + \delta}, \tag{24}$$

since $(x-\delta)/(x+\delta)$ is nondecreasing for $x > \delta$. The nonzero-reference and relative-error conditions are essential to this argument; an absolute deviation bound alone is insufficient.

A.5 Proof of Proposition 2

Part (a): Pointwise bound. Since $0 \le G_j \le 1$, the triangle inequality gives

$$\|B_{\mathrm{aligned},j,:} - B^*_{j,:}\|_2 = \|G_j(B^g_{j,:} - B^*_{j,:}) + (1 - G_j)(B^{(i)}_{j,:} - B^*_{j,:})\|_2 \le G_j e^g_j + (1 - G_j)e^l_j.$$

Part (b): Comparison with fixed-weight bounds. Assume the coherence–quality condition and positive margin $m = \min_j |\rho_j - \theta| > 0$ from the proposition. The finite-sharpness comparison also requires strict preferences on both sides of the threshold: some $j_H \in \mathcal{H}$ satisfies $e^g_{j_H} < e^l_{j_H}$, and some $j_L \in \mathcal{L}$ satisfies $e^l_{j_L} < e^g_{j_L}$, where $\mathcal{H} = \{j : \rho_j > \theta\}$ and $\mathcal{L} = \{j : \rho_j < \theta\}$.

Let $s = \sigma(\tau m) \in (1/2, 1)$, and define the squared DDCA bound $U_\tau = \sum_j (G_j e^g_j + (1 - G_j)e^l_j)^2$. On $\mathcal{H}$, $G_j \ge s$ and $e^g_j \le e^l_j$; on $\mathcal{L}$, $G_j \le 1 - s$ and $e^l_j \le e^g_j$. Consequently,

$$\begin{aligned} U_\tau &\le H(s), \\ H(s) &= \sum_{j\in\mathcal{H}} \big(se^g_j + (1 - s)e^l_j\big)^2 + \sum_{j\in\mathcal{L}} \big((1 - s)e^g_j + se^l_j\big)^2. \end{aligned} \tag{25}$$

As $\tau \to \infty$, $s \to 1$ and

$$H(s) \longrightarrow E_\star := \sum_j \min\big\{(e^g_j)^2, (e^l_j)^2\big\}.$$

For a fixed coefficient $\alpha \in [0, 1]$, let $F(\alpha) = \sum_j (\alpha e^g_j + (1 - \alpha)e^l_j)^2$. Every summand is at least the smaller squared error, so $F(\alpha) \ge E_\star$. At row j_H, equality would require $\alpha = 1$; at row j_L, equality would require $\alpha = 0$. Hence $F(\alpha) > E_\star$ for every $\alpha \in [0, 1]$. Because F is continuous on this compact interval,

$$\gamma := \min_{\alpha\in[0,1]} \big(F(\alpha) - E_\star\big) > 0.$$

The convergence of $H(\sigma(\tau m))$ to $E_\star$ therefore gives a finite τ_0 such that, for all $\tau \ge \tau_0$,

$$U_\tau \le H(\sigma(\tau m)) < E_\star + \gamma \le F(\alpha) \qquad \text{for every } \alpha \in [0, 1].$$

This proves the comparison with one threshold τ_0 that applies uniformly to all fixed mixing coefficients. It does not imply that the default τ in the experiments exceeds this threshold.

Part (c): Sharp-gating limit. The positive margin implies $G_j \to \mathbb{I}[\rho_j > \theta]$. Under the coherence–quality condition, each row then selects the global or local candidate with the smaller reference error, so $U_\tau \to E_\star$. This limit requires the monotonicity and margin conditions, but not the two-sided strictness used in part (b). It is optimal among row-wise choices between the two candidates, not among all possible interpolations, and it compares error bounds rather than realized task losses.

A.6 Proof of Proposition 3

Let $\mathcal{L}_{\text{task}} > 0$, and hold λ and $\mathcal{L}_{\text{DCR}}$ fixed. For the additive objective,

$$R_{\text{add}} = \frac{\lambda \mathcal{L}_{\text{DCR}}}{\mathcal{L}_{\text{task}}}.$$

Replacing $\mathcal{L}_{\text{task}}$ by $K\mathcal{L}_{\text{task}}$, with $K > 0$, divides this ratio by K. For the multiplicative objective, the additional loss is $\lambda \mathcal{L}_{\text{task}} \mathcal{L}_{\text{DCR}}$, so

$$R_{\text{mul}} = \frac{\lambda \mathcal{L}_{\text{task}} \mathcal{L}_{\text{DCR}}}{\mathcal{L}_{\text{task}}} = \lambda \mathcal{L}_{\text{DCR}},$$

which is unchanged by the same rescaling. For trainable parameters ϑ, differentiating the objective gives

$$\nabla_\vartheta \mathcal{L}_{\text{total}} = (1 + \lambda \mathcal{L}_{\text{DCR}}) \nabla_\vartheta \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{task}} \nabla_\vartheta \mathcal{L}_{\text{DCR}}.$$

Thus the proposition establishes loss-ratio invariance under positive rescaling, not identical optimization dynamics across different tasks, nor a guarantee that λ never requires tuning.

A.7 Design Rationale for the DCR Reference

Shared reference and personalized initialization. DCR uses $\mathcal{W}^g = B^g A^g$, formed from the unblended broadcast factors and held fixed within a round. Clients receiving the same broadcast therefore share the same reference. In contrast, the DDCA-initialized product $B_{\text{aligned}} A_{\text{aligned}}$ depends on each client's retained factors. Using that product as the reference would regularize clients toward different personalized initializations; the unblended product separates the shared reference from the client-specific starting point.

Complementary roles. DDCA mixes the rows of each factor before local training, whereas DCR compares the current low-rank product with the fixed reference during training. Local magnitudes remain trainable and are not communicated. This provides selective initialization and continued directional guidance while retaining local adaptation. DCR is a soft cosine penalty; it does not impose a hard directional cone or guarantee monotonic improvement in pairwise client coherence.

A.8 Discussion of the Monotonicity Assumption in Proposition 2

The coherence–quality condition assumes

$$\rho_j > \theta \Rightarrow e_j^g \leq e_j^l, \qquad \rho_j < \theta \Rightarrow e_j^l \leq e_j^g.$$

These errors are distances to the client-specific row reference $B^*_{j,:}$ in Proposition 2, not distances to a population optimum. The condition is a modeling assumption linking the observed coherence score to the relative quality of two candidate rows; it does not follow from cosine similarity or federated averaging alone.

The intended intuition is that global–local agreement can indicate reusable shared structure, while disagreement may indicate a local adaptation worth retaining. Neither interpretation is guaranteed: coherence alone does not establish which candidate is closer to $B^*_{j,:}$. The finite-sharpness comparison additionally requires a positive threshold margin and strict preferences on both sides, as stated in the proposition. Note also that coherence is measured against the global direction B^g, whose quality can itself be reduced by directional cancellation (Theorem 1); the assumption should therefore be read as a working hypothesis about the *relative* ordering of e_j^g and e_j^l rather than an absolute quality certificate.

DDCA does not compute $B^*_{j,:}$ or the reference errors during training. A direct empirical test of this assumption would therefore require an explicit procedure for approximating the row reference and measuring the corresponding errors. The ablation in Table 3 evaluates the benefit of DDCA but does not directly validate this row-wise assumption. The proposition should be read as a conditional analysis of the gate, not an unconditional certificate of its selection quality.

Algorithm 1 FedCoDi Training Algorithm

Server Input: Initial A^0, B^0, learning rate η, rounds T.
Client Input: Local magnitude $M^{(i)}$, data $\mathcal{D}_i$, hyperparameters τ, θ, λ.
Default hyperparameters: $\tau = 10$, $\theta = 0.3$, $\lambda = 0.05$.
for round $t = 1$ **to** T **do**
  Select $\mathcal{S}_t$ and normalized aggregation weights $\{w_i^t\}_{i \in \mathcal{S}_t}$.
  Server broadcasts A^{t-1}, B^{t-1} to clients in $\mathcal{S}_t$.
  for each client $i \in \mathcal{S}_t$ in parallel **do**
    Cache $\mathcal{W}^g \leftarrow B^{t-1}A^{t-1}$ and hold it fixed.
    if $t > 1$ **then**
      Retrieve the retained local factors $A^{(i)}, B^{(i)}$.
      Apply row-wise DDCA to the broadcast and local factors.
      Load $A_{\text{aligned}}, B_{\text{aligned}}$ as $A^{(i)}, B^{(i)}$.
    else
      Initialize $A^{(i)} \leftarrow A^0, B^{(i)} \leftarrow B^0$.
    end if
    for local update step $s = 1$ **to** S **do**
      Compute the task loss on a local batch from $\mathcal{D}_i$.
      Compute $\mathcal{L}_{\text{DCR}}$ using the current local product and the cached unblended reference.
      $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{task}}(1 + \lambda\mathcal{L}_{\text{DCR}})$.
      Update $A^{(i)}, B^{(i)}, M^{(i)}$, and the head when present, using SGD with learning rate η.
    end for
    Retain the updated local factors and magnitude for later rounds.
    Upload only $A^{(i)}, B^{(i)}$ to the server.
  end for
  $A^t \leftarrow \sum_{i \in \mathcal{S}_t} w_i^t A^{(i)}$, $B^t \leftarrow \sum_{i \in \mathcal{S}_t} w_i^t B^{(i)}$.
end for

B Training Algorithm

Algorithm 1 summarizes the client-side operations and factor-wise aggregation. Let $\mathcal{S}_t$ denote the clients selected in round t, with $w_i^t \geq 0$ and $\sum_{i \in \mathcal{S}_t} w_i^t = 1$. The cached reference is formed before DDCA and remains fixed throughout local training. The client retains its trained factor state for subsequent DDCA steps, while the magnitude vector is updated locally and never included in the uploaded payload.