ON THE CONSISTENCY OF FEDERATED LORA UNDER SECURE AGGREGATION VIA FUNCTIONAL SKETCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated low-rank adaptation (LoRA) with independently trained two-factor adapters faces two coupled aggregation inconsistencies. Independently averaging factors generally differs from averaging their products, while additive masking before local factor multiplication introduces non-canceling cross terms. We propose SAGE-LORA, a functional-sketch framework that applies secure aggregation (SecAgg) to linear measurements of locally paired products rather than factor coordinates. Budget-aware paired blocks support heterogeneous local ranks, while fixed-width sketches decouple per-client measurement size from local rank and cohort size. An anchored right sketch identifies a cohort-level range, and a range-locked left sketch from the same surviving cohort enables fixed-rank reconstruction. Clients evaluate both sketches through factor chains before masking, without materializing dense updates. We establish exact measurement aggregation and sequential aggregate confidentiality under the universal SecAgg assumptions, and derive reconstruction and time-averaged stationarity bounds that distinguish approximation from optimization error. Across eight commonsense tasks with Qwen2.5-1.5B-Instruct and rank-8 adapters, SAGE-LORA with SecAgg achieves 76.37% mean accuracy, within 0.66 percentage points of unprotected FlexLoRA and over 5 points above unprotected FLoRA and FSLoRA. Under the evaluated representations and precisions, logical communication is 67.3% and 60.1% lower than FLoRA and FSLoRA, respectively, including second-stage projection traffic. The paired mean accuracy change from plaintext is only 0.10 percentage points. On five GLUE tasks, plaintext SAGE-LORA improves the macro score over FLoRA, FSLoRA, and FlexLoRA by 0.42–2.40 points. These results support SAGE-LORA as a practical exploration of SecAgg for federated LoRA that reconciles product consistency, aggregate confidentiality, and compact communication.

1 INTRODUCTION

Federated fine-tuning enables a population of clients to adapt a shared pretrained model while keeping raw examples local. For large language models (LLMs), full-model fine-tuning is often infeasible on end devices because parameter, gradient, activation, and optimizer states exceed local memory and energy budgets (Brown et al., 2020; Xu et al., 2024). Parameter-efficient fine-tuning (PEFT) reduces this burden by freezing the backbone and optimizing compact auxiliary parameters (Houlsby et al., 2019; Lester et al., 2021). Low-rank adaptation (LoRA) is particularly suitable because it writes an adapted weight as $W_0 + BA$, where $B \in \mathbb{R}^{d^o \times r}$ and $A \in \mathbb{R}^{r \times d^i}$ with $r \ll \min(d^o, d^i)$ (Hu et al., 2021). Since the trainable state scales as $r(d^o + d^i)$, each client can select a rank that matches its computation, memory, and communication budget (Cho et al., 2024; Fang et al., 2025). This resource adaptivity is central to heterogeneous on-device learning, although it removes the common factor layout assumed by conventional federated averaging.

These constraints already arise in deployed model customization. Google has trained Gboard language models across heterogeneous phones and has reported production configurations in which secure aggregation exposes only cohort-level aggregates (Bonawitz et al., 2019; Xu et al., 2023; Bonawitz et al., 2017). OpenAI reports that Indeed and SK Telecom used fine-tuning for personalized job recommendations and telecommunications customer service, showing that production adaptation relies on proprietary behavior signals and domain knowledge (OpenAI, 2024). Microsoft's disclosure

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.

of an AI-development storage exposure, together with recent reconstruction attacks on transformer and adapter updates, further shows that centralized infrastructure and communicated updates cannot be assumed confidential (Microsoft Security Response Center, 2023; Petrov et al., 2024; Chen et al., 2026). These deployments make heterogeneous adaptation and aggregate-only visibility simultaneous requirements, which in LoRA expose two distinct inconsistencies at the aggregation interface.

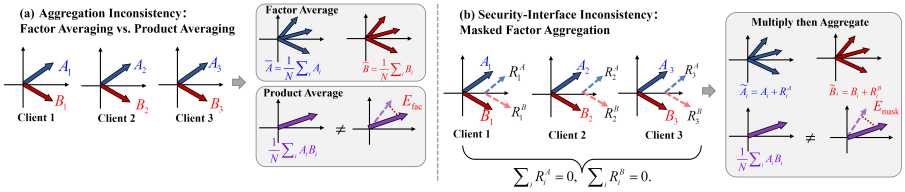

Figure 1: Two Inconsistencies in Secure Federated LoRA Aggregation.

As illustrated in Figure 1, the first difficulty is an *aggregation inconsistency*. For a common factor layout, let client i produce $D_i = B_i A_i$ with normalized data weight α_i. Direct FedLoRA, as instantiated by FedIT (Zhang et al., 2024), averages the factors and then multiplies them, whereas product-space model averaging aggregates the client products. Their difference is

$$\begin{aligned} E_{\text{fac}} &= \Big(\sum_i \alpha_i B_i\Big)\Big(\sum_j \alpha_j A_j\Big) - \sum_i \alpha_i B_i A_i \\ &= \sum_{i\neq j} \alpha_i \alpha_j B_i A_j + \sum_i (\alpha_i^2 - \alpha_i) B_i A_i. \end{aligned} \tag{1}$$

Equation (1) measures deviation from the weighted product-average target. This deviation can persist at a common rank because LoRA factors are non-identifiable, with $BA = (B\mathcal{T})(\mathcal{T}^{-1}A)$ for any invertible $\mathcal{T}$. Unequal ranks additionally require a shared embedding for factor averaging. Freezing a common factor, as in FFA-LoRA, makes aggregation linear but restricts updates to its fixed row or column subspace (Sun et al., 2024). Our setting retains two-factor local adaptation, allowing each client to adjust both update subspaces within its rank budget.

The second difficulty is a *security-interface inconsistency*. Additive secure aggregation (SecAgg) protects individual contributions while revealing their sum (Bonawitz et al., 2017; Liu et al., 2022). Its cancellation property must be preserved when choosing the aggregation representation. For dimensionally compatible factors, suppose additive masks satisfy $\sum_i R_i^A = \sum_i R_i^B = 0$. Applying these masks before local factor multiplication gives

$$\begin{aligned} \sum_i (B_i + R_i^B)(A_i + R_i^A) &= \sum_i B_i A_i + E_{\text{mask}}, \\ E_{\text{mask}} &= \sum_i \big(B_i R_i^A + R_i^B A_i + R_i^B R_i^A\big), \end{aligned} \tag{2}$$

which is generally nonzero. Equation (2) shows that additive cancellation does not carry through the bilinear factor map. A product-consistent SecAgg interface should therefore evaluate linear measurements of each locally paired product before applying masks. This preserves the additive aggregation structure without communicating dense product matrices.

Existing methods address different aspects of federated LoRA aggregation. FLoRA preserves product sums by stacking factors, whereas FlexLoRA compresses the product average through truncated SVD (Wang et al., 2024; Bai et al., 2024). FSLoRA adapts to heterogeneous resources through sparse updates of shared factors and supports their secure aggregation (Fang et al., 2025). FedRot-LoRA aligns factors before averaging, while Fed-SB makes aggregation linear by training a core within fixed outer subspaces (Zhang et al., 2026; Singhal et al., 2025). Privacy-preserving approaches further investigate secure aggregation and selective encryption (Bossy et al., 2026; Liu et al., 2026). FedASK uses two-stage product sketching to address the trade-off between DP noise amplification and adapter learnability (Wen et al., 2025). SAGE-LoRA instead focuses on adapting additive SecAgg to independently trained, unequal-rank LoRA products. It combines public-adapter anchoring, common measurement dimensions, and two secure aggregates bound to the same participating cohort, so

that reconstruction operates on a consistent product average while individual contributions remain protected under the stated SecAgg assumptions.

To this end, we propose SAGE-LORA, a functional-sketch framework for product-consistent secure aggregation in heterogeneous federated LoRA. Its central design separates local bilinear adaptation from protected linear aggregation. Each client trains budget-feasible paired blocks and computes $D_i\Omega$ and ΨD_i through the factor chains $B_i(A_i\Omega)$ and $(\Psi B_i)A_i$, without materializing D_i. The right projection incorporates the global adapter's public factor subspace and probes complementary directions. The first secure aggregate identifies a cohort-level range, from which the server constructs the left projection. The same retained clients compute the second measurement from unchanged local products and weights, enabling fixed-rank reconstruction from two consistent aggregates. Applying masks after measurement formation keeps all protected aggregation operations additive. The interface supports both complete-adapter and residual updates, with local ranks controlling learning capacity and public sketch widths controlling measurement size.

Our contributions are as follows.

- **Product-consistent SecAgg for heterogeneous LoRA.** We design a fixed-width measurement interface that preserves local factor pairing while supporting independently trained, unequal-rank adapters. By moving the protected aggregation to linear product measurements, SAGE-LORA avoids factor-averaging mismatch and pre-product mask cross terms without requiring dense client updates.
- **Anchored reconstruction with explicit guarantees.** We combine public-adapter anchoring with cohort-bound, range-locked measurements for fixed-rank reconstruction. We establish exact measurement aggregation, bound missed-range energy and rank-truncation error, and derive a time-averaged stationarity bound with update-mode-specific direction estimates. A sequential security analysis establishes aggregate confidentiality under the stated SecAgg assumptions.
- **Competitive utility with quantified communication savings.** On eight commonsense tasks with rank-8 adapters, SAGE-LORA with SecAgg reaches 76.37% mean accuracy, within 0.66 percentage points of unprotected FlexLoRA and over 5 points above unprotected FLoRA and FSLoRA. Including the second-stage projection, logical communication is reduced by 67.3% and 60.1% relative to the latter two protocols under the evaluated representations and precisions. The paired mean accuracy change from plaintext is 0.10 percentage points. GLUE experiments and matched ablations further assess transfer utility and reconstruction fidelity.

2 PROBLEM BACKGROUND

2.1 FEDERATED LORA OBJECTIVE AND SYSTEM MODEL

We consider N clients with private datasets $\{\mathcal{D}_i\}_{i=1}^N$ and one coordinating server. For an adapted layer with frozen weight $W_0 \in \mathbb{R}^{d^o \times d^i}$, the population objective is

$$\min_{M \in \mathbb{R}^{d^o \times d^i}} F(M), \qquad F(M) = \sum_{i=1}^{N} p_i F_i(M), \qquad F_i(M) = \mathbb{E}_{\xi \sim \mathcal{D}_i}[\ell(W_0 + M; \xi)], \tag{3}$$

where $p_i \geq 0$ and $\sum_i p_i = 1$, following the standard federated objective (McMahan et al., 2017). At round t, the server represents the rank-k global adapter as $M^t = B^t A^t$, where $B^t \in \mathbb{R}^{d^o \times k}$ and $A^t \in \mathbb{R}^{k \times d^i}$. It also specifies a pool $\mathcal{G} = \{1, \ldots, G\}$ of paired LoRA blocks. Index $j \in \mathcal{G}$ identifies a block of width r_j, with $B_{i,j}^t \in \mathbb{R}^{d^o \times r_j}$ and $A_{i,j}^t \in \mathbb{R}^{r_j \times d^i}$. Client i activates $G_i^t \in \{0, \ldots, G\}$ blocks according to its resource budget and obtains

$$D_i^t = \sum_{j=1}^{G_i^t} B_{i,j}^t A_{i,j}^t, \qquad r_i^t = \sum_{j=1}^{G_i^t} r_j, \qquad \operatorname{rank}(D_i^t) \leq r_i^t. \tag{4}$$

Here D_i^t denotes the uploaded paired product, not necessarily a residual. Let $\mathcal{U}_t$ be the retained clients, n_i their protected sample counts, and $\alpha_i^t = n_i / \sum_{u \in \mathcal{U}_t} n_u$. With a public base P^t, the ideal

target is

$$S^t = P^t + \eta_{\mathrm{srv}} \bar{D}^t, \qquad \bar{D}^t = \sum_{i \in \mathcal{U}_t} \alpha_i^t D_i^t. \tag{5}$$

The *complete-update* mode uses $P^t = 0$, $\eta_{\mathrm{srv}} = 1$, and $D_i^t = C_i^t$, a complete client adapter initialized from the global state; its product average replaces M^t. The *residual* mode uses $P^t = M^t$ and trains fresh D_i^t on top of frozen M^t. The modes share the measurement interface, not the local objective. Population weights p_i and realized weights α_i^t may differ. Throughout, $\|\cdot\|$ is Euclidean for vectors and Frobenius for matrices, and $[X]_k$ is a best rank-k approximation. Appendix B compares aggregation rules against the common product sum $\bar{D}$.

2.2 Threat Model and Secure Aggregation

We adopt the semi-honest model used in SHE-LoRA (Liu et al., 2026) and rely on the selected secure aggregation protocol for its declared collusion and dropout thresholds (Bonawitz et al., 2017; Liu et al., 2022). The adversary observes public model states, projection matrices, protocol metadata, and the released cohort aggregates, but should learn no additional information about an honest client's data, LoRA factors, local rank, block selection, or individual sketches.

SAGE-LoRA uses sequential secure aggregations because the second projection depends on the first aggregate. The first stage fixes the survivor set $\mathcal{U}_t$; the same clients supply the left sketch. The formal protocol requires Stage II to authenticate that set and its protected count before releasing an aggregate, or abort with fresh retry state. This binding has zero tolerance to additional dropout after Stage I: if any accepted client fails to complete Stage II, the round produces no model update. Per-stage dropout recovery therefore does not remove the inter-stage availability constraint, whose effect on round-completion rates is not measured here. Appendix E separates this requirement from the evaluated reference implementation. The guarantee is conditional aggregate confidentiality, not differential privacy; malicious deviations, side channels, and released-model leakage are outside its scope.

3 SAGE-LoRA

The method combines budget-aware local adaptation, causal product-sketch aggregation, and stable global reconstruction. As illustrated in Figure 2, each participating client selects and trains a budget-feasible set of paired LoRA blocks, and then evaluates a right product sketch without materializing the dense local update. The first secure aggregation identifies the cohort-level range, from which the server derives and distributes the range-locked projection Ψ^t. The same survivor set subsequently contributes the complementary left sketches, allowing the server to reconstruct balanced rank-k global factors for the next communication round. This workflow preserves heterogeneous local LoRA capacity while keeping the communicated measurements fixed-dimensional and product-consistent.

3.1 Budget-Aware Paired-Block Local Adaptation

Given profiled block costs $c_{i,j}^{\mathrm{tr}}$ and local budget C_i^{tr}, client i activates the largest feasible prefix

$$G_i^t = \max\left\{ G' \in \{0, \dots, G\} : \sum_{j=1}^{G'} c_{i,j}^{\mathrm{tr}} \le C_i^{\mathrm{tr}} \right\}, \qquad r_i^t = \sum_{j=1}^{G_i^t} r_j. \tag{6}$$

Writing $\widehat{F}_i$ for the empirical loss on $\mathcal{D}_i$, the client trains the active block pairs by

$$\min_{\{B_{i,j}^t, A_{i,j}^t\}_{j \le G_i^t}} \widehat{F}_i\left(P^t + \sum_{j=1}^{G_i^t} B_{i,j}^t A_{i,j}^t \right). \tag{7}$$

For complete updates, the client uses the explicit rank-feasible initializer in Equation (20): it retains the first $\min\{r_i^t, k\}$ components of the balanced global factors and assigns any extra slots zero initial product, without a local SVD. Residual updates instead optimize fresh blocks with M^t frozen. Keeping each $B_{i,j}$ and $A_{i,j}$ paired preserves its product without common factor coordinates. The complete round procedure is given in Algorithm 2 in Appendix G.

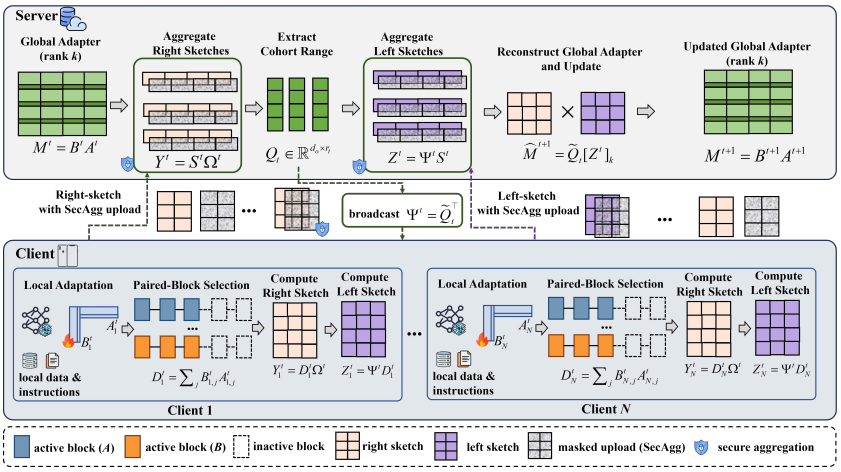

Figure 2: Workflow of SAGE-LORA. The evaluated reconstruction uses the Stage-I-range decoder in Equation (14).

3.2 CAUSAL RANGE-LOCKED PRODUCT SKETCHING

We write the sketch equations in the chosen layer orientation. For nonzero $M^t = B^t A^t$, let V^t be an orthonormal basis, publicly completed to k columns, for the oriented public factor subspace $\mathrm{range}((A^t)^\top)$, which contains $\mathrm{row}(M^t)$. Transposition exchanges the original row and column roles. The anchor uses M^t, not P^t; with a zero global product the probe is purely random. Otherwise,

$$\Omega^t = [V^t, V_\perp^t] \in \mathbb{R}^{d^i \times s_\mathrm{R}}, \qquad V_\perp^t \in \mathbb{R}^{d^i \times (s_\mathrm{R}-k)}, \qquad k \le s_\mathrm{R} \le d^i. \tag{8}$$

Each participating client receives Ω^t and evaluates the product sketch through its low-rank factors,

$$Y_i^t = D_i^t \Omega^t = \sum_{j=1}^{G_i^t} B_{i,j}^t (A_{i,j}^t \Omega^t), \tag{9}$$

$$Y^t = P^t \Omega^t + \eta_\mathrm{srv} \sum_{i \in \mathcal{U}_t} \alpha_i^t Y_i^t = S^t \Omega^t. \tag{10}$$

Thus, the dense matrix D_i^t is never formed, and secure aggregation preserves the local B–A pairing. From Y^t, the server computes a fixed-width thin-QR basis $Q_t \in \mathbb{R}^{d^o \times s_\mathrm{R}}$. The thresholded effective rank $\widehat{r}_t \le s_\mathrm{R}$ is a numerical diagnostic. If $\widehat{r}_t < k$, the server aborts before Stage II and retries under a fresh round identifier and probe; all reconstruction statements condition on an accepted round with $\widehat{r}_t \ge k$. The server completes Q_t with $H_t \in \mathbb{R}^{d^o \times (s_\mathrm{L} - s_\mathrm{R})}$ to $\widetilde{Q}_t = [Q_t, H_t] \in \mathbb{R}^{d^o \times s_\mathrm{L}}$, where $s_\mathrm{R} \le s_\mathrm{L} \le d^o$ (Halko et al., 2011), and defines

$$\Psi^t = [Q_t, H_t]^\top = \widetilde{Q}_t^\top, \qquad G_t := \Psi^t Q_t = \begin{bmatrix} I_{s_\mathrm{R}} \\ 0_{(s_\mathrm{L}-s_\mathrm{R}) \times s_\mathrm{R}} \end{bmatrix}. \tag{11}$$

The minimal functional choice is $s_\mathrm{L} = s_\mathrm{R}$, for which H_t is empty. When $s_\mathrm{L} > s_\mathrm{R}$, the completion rows retain the evaluated bilateral payload geometry but are compatibility measurements: the exact decoder below discards them, so they are not a claimed source of reconstruction or utility gain. The same survivor set evaluates and securely aggregates the left sketches

$$Z_i^t = \Psi^t D_i^t = \sum_{j=1}^{G_i^t} (\Psi^t B_{i,j}^t) A_{i,j}^t, \tag{12}$$

$$Z^t = \Psi^t P^t + \eta_\mathrm{srv} \sum_{i \in \mathcal{U}_t} \alpha_i^t Z_i^t = \widetilde{Q}_t^\top S^t. \tag{13}$$

The first secure aggregation identifies the cohort-level range, whereas the second recovers the coordinates of the same product average in that range. This causal dependence is the reason that SAGE-LORA uses two sequential secure aggregations rather than one non-adaptive aggregation.

3.3 Stable Global Adapter Reconstruction

The range lock makes G_t column-orthonormal, so its pseudoinverse stably selects the coordinates associated with the first-stage range,

$$X^t = G_t^\dagger Z^t = Q_t^\top S^t, \qquad \widehat{M}^{t+1} = Q_t[X^t]_k = [Q_t Q_t^\top S^t]_k. \tag{14}$$

If $X^t = U_X \Sigma_X V_X^\top$, a balanced factorization is $B^{t+1} = Q_t U_{X,k} \Sigma_{X,k}^{1/2}$ and $A^{t+1} = \Sigma_{X,k}^{1/2} V_{X,k}^\top$. The server sets $M^{t+1} = B^{t+1} A^{t+1}$ using the reduced $X^t \in \mathbb{R}^{s_\mathrm{R} \times d^i}$. For fixed Q_t, $G_t^\dagger = [I_{s_\mathrm{R}}, 0_{s_\mathrm{R} \times (s_\mathrm{L} - s_\mathrm{R})}]$ discards all H_t rows: increasing s_L alone cannot improve this exact decoder. Those rows remain in the evaluated transport and its reported cost, not in the returned adapter.

Efficiency. For a chosen orientation of an $m \times n$ layer, sketch uploads contain $ms_\mathrm{R} + s_\mathrm{L} n$ scalars; transposition exchanges m, n in both upload and projection downlink. One orientation must govern both directions (Appendix A). Upload shape is independent of client rank and cohort size. Consequently, for a very small r_i^t, the sketch can exceed the raw-factor payload whenever $ms_\mathrm{R} + s_\mathrm{L} n > (m+n) r_i^t$; fixed width bounds the server interface rather than guaranteeing a smaller uplink for every client. Factor-chain evaluation costs $O(r_i^t (s_\mathrm{R} + s_\mathrm{L})(m+n))$ without a dense update. Sage adds a second synchronization and a projection downlink; avoiding dense products does not imply a speed advantage over factor-space FlexLoRA (Bai et al., 2024). Reported native bytes also reflect FP32 adapter endpoints versus BF16 sketches, and are not a common-precision or packet-level comparison. The byte reductions therefore combine representation and datatype effects; Table 5 reports scalar counts, while an equal-precision payload ablation remains unmeasured.

4 Theoretical Analysis

We retain only the main optimization consequence here. Structural correctness, reconstruction error, heterogeneity specializations, and optimization proofs are deferred to Appendix C, while the formal confidentiality definition and proof are given in Appendix E. With the same survivor set in both secure aggregations, linearity gives

$$Y^t = S^t \Omega^t, \qquad Z^t = \widetilde{Q}_t^\top S^t, \tag{15}$$

so neither the factor-aggregation term E_fac nor the factor-mask term E_mask enters the reconstructed update.

Let $x^t = \mathrm{vec}(M^t)$, identify F with its vectorized representation, and write one round as

$$x^{t+1} = x^t - \eta g^t + e^t, \tag{16}$$

where $g^t = \mathrm{vec}(M^t - \widetilde{S}^t)/\eta$ and $e^t = \mathrm{vec}(\widehat{M}^{t+1} - \widetilde{S}^t)$. Here $\widetilde{S}^t = S^t$ unless optional block sampling is used. This identity covers both modes; the assumptions below constrain, rather than establish, the quality of that direction.

Assumption 4.1 (Smooth objective)**.** *F is lower bounded by $F_\star$ and L-smooth in the product variable.*

Assumption 4.2 (Federated direction)**.** *For history $\mathcal{F}_t$, let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$. With deterministic envelopes β_t, σ_t, assume*

$$\mathbb{E}_t[g^t] = \nabla F(x^t) + b^t, \quad \|b^t\|^2 \le \beta_t^2, \quad \mathbb{E}_t \|g^t - \mathbb{E}_t g^t\|^2 \le \sigma_t^2. \tag{17}$$

Assumption 4.3 (Bounded reconstruction error)**.** *For a deterministic envelope ζ_t,*

$$\mathbb{E}_t \|e^t\|^2 \le \eta^2 \zeta_t^2. \tag{18}$$

The η^2 scaling follows the normalization $\widetilde{S}^t - M^t = -\eta g^t$: bounded relative reconstruction gives $e^t = O(\eta)$. It remains an assumption for complete-update training. Here β_t, σ_t bound direction bias and variance, while ζ_t bounds reconstruction distortion. Appendix C.9 gives separate residual- and complete-update envelopes; the latter permits displacement rank $r_i^t + k$.

Table 1: Eight-task commonsense accuracy (%, homogeneous rank 8). Each cell gives the mean above its ± sample SD over three independent seeds; Avg. is the within-seed task macro-average. Bold denotes the largest mean in each column.

Method	ARC-c	ARC-e	BoolQ	HellaS.	OBQA	PIQA	SIQA	WinoG.	Avg.
FedIT	75.47 ±0.19	87.84 ±0.27	**78.21** ±0.08	75.03 ±0.21	**75.73** ±1.47	79.16 ±0.34	73.27 ±0.34	69.56 ±0.45	76.78 ±0.33
Zero-padding	75.47 ±0.19	87.84 ±0.27	**78.21** ±0.08	75.03 ±0.21	**75.73** ±1.47	79.16 ±0.34	73.27 ±0.34	69.56 ±0.45	76.78 ±0.33
FLoRA	75.47 ±0.39	87.43 ±0.10	71.07 ±0.35	59.88 ±0.63	68.07 ±0.50	75.88 ±0.08	70.32 ±0.32	60.17 ±0.12	71.03 ±0.03
FlexLoRA	75.03 ±0.84	**88.71** ±0.27	78.04 ±0.35	**76.32** ±0.21	75.33 ±1.15	79.13 ±0.64	**73.47** ±0.58	**70.17** ±0.75	**77.03** ±0.14
FS-LoRA	**75.70** ±0.84	87.49 ±0.20	71.52 ±0.18	60.09 ±0.31	68.07 ±0.42	76.26 ±0.06	70.13 ±0.19	61.43 ±0.28	71.34 ±0.09
FedASK	73.24 ±0.33	87.72 ±0.46	77.48 ±0.44	73.84 ±0.36	73.27 ±0.31	77.73 ±0.42	73.39 ±0.18	68.48 ±0.73	75.64 ±0.40
FedRot-LoRA	74.47 ±0.51	88.30 ±0.27	78.17 ±0.40	75.20 ±0.42	75.07 ±1.22	**79.27** ±0.44	73.10 ±0.33	69.48 ±0.75	76.63 ±0.30
SAGE-LORA	74.02 ±0.84	88.30 ±0.54	77.65 ±0.24	74.55 ±0.76	74.00 ±1.11	78.45 ±0.29	73.27 ±0.03	69.88 ±0.79	76.27 ±0.41

Theorem 4.4 (Time-averaged stationarity of SAGE-LORA)**.** *Under Assumptions 4.1–4.3, let* $0 < \eta \le 1/(8L)$ *and define* $\bar{\beta}_T^2 = T^{-1}\sum_{t=0}^{T-1}\beta_t^2$, $\bar{\sigma}_T^2 = T^{-1}\sum_{t=0}^{T-1}\sigma_t^2$, *and* $\bar{\zeta}_T^2 = T^{-1}\sum_{t=0}^{T-1}\zeta_t^2$. *Then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x^t)\|^2 \le \frac{4(F(x^0)-F_\star)}{\eta T} + 8\bar{\beta}_T^2 + 4L\eta\bar{\sigma}_T^2 + 4(1+L\eta)\bar{\zeta}_T^2. \tag{19}$$

This bounds the time average of expected squared gradient norms, not last-iterate or parameter convergence. If $\beta_t = \zeta_t = 0$, variance is bounded, and $\eta = \Theta(T^{-1/2})$, that average is $O(T^{-1/2})$. Persistent bias or distortion leaves a nonzero upper-bound floor. A separate PL corollary bounds the expected objective gap (Appendix C.3); confidentiality remains conditional on the SecAgg assumptions in Appendix E.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate Qwen2.5-1.5B-Instruct (Yang et al., 2024) on the eight-task commonsense suite used by FSLoRA (Fang et al., 2025) and five GLUE tasks (Wang et al., 2019). We report equal-task mean accuracy for commonsense reasoning and the mean of the primary GLUE scores: Matthews correlation for CoLA, accuracy for SST-2, QNLI and RTE, and mean accuracy/F1 for MRPC. All evaluations use the complete validation sets.

Training and implementation. Experiments used heterogeneous NVIDIA GPUs, including RTX A6000, L20, and A40. All methods use 50 clients, 10 participants per round, 100 rounds, 20 local optimizer steps, learning rate 2×10^{-4}, and effective batch size 16. We train BF16 LoRA on `q_proj`/`v_proj` with a frozen backbone, length 512, and no quantization. Baselines are FedIT, zero-padding, FLoRA, FlexLoRA, FSLoRA random-k, FedASK, and FedRot-LoRA; within-cohort partitions and participant schedules are shared, with method-specific global ranks. Sage uses complete-adapter updates for homogeneous ranks and profile-specific residual adaptation for heterogeneous ranks. GLUE classifier heads use sample-weighted FedAvg. Unless noted, results are means and sample standard deviations over three independent random seeds, averaging tasks within seeds. Full settings and accounting appear in Appendices F.1 and F.4.

5.2 TASK UTILITY AND COMMUNICATION

Commonsense accuracy. SAGE-LoRA achieves 76.27%, improving over FLoRA, FSLoRA, and FedASK by 5.24, 4.93, and 0.63 percentage points (Table 1). FlexLoRA and FedRot-LoRA remain competitive, with SAGE-LoRA within 0.76 points of both. Rather than retaining stacked or larger shared factors, SAGE-LoRA uses the public adapter to guide range estimation and reconstructs the aggregated products as a rank-8 adapter. The results show that this compact representation retains

Table 2: GLUE-5 transfer subset (validation): three-seed mean and sample SD, multiplied by 100. CoLA uses MCC, MRPC uses (accuracy + F1)/2, and other tasks use accuracy; Avg. averages these five primary metrics within each seed. Bold marks the highest column mean.

Method	CoLA	SST-2	MRPC	QNLI	RTE	Avg.
FedIT	$57.01_{\pm 1.67}$	$95.53_{\pm 0.72}$	$84.90_{\pm 0.92}$	$\mathbf{93.27}_{\pm 0.35}$	$81.47_{\pm 0.91}$	$82.44_{\pm 0.58}$
Zero-padding	$\mathbf{57.08}_{\pm 1.68}$	$95.72_{\pm 0.70}$	$\mathbf{86.12}_{\pm 0.93}$	$93.14_{\pm 0.41}$	$\mathbf{81.71}_{\pm 2.08}$	$\mathbf{82.75}_{\pm 0.64}$
FLoRA	$52.95_{\pm 2.88}$	$94.15_{\pm 0.30}$	$81.55_{\pm 1.08}$	$89.63_{\pm 0.08}$	$75.81_{\pm 2.60}$	$78.82_{\pm 0.61}$
FlexLoRA	$51.39_{\pm 1.99}$	$95.49_{\pm 0.18}$	$83.71_{\pm 0.75}$	$93.00_{\pm 0.11}$	$80.39_{\pm 1.63}$	$80.79_{\pm 0.79}$
FS-LoRA	$51.77_{\pm 0.39}$	$94.42_{\pm 0.18}$	$84.26_{\pm 0.48}$	$90.23_{\pm 0.19}$	$78.58_{\pm 1.10}$	$79.85_{\pm 0.12}$
FedASK	$53.75_{\pm 2.14}$	$94.27_{\pm 0.46}$	$83.68_{\pm 1.03}$	$91.58_{\pm 0.66}$	$80.14_{\pm 1.30}$	$80.68_{\pm 1.12}$
FedRot-LoRA	$54.87_{\pm 2.43}$	$\mathbf{95.80}_{\pm 0.53}$	$85.62_{\pm 0.81}$	$93.15_{\pm 0.31}$	$80.39_{\pm 1.78}$	$81.96_{\pm 0.66}$
SAGE-LORA	$53.34_{\pm 1.98}$	$95.45_{\pm 0.24}$	$84.29_{\pm 1.60}$	$92.85_{\pm 0.27}$	$80.14_{\pm 1.65}$	$81.22_{\pm 1.01}$

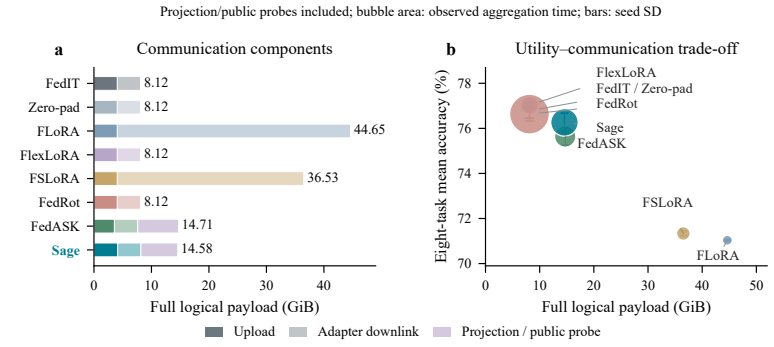

Figure 3: Homogeneous-rank utility and full logical traffic over 100 rounds. (a) Upload, adapter download, and required projection. (b) Three-seed accuracy with SD; bubble area encodes observed aggregation time on mixed hardware. FedIT and zero-padding coincide.

useful client updates, providing an effective learning component for the linear SecAgg interface in Equation (15).

Transfer performance. SAGE-LoRA obtains 81.22 on GLUE-5, exceeding FLoRA, FSLoRA, and FedASK by 2.40, 1.37, and 0.54 points (Table 2). Its gains over FedASK are largest on QNLI (1.27 points) and SST-2 (1.18). FedIT and zero-padding remain stronger overall, so the benefit is competitive transfer rather than universal accuracy leadership. All methods use the same head-aggregation rule; SAGE-LoRA adapts each task through its compact product reconstruction without changing that rule. This supports reuse of the aggregation interface across independently trained classification objectives.

Communication overhead. SAGE-LoRA uses 14.58 GiB, reducing full logical traffic by 67.3% and 60.1% relative to FLoRA and FSLoRA (Figure 3). The main saving is downlink: FLoRA broadcasts stacked factors and FSLoRA distributes rank-64 factors, whereas SAGE-LoRA returns rank eight, independent of total client rank. Projection traffic is included. FedASK has a similar total (14.71 GiB), while compact factor baselines use 8.12 GiB. Thus, fixed-rank reconstruction controls the communication cost of combining client products, although these native-byte savings also reflect transport precision (Tables 5 and 11).

5.3 Heterogeneity and Mechanism Ablations

Heterogeneous ranks. SAGE-LoRA reaches 74.46% and 73.45% under balanced and heavy-tail ranks, improving over FedASK by approximately 0.60 and 0.95 points (Table 8). It remains 0.97 points below balanced FlexLoRA and 2.64 below heavy-tail FedRot-LoRA. Budget-feasible paired blocks let clients train at different ranks, while common sketch widths keep their messages

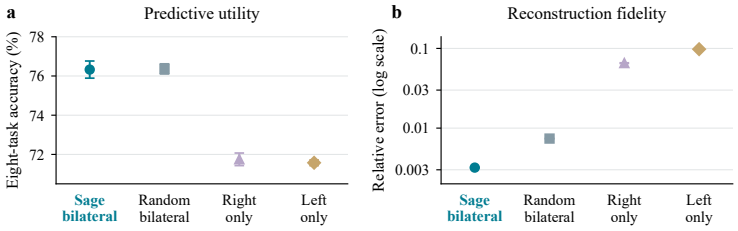

Figure 4: Matched interventions at fixed upload and returned rank. Points show three-seed means ± SD (some error bars are smaller than markers). Bilateral measurements improve (a) accuracy and (b) reconstruction over one-sided controls; adaptation improves fidelity, not accuracy, over random bilateral measurement.

compatible with secure summation. These results support resource-dependent local training without rank-dependent message shapes. The remaining gaps are consistent with the separate local-capacity and reconstruction limits analyzed in Appendix C.9.

Reconstruction quality. At fixed upload and returned rank, bilateral SAGE-LoRA achieves 76.33%, versus 71.75% for right-only and 71.57% for left-only measurement (Figure 4 and Table 13). The first sketch identifies an aggregate range; the second supplies its coefficients. Compared with random bilateral measurement, adaptation reduces excess reconstruction error from 0.00442 to 0.00021, while accuracy is nearly unchanged (76.36% versus 76.33%). The benefit is therefore more faithful reconstruction within the same rank budget, consistent with the range and truncation terms in Equation (36), rather than higher output capacity.

5.4 Security–Utility Trade-offs

Utility under SecAgg. Secure SAGE-LoRA achieves 76.37%, exceeding FedASK-DP by 0.73 points and remaining within 0.10 points of selective CKKS (Table 14). Its paired change from plaintext is only $+0.10 \pm 0.32$ points. Clients compute product sketches before masking, so SecAgg sums linear measurements without introducing bilinear mask terms. This explains why protection can retain the plaintext learning behavior and supports Equation (15). SecAgg and DP provide different guarantees; implementation costs and the reference's security scope are detailed in Appendices E.2 and F.6.

6 Conclusion

SAGE-LoRA adapts additive SecAgg to federated LoRA through fixed-width linear measurements of locally paired products. Public-adapter anchoring and cohort-bound, range-locked aggregation support fixed-rank reconstruction while preserving heterogeneous local capacity. The analysis separates measurement consistency, reconstruction error, optimization effects, and aggregate confidentiality. Experiments show competitive secured utility, a small paired accuracy change from plaintext, and reduced logical communication relative to the evaluated FLoRA and FSLoRA protocols. Together, these results support functional sketching as a practical approach to integrating SecAgg with flexible federated LoRA adaptation.

7 Limitations

SAGE-LoRA aborts a round if any Stage-I survivor fails to complete Stage II. Fixed-width sketches may exceed raw-factor uploads at very low ranks. Performance trails strong baselines in the evaluated heterogeneous-rank settings, indicating that aggregation compatibility alone does not ensure effective adaptation.

AI USE STATEMENT

Generative AI tools assisted with the literature review, language polishing, and guidance on implementing experimental code. The authors are responsible for verifying the scientific content and all assertions made.

ETHICS STATEMENT

This work studies privacy-preserving aggregation of federated model updates and does not involve new human-subject data collection. Any reconstruction attack should be used only for defensive evaluation on appropriately licensed or approved data.

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. *Advances in Neural Information Processing Systems*, 37:14457–14483, 2024.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017.

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečný, Stefano Mazzocchi, Brendan McMahan, et al. Towards federated learning at scale: System design. *Proceedings of machine learning and systems*, 1: 374–388, 2019.

Thierry Bossy, Julien Vignoud, Tahseen Rabbani, Juan R. Troncoso Pastoriza, and Martin Jaggi. Mitigating unintended memorization with LoRA in federated learning for LLMs. *Transactions on Machine Learning Research*, 2026.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Silong Chen, Yuchuan Luo, Guilin Deng, Yi Liu, Ming Xu, Shaojing Fu, and Xiaohua Jia. Reconstructing training data from adapter-based federated large language models. In *Proceedings of the ACM Web Conference 2026*, pp. 2602–2613, 2026.

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*, pp. 12903–12913, 2024.

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. *arXiv preprint arXiv:2501.19389*, 2025.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2): 217–288, 2011.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 3045–3059, 2021.

Jianmin Liu, Li Yan, Borui Li, Lei Yu, and Chao Shen. She-lora: Selective homomorphic encryption for federated tuning with heterogeneous lora. In *International Conference on Learning Representations*, volume 2026, pp. 154701–154733, 2026.

Zizhen Liu, Si Chen, Jing Ye, Junfeng Fan, Huawei Li, and Xiaowei Li. Sash: Efficient secure aggregation based on shprg for federated learning. In *Uncertainty in artificial intelligence*, pp. 1243–1252. PMLR, 2022.

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.

Microsoft Security Response Center. Microsoft mitigated exposure of internal information in a storage account due to an overly permissive SAS token. `https://www.microsoft.com/en-us/msrc/blog/2023/09/microsoft-mitigated-exposure-of-internal-information-in-a-storage-account-due-to-overly-permissive-sas-token`, 2023. Published September 18, 2023.

OpenAI. Introducing improvements to the fine-tuning API and expanding our custom models program. `https://openai.com/index/introducing-improvements-to-the-fine-tuning-api-and-expanding-our-custom-models-program/`, 2024. Published April 4, 2024.

Ivo Petrov, Dimitar I Dimitrov, Maximilian Baader, Mark N Müller, and Martin Vechev. DAGER: Exact gradient inversion for large language models. *Advances in neural information processing systems*, 37:87801–87830, 2024.

Raghav Singhal, Kaustubh Ponkshe, Rohit Vartak, Lav R Varshney, and Praneeth Vepakomma. Fed-sb: A silver bullet for extreme communication efficiency and performance in (private) federated lora fine-tuning. *arXiv preprint arXiv:2502.15436*, 2025.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving lora in privacy-preserving federated learning. In *International conference on learning representations*, volume 2024, pp. 17978–17994, 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.

Shuai Wang, Youliang Tian, Jinbo Xiong, Renwan Bi, Jianfeng Ma, and Yan Zhang. Pridfl: Computation-optimized secure aggregation with byzantine-resilient in decentralized federated learning. *IEEE Transactions on Dependable and Secure Computing*, 2025.

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

Ming Wen, Jiaqi Zhu, Yuedong Xu, Yipeng Zhou, and Dingding Han. Differentially private federated low rank adaptation beyond fixed-matrix. In *Advances in Neural Information Processing Systems*, volume 38, pp. 114755–114786, 2025.

Mengwei Xu, Dongqi Cai, Yaozong Wu, Xiang Li, and Shangguang Wang. FwdLLM: Efficient federated finetuning of large language models with perturbed inferences. In *2024 USENIX Annual Technical Conference (USENIX ATC 24)*, pp. 579–596, 2024.

Zheng Xu, Yanxiang Zhang, Galen Andrew, Christopher Choquette, Peter Kairouz, Brendan McMahan, Jesse Rosenstock, and Yuanbo Zhang. Federated learning of gboard language models with differential privacy. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 629–639, 2023.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Haoran Zhang, Dongjun Kim, Seohyeon Cha, and Haris Vikalo. Fedrot-lora: Mitigating rotational misalignment in federated lora. *arXiv preprint arXiv:2602.23638*, 2026.

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

REPRODUCIBILITY STATEMENT

This appendix provides the aggregation rules, proofs, complexity analysis, numerical requirements, additional measurements, and model-scale protocol.

A COMPUTATION, MEMORY, AND COMMUNICATION COMPARISON

We represent client computation and memory, server aggregation overhead, and communication separately (Zhang et al., 2024; Wang et al., 2024; Bai et al., 2024; Fang et al., 2025; Zhang et al., 2026; Singhal et al., 2025). Let $m = d^o$, $n = d^i$, $c = |\mathcal{U}_t|$, H be the number of local optimization steps, r_i the local rank of client i, $r_\Sigma = \sum_{u\in\mathcal{U}_t} r_u$ the cumulative participating rank, r a common or maximum baseline rank, and $\mathsf{sz}(r) = (m+n)r$ the number of scalars in a rank-r LoRA adapter. Let $\mathcal{M}_0$ denote the common memory required by the frozen backbone and its training activations. The tables omit costs shared by all methods and constant-size protocol metadata. Memory and computation use the direct orientation; transposition exchanges m, n. For communication, a single fixed orientation $\tau \in \{0, 1\}$ sets $(m_\tau, n_\tau) = (m, n)$ or (n, m) for both directions, with $s_{\mathrm{R}} \le n_\tau$ and $s_{\mathrm{R}} \le s_{\mathrm{L}} \le m_\tau$. Server-memory entries follow a materialized-update convention; streaming can reduce some costs.

Client cost. Table 3 reports adapter-related memory and operations in addition to the common backbone computation. For SAGE-LORA, $\mathcal{W}_{\mathrm{sk}}$ denotes the sketch workspace, bounded by $m_\tau s_{\mathrm{R}} + s_{\mathrm{L}} n_\tau$ scalars for the selected orientation; an implementation can stream it blockwise.

Table 3: Client memory and additional computation per adapted layer.

Method	Adapter-related memory	Computation per round
FedIT	$\mathcal{M}_0 + \mathsf{sz}(r)$	$O(Hr(m+n))$
FLoRA	$\mathcal{M}_0 + \max\{\mathsf{sz}(r_\Sigma), mn\}$	$O(Hr_i(m+n) + r_\Sigma mn + mn)$
FlexLoRA	$\mathcal{M}_0 + \mathsf{sz}(r_i)$	$O(Hr_i(m+n))$
FSLoRA	$\mathcal{M}_0 + \mathsf{sz}(r_i)$	$O(Hr_i(m+n))$
FedRot-LoRA	$\mathcal{M}_0 + \mathsf{sz}(r)$	$O(Hr(m+n) + (m+n)r^2 + r^3)$
Fed-SB	$\mathcal{M}_0 + \mathsf{sz}(r) + r^2$	$O(H(r(m+n) + r^2))$
SAGE-LORA	$\mathcal{M}_0 + \mathsf{sz}(k) + \mathsf{sz}(r_i) + \mathcal{W}_{\mathrm{sk}}$	$O(H(k+r_i)(m+n) + r_i(s_{\mathrm{R}} + s_{\mathrm{L}})(m+n) + nks_{\mathrm{R}})$

FLoRA may avoid storing the full stacked adapter by merging it into a dense layer, shifting the cost to dense client computation and storage. The SAGE-LORA entry describes residual training: it counts both the frozen rank-k global adapter and trained rank-r_i blocks. Complete-update training instead fine-tunes the local adapter, with training term $O(Hr_i(m+n))$; retaining the global factors for probe regeneration still fits the displayed memory upper bound. Both modes add two factor-chain sketch evaluations and regenerate the public probe without forming dense D_i.

Server cost. The SAGE-LORA server cost depends on the public sketch widths rather than r_Σ and does not include an SVD of an $m \times n$ product. The stated bound covers low-rank global-adapter products, range extraction, orthogonal completion, and the reduced decomposition; secure aggregation decoding is linear in the sketch size. Regenerating the right probe from a public seed costs $O(nks_{\mathrm{R}})$ in the direct orientation; broadcasting it instead exchanges this computation for ns_{R} additional downlink scalars.

Table 4: Server-side aggregation memory and additional high-cost computation per adapted layer.

Method	Aggregation memory	Additional server computation
FedIT	$O(c\mathsf{sz}(r))$	$O(c\mathsf{sz}(r))$
FLoRA	$O(\sum_i \mathsf{sz}(r_i))$	$O(\sum_i \mathsf{sz}(r_i))$
FlexLoRA	$O(\max\{\sum_i \mathsf{sz}(r_i), 2mn\})$	$O(r_\Sigma mn + cmn + mn\min\{m,n\})$
FSLoRA	$O(\sum_i \mathsf{sz}(r_i))$	$O(c\mathsf{sz}(r))$
FedRot-LoRA	$O(c\mathsf{sz}(r))$	$O(c\mathsf{sz}(r))$ after client alignment
Fed-SB	$O(cr^2 + \mathsf{sz}(r))$	$O(cr^2)$
SAGE-LORA	$O(ms_\mathrm{R} + s_\mathrm{L}n + ms_\mathrm{L} + \mathsf{sz}(k))$	$O((m+n)k(s_\mathrm{R}+s_\mathrm{L}) + ms_\mathrm{R}^2 + (m+n)s_\mathrm{L}^2)$

Table 5: Model-parameter communication per round and adapted layer. Scalar counts are reported unless bits are stated explicitly. The right projection is reconstructed from the global adapter and a public seed; constant-size metadata and cryptographic setup traffic are omitted.

Method	Client uplink	Client downlink	Dependence highlighted
FedIT	$\mathsf{sz}(r)$	$\mathsf{sz}(r)$	common rank
FLoRA	$\mathsf{sz}(r_i)$	$\min\{\mathsf{sz}(r_\Sigma), mn\}$	cohort rank sum
FlexLoRA	$\mathsf{sz}(r_i)$	$\mathsf{sz}(r_i)$	dense server product/SVD
FSLoRA	$\mathsf{sz}(r_i)$	$\mathsf{sz}(r)$ scalars $+ O(r)$ bits	local rank in uplink
FedRot-LoRA	$\mathsf{sz}(r)$	$\mathsf{sz}(r)$	common rank
Fed-SB	r^2	r^2	shared outer subspace
SAGE-LORA	$m_\tau s_\mathrm{R} + s_\mathrm{L}n_\tau$	$\mathsf{sz}(k) + m_\tau s_\mathrm{L}$	one fixed orientation τ

Communication. The two feasible communication pairs are $(ms_\mathrm{R} + s_\mathrm{L}n, \mathsf{sz}(k) + ms_\mathrm{L})$ and $(ns_\mathrm{R} + s_\mathrm{L}m, \mathsf{sz}(k) + ns_\mathrm{L})$. Their uplink and downlink minima need not use the same orientation and cannot be combined. If both orientations are feasible, minimizing their sum gives $\mathsf{sz}(k) + (m+n)s_\mathrm{L} + s_\mathrm{R}\min\{m,n\}$. This scalar-count comparison is separate from native byte accounting: the evaluated adapter endpoints and projection use FP32, while plaintext sketch values use BF16. Reported byte savings therefore reflect both representation and precision; the secure reference's encoded channels and packet traffic have different scopes.

Rank-feasible complete-update initialization. Because the server returns balanced factors ordered by decreasing singular value, client i can set $\check{r}_i^t = \min\{r_i^t, k\}$ and initialize

$$B_{i,0}^t = \left[B_{:,1:\check{r}_i^t}^t,\ 0_{d^o \times (r_i^t - \check{r}_i^t)}\right], \qquad A_{i,0}^t = \begin{bmatrix} A_{1:\check{r}_i^t,:}^t \\ R_i^t \end{bmatrix}, \tag{20}$$

where $R_i^t \in \mathbb{R}^{(r_i^t - \check{r}_i^t)\times d^i}$ is a seeded LoRA initialization. Then $B_{i,0}^t A_{i,0}^t = [M^t]_{\check{r}_i^t}$; extra capacity has zero initial product but receives gradients through R_i^t. The concatenated factors are partitioned across the active blocks, so no client-side SVD is required.

B Representative Federated-LoRA Aggregation Rules and Failure Modes

This appendix makes the aggregation object of representative baselines explicit. Let client i return $B_i \in \mathbb{R}^{m\times r_i}$, $A_i \in \mathbb{R}^{r_i \times n}$, product $D_i = B_i A_i$, and normalized weight α_i. For paired blocks, $B_i = [B_{i,1}, \ldots, B_{i,G_i}]$ and $A_i = [A_{i,1}^\top, \ldots, A_{i,G_i}^\top]^\top$, so $B_i A_i = \sum_j B_{i,j} A_{i,j}$. The common product reference is $\bar{D} = \sum_i \alpha_i D_i$, giving target $S = P + \eta_\mathrm{srv}\bar{D}$ when a public base is added. Here D_i may be a complete adapter or a fresh residual; their training objectives remain distinct.

B.1 FEDIT / DIRECT FEDLORA

For a common rank r, FedIT averages the factors independently:

$$\bar{B} = \sum_i \alpha_i B_i, \qquad \bar{A} = \sum_i \alpha_i A_i, \qquad D_{\mathrm{FedIT}} = \bar{B}\bar{A}. \tag{21}$$

This rule preserves rank r and communicates $(m+n)r$ scalars, but it does not generally equal $\bar{D}$ (Zhang et al., 2024).

B.2 FROZEN-FACTOR LINEARIZATION

If all clients share a fixed A_0 and train only B_i, then

$$\sum_i \alpha_i B_i A_0 = \left(\sum_i \alpha_i B_i\right) A_0. \tag{22}$$

Thus FFA-LoRA removes both the factor-average discrepancy and the need for a bilinear server operation (Sun et al., 2024). The trade-off is structural rather than arithmetic: every update has row space contained in $\mathrm{row}(A_0)$ (or the analogous fixed column space if B is frozen). SAGE-LORA addresses the setting in which both factors are locally trainable and this linearization is not imposed.

B.3 FLORA

FLoRA preserves each local pairing by stacking weighted factors. An equivalent normalized construction is

$$B_{\mathrm{FLoRA}} = [\sqrt{\alpha_1}B_1, \ldots, \sqrt{\alpha_c}B_c], \qquad A_{\mathrm{FLoRA}} = \begin{bmatrix} \sqrt{\alpha_1}A_1 \\ \vdots \\ \sqrt{\alpha_c}A_c \end{bmatrix}, \tag{23}$$

so that $B_{\mathrm{FLoRA}}A_{\mathrm{FLoRA}} = \sum_i \alpha_i B_i A_i = \bar{D}$. The aggregation is exact (Wang et al., 2024), but the effective rank is at most $\sum_i r_i$ and the stacked downlink grows with cohort size unless the product is merged into a dense weight.

B.4 FLEXLORA

FlexLoRA forms the product aggregate and reparameterizes it through truncated SVD:

$$\bar{D} = U\Sigma V^\top, \qquad B_{(r)} = U_{:,1:r}\Sigma_{1:r}/a_r, \qquad A_{(r)} = V^\top_{:,1:r}. \tag{24}$$

Here $a_r = \alpha_{\mathrm{LoRA}}/r$ is the adapter scale; this asymmetric factorization follows the official rule (Bai et al., 2024). Before truncation, aggregation of the scaled products is exact. At rank r, the approximation error is the spectral tail $\|\bar{D} - [\bar{D}]_r\|_F^2 = \sum_{j>r}\sigma_j^2(\bar{D})$. Heterogeneous clients receive factors truncated to their feasible ranks. A dense SVD is one realization; our evaluated factor-space QR/core SVD produces the same truncated product without materializing the dense aggregate. The direct-materialization cost in Appendix A is therefore not a lower bound on optimized FlexLoRA implementations.

B.5 FSLORA

FSLoRA maintains shared global factors (B^t, A^t) of maximum rank r. Client i samples a Random-k_i diagonal sketch and updates only the selected columns of B and rows of A. After zero-padding the sparse factor updates to the global shapes, the server applies

$$\left(B^{t+1}, A^{t+1}\right) = \left(B^t, A^t\right) + \sum_i \alpha_i \left(\Delta B_i^t, \Delta A_i^t\right). \tag{25}$$

The client uplink scales with $k_i(m+n)$, while the downlink contains the shared rank-r factors and sketch indices. Its optimization and convergence analysis are defined in the shared sketched factor coordinates (Fang et al., 2025).

B.6 FEDROT-LORA

FedRot-LoRA first aligns each common-rank local factorization by an orthogonal matrix O_i obtained from a small Procrustes SVD:

$$\widetilde{B}_i = B_i O_i, \qquad \widetilde{A}_i = O_i^\top A_i, \qquad \widetilde{B}_i \widetilde{A}_i = B_i A_i. \tag{26}$$

The server then averages the aligned factors,

$$B^{t+1} = \sum_i \alpha_i \widetilde{B}_i, \qquad A^{t+1} = \sum_i \alpha_i \widetilde{A}_i. \tag{27}$$

Alignment reduces rotational dispersion and tightens an aggregation-error bound, but Equation (27) is still a factor-product rule and is not identically equal to $\bar{D}$ in general (Zhang et al., 2026).

B.7 FED-SB

Fed-SB fixes shared outer factors B and A and lets client i train only a core $\Phi_i \in \mathbb{R}^{r\times r}$. The server computes

$$\bar{\Phi} = \sum_i \alpha_i \Phi_i, \qquad B\bar{\Phi}A = \sum_i \alpha_i B\Phi_i A. \tag{28}$$

The update is linear and exactly averageable with r^2 communicated scalars, while expressivity is restricted to the shared outer subspaces (Singhal et al., 2025).

B.8 SAGE-LORA

For the two public linear maps $\mathcal{L}_\Omega(D) = D\Omega$ and $\mathcal{L}_\Psi(D) = \Psi D$, SAGE-LORA aggregates

$$\sum_i \alpha_i \mathcal{L}_\Omega(D_i) = \mathcal{L}_\Omega(\bar{D}), \qquad \sum_i \alpha_i \mathcal{L}_\Psi(D_i) = \mathcal{L}_\Psi(\bar{D}). \tag{29}$$

The server adds the public terms $P\Omega$ and ΨP and applies η_{srv}, recovering measurements of $S = P + \eta_{\mathrm{srv}}\bar{D}$. Thus the interface is exact for either update mode; approximation comes from reconstruction, optional sampling, and finite precision.

B.9 WHY DIRECT FACTOR AGGREGATION PRODUCES AN ERROR TERM

For a common rank r, or after a fixed embedding into common factor coordinates, expanding Equation (21) gives

$$\begin{aligned} E_{\mathrm{fac}} &= D_{\mathrm{FedIT}} - \bar{D} \\ &= \sum_{i\neq j} \alpha_i \alpha_j B_i A_j + \sum_i (\alpha_i^2 - \alpha_i) B_i A_i \\ &= -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j (B_i - B_j)(A_i - A_j). \end{aligned} \tag{30}$$

The final identity shows that the error is governed by cross-client factor dispersion, including arbitrary rotations of equivalent factorizations. It vanishes in special linearized cases, such as a factor fixed identically across clients, but not for general two-factor local training. This is the aggregation perturbation isolated in FedRot-LoRA-style convergence bounds. SAGE-LORA removes it structurally because the common linear map is applied only after each $B_i A_i$ has been formed.

B.10 WHY FACTOR-WISE SECURE-AGGREGATION RANDOMNESS DOES NOT CANCEL

Suppose clients attempt to hide factors with additive masks R_i^B and R_i^A satisfying $\sum_i R_i^B = 0$ and $\sum_i R_i^A = 0$. The product sum is

$$\sum_i (B_i + R_i^B)(A_i + R_i^A) = D_\Sigma + E_{\mathrm{mask}}, \tag{31}$$

where $D_\Sigma := \sum_i B_i A_i$ and

$$E_{\text{mask}} = \sum_i \left(B_i R_i^A + R_i^B A_i + R_i^B R_i^A \right). \tag{32}$$

Zero-sum masks do not imply $E_{\text{mask}} = 0$ because the masks are multiplied by client-dependent factors and by one another. For two clients with $R_1^A = \Lambda_A$, $R_2^A = -\Lambda_A$, $R_1^B = \Lambda_B$, and $R_2^B = -\Lambda_B$,

$$E_{\text{mask}} = (B_1 - B_2)\Lambda_A + \Lambda_B(A_1 - A_2) + 2\Lambda_B\Lambda_A, \tag{33}$$

which remains nonzero even when the factor differences vanish. A zero expectation under a particular random-mask distribution would not provide per-round correctness because the realized variance remains. SAGE-LoRA instead masks only after the linear product measurements are formed. If $\sum_i R_i = 0$, then

$$\sum_i \left(\mathcal{L}(D_i) + R_i \right) = \mathcal{L}\left(\sum_i D_i \right) \tag{34}$$

for either $\mathcal{L}_\Omega$ or $\mathcal{L}_\Psi$, so the secure-aggregation randomness cancels exactly in the intended additive domain.

C Proofs and Additional Theoretical Results

C.1 Product-Space Correctness and Reconstruction Error

Theorem C.1 (Product-space aggregation correctness)**.** *In exact arithmetic, if both secure aggregations use the same clients and weights, then*

$$Y^t = S^t \Omega^t, \qquad Z^t = \widetilde{Q}_t^\top S^t. \tag{35}$$

For either complete or residual updates, every contribution is a locally formed product followed by a common linear map; the public base is added by the server. No cross-client term $B_i A_j$ is created. Neither E_{fac} nor the pre-product factor-mask perturbation E_{mask} is introduced. Clipping, quantization, and any almost-homomorphic discrepancy are separate transport errors accounted for by $\varepsilon_{\text{fp}}^t$.

Theorem C.2 (Conditional range-locked reconstruction)**.** *Condition on the right aggregate and the resulting orthonormal $Q \in \mathbb{R}^{d_{\text{o}} \times s_{\text{R}}}$. Let $\Psi = [Q, H]^\top$, $G = \Psi Q$, $Z = \Psi S$, $X = G^\dagger Z = Q^\top S$, and $\mathcal{R}_k(S) = Q[X]_k$. Then*

$$\begin{aligned} \mathcal{R}_k(S) &= [QQ^\top S]_k, \\ \|S - \mathcal{R}_k(S)\|_F^2 &= \|(I - QQ^\top)S\|_F^2 \\ &\quad + \|QQ^\top S - [QQ^\top S]_k\|_F^2. \end{aligned} \tag{36}$$

The two terms are the missed-range energy and the rank-k truncation within the captured range. If $\operatorname{rank}(S) \le k$ and $\operatorname{range}(S) \subseteq \operatorname{range}(Q)$, then $\mathcal{R}_k(S) = S$.

Proposition C.3 (Anchored-probe missed-range bound)**.** *In exact arithmetic, suppose $\Omega^\top \Omega = I$, $\operatorname{row}(M) \subseteq \operatorname{range}(\Omega)$, $Q^\top Q = I$, and $\operatorname{range}(S\Omega) \subseteq \operatorname{range}(Q)$. Writing $\Delta = S - M$ gives*

$$\begin{aligned} (I - QQ^\top)S &= (I - QQ^\top)\Delta(I - \Omega\Omega^\top), \\ \|(I - QQ^\top)S\|_F &\le \|\Delta(I - \Omega\Omega^\top)\|_F \le \|\Delta\|_F. \end{aligned} \tag{37}$$

This applies to $\Delta = \sum_i \alpha_i C_i - M$ for complete updates and $\Delta = \eta_{\text{srv}} \sum_i \alpha_i D_i$ for residual updates. It is a deterministic range bound, not a claim of smaller error than random sketching or improved task utility. The zero-product initialization satisfies the row-space condition for any orthonormal probe.

Define

$$\varepsilon_{\text{repr}}^2(S) := \|(I - QQ^\top)S\|_F^2 + \|QQ^\top S - [QQ^\top S]_k\|_F^2. \tag{38}$$

Let $S_*^t = S^t$ be the full-block target. If optional probability-corrected block sampling is used, define

$$\widetilde{S}^t = P^t + \eta_{\text{srv}} \sum_{i \in \mathcal{U}_t} \alpha_i^t \widetilde{D}_i^t, \tag{39}$$

and let $\mathcal{V}_t$ be the round-indexed conditional sampling variance defined in Equation (79).

Proposition C.4 (Per-round representation and total error budget). *Let* $e^t = \mathrm{vec}(\widehat{M}^{t+1} - \widetilde{S}^t)$. *If clipping, quantization, and secure aggregation perturb exact reconstruction by at most* $\varepsilon_{\mathrm{fp}}^t$, *then*

$$\mathbb{E}_t\|e^t\|^2 \le 2\mathbb{E}_t\varepsilon_{\mathrm{repr}}^2(\widetilde{S}^t) + 2(\varepsilon_{\mathrm{fp}}^t)^2, \tag{40}$$

and

$$\mathbb{E}_t\|\widehat{M}^{t+1} - S_*^t\|_F^2 \le 3\eta_{\mathrm{srv}}^2\mathbb{E}_t\mathcal{V}_t + 3\mathbb{E}_t\varepsilon_{\mathrm{repr}}^2(\widetilde{S}^t) + 3(\varepsilon_{\mathrm{fp}}^t)^2. \tag{41}$$

C.2 Residual-Mode Heterogeneous Local Adaptation Terms

This specialization assumes $P^t = M^t$ and fresh residual products. It does not establish direction bounds for complete-adapter training. For one adapted layer, define the matrix-shaped local gradient $\Gamma_i^t = \nabla_M F_i(M^t)$ and let $[\Gamma_i^t]_{r_i^t}$ be its best rank-r_i^t Frobenius approximation. Under vectorization, $\nabla F(x^t) = \sum_i p_i \,\mathrm{vec}(\Gamma_i^t)$. Define the full and optionally sampled product directions by

$$g_i^t = -\frac{\eta_{\mathrm{srv}}}{\eta}\,\mathrm{vec}(D_i^t), \qquad \widetilde{g}_i^t = -\frac{\eta_{\mathrm{srv}}}{\eta}\,\mathrm{vec}(\widetilde{D}_i^t). \tag{42}$$

Let w_i^t denote the realized aggregation coefficient. The implemented survivor-normalized rule uses $w_i^t = \alpha_i^t$ for $i \in \mathcal{U}_t$ and zero otherwise, so that

$$g^t = \sum_i w_i^t \widetilde{g}_i^t, \qquad x^t - \eta g^t = \mathrm{vec}\left(M^t + \eta_{\mathrm{srv}}\sum_i w_i^t \widetilde{D}_i^t\right). \tag{43}$$

For $w_i^t = \alpha_i^t$, the right-hand side is $\mathrm{vec}(\widetilde{S}^t)$. For the unbiased specializations below, $w_i^t = p_i$ under full participation and $w_i^t = p_i I_i^t / q_i^t$ under independent participation. Here $I_i^t \in \{0,1\}$ is the participation indicator and $q_i^t = \Pr(I_i^t = 1 \mid \mathcal{F}_t)$. Define

$$\tau_{i,t}^2(r_i^t) := \|\Gamma_i^t - [\Gamma_i^t]_{r_i^t}\|_F^2 = \sum_{j > r_i^t} \sigma_j^2(\Gamma_i^t), \tag{44}$$

$$d_{i,t} := \|\mathbb{E}_t g_i^t - \mathrm{vec}([\Gamma_i^t]_{r_i^t})\|, \qquad \nu_{i,t}^2 := \mathbb{E}_t\|g_i^t - \mathbb{E}_t g_i^t\|^2. \tag{45}$$

Here $\tau_{i,t}$ measures rank-limited capacity, $d_{i,t}$ local factor optimization and client drift, and $\nu_{i,t}$ local stochasticity.

Proposition C.5 (A sufficient specialization of the heterogeneity terms). *Let*

$$\sigma_{\mathrm{blk},t}^2 := \mathbb{E}_t\left\|-\frac{\eta_{\mathrm{srv}}}{\eta}\sum_i w_i^t(\widetilde{D}_i^t - D_i^t)\right\|_F^2. \tag{46}$$

Under conditional independence and full participation,

$$\beta_t \le \sum_i p_i(\tau_{i,t}(r_i^t) + d_{i,t}), \qquad \sigma_t^2 \le \sum_i p_i^2\nu_{i,t}^2 + \sigma_{\mathrm{blk},t}^2. \tag{47}$$

If clients participate independently with probabilities $q_i^t > 0$, *independently of their potential local random directions conditional on* $\mathcal{F}_t$, *and a Horvitz–Thompson estimator is used, the same bias bound holds and*

$$\sigma_t^2 \le \sum_i p_i^2\left[\frac{\nu_{i,t}^2}{q_i^t} + \left(\frac{1}{q_i^t} - 1\right)\|\mathbb{E}_t g_i^t\|^2\right] + \sigma_{\mathrm{blk},t}^2. \tag{48}$$

Other cohort-weighting rules add the participation bias

$$\Delta_{\mathrm{part},t} := \left\|\mathbb{E}_t g^t - \sum_i p_i \mathbb{E}_t g_i^t\right\|, \qquad \beta_t \le \sum_i p_i(\tau_{i,t}(r_i^t) + d_{i,t}) + \Delta_{\mathrm{part},t}. \tag{49}$$

C.3 Additional Convergence Results

Corollary C.6 (PL expected-objective bound). *If* F *also satisfies* $\|\nabla F(x)\|^2 \ge 2\mu(F(x) - F_\star)$, $0 < \mu\eta/2 < 1$, *and* $\beta_t^2 \le \beta^2$, $\sigma_t^2 \le \sigma^2$, $\zeta_t^2 \le \zeta^2$, *then*

$$\mathbb{E}[F(x^T) - F_\star] \le (1 - \mu\eta/2)^T(F(x^0) - F_\star) + \frac{4\beta^2}{\mu} + \frac{2L\eta\sigma^2}{\mu} + \frac{2(1 + L\eta)\zeta^2}{\mu}. \tag{50}$$

C.4 Proof of Product-Space Aggregation Correctness

The first secure aggregation releases the weighted uploaded-product measurement. For either public-base choice, linearity gives

$$P^t\Omega^t + \eta_{\mathrm{srv}} \sum_i \alpha_i^t Y_i^t = P^t\Omega^t + \eta_{\mathrm{srv}} \sum_i \alpha_i^t D_i^t \Omega^t$$
$$= \left(P^t + \eta_{\mathrm{srv}} \sum_i \alpha_i^t D_i^t \right) \Omega^t = S^t \Omega^t. \tag{51}$$

Conditioning on this transcript fixes $\widetilde{Q}_t$ and $\Psi^t = \widetilde{Q}_t^\top$. Identical-cohort linearity gives

$$\Psi^t P^t + \eta_{\mathrm{srv}} \sum_i \alpha_i^t Z_i^t = \Psi^t P^t + \eta_{\mathrm{srv}} \sum_i \alpha_i^t \Psi^t D_i^t$$
$$= \Psi^t S^t = \widetilde{Q}_t^\top S^t. \tag{52}$$

Every uploaded product is formed as $\alpha_i^t B_{i,j}^t A_{i,j}^t$ before either common linear map. The protocol never multiplies a sum of client B factors by a sum of client A factors, so no cross-client term arises. This proves Theorem C.1.

C.5 Proof of Conditional Range-Locked Reconstruction

Because $\Psi = [Q, H]^\top$ with $H^\top Q = 0$, $G = \Psi Q = [I_{s_{\mathrm{R}}}; 0_{(s_{\mathrm{L}} - s_{\mathrm{R}}) \times s_{\mathrm{R}}}]$ and therefore $G^\dagger Z = G^\dagger \Psi S = Q^\top S$. Left multiplication by Q is an isometry from $\mathbb{R}^{s_{\mathrm{R}} \times d^i}$ into matrices with columns in $\mathrm{range}(Q)$, so the rank-k truncated SVD commutes with this isometry:

$$Q[G^\dagger Z]_k = Q[Q^\top S]_k = [QQ^\top S]_k. \tag{53}$$

Moreover, $(I - QQ^\top)S$ is Frobenius-orthogonal to every matrix whose columns lie in $\mathrm{range}(Q)$. Applying Pythagoras to $S - \mathcal{R}_k(S) = (I - QQ^\top)S + [QQ^\top S - \mathcal{R}_k(S)]$ yields Equation (36). If S lies in the retained range and has rank at most k, both terms are zero. This proves Theorem C.2. For fixed Q, $G^\dagger = [I_{s_{\mathrm{R}}}, 0_{s_{\mathrm{R}} \times (s_{\mathrm{L}} - s_{\mathrm{R}})}]$ makes the decoder independent of $H^\top S$. These extra rows still contribute to transport and authorized aggregate leakage. The exact-arithmetic statement does not remove the need to check finite-precision orthogonality and the reduced solve.

C.6 Proof of the Anchored-Probe Bound

Since $(I - QQ^\top)S\Omega = 0$, $(I - QQ^\top)S = (I - QQ^\top)S(I - \Omega\Omega^\top)$. The row-space condition gives $M(I - \Omega\Omega^\top) = 0$. Substitute $S = M + \Delta$ and use the contraction of orthogonal projections in Frobenius norm to obtain Equation (37). This argument does not require $\mathrm{range}(M) \subseteq \mathrm{range}(Q)$, which need not hold. A noisy Stage-I sketch requires an additional perturbation term; the exact bound alone does not control that case.

C.7 Proof of the Per-Round Representation and Total Error Budgets

Condition on $\mathcal{F}_t$ and on the locally trained paired blocks. Let $\mathcal{R}_k$ denote exact conditional range-locked reconstruction. Relative to the realized probability-corrected target,

$$\widehat{M}^{t+1} - \widetilde{S}^t = \underbrace{(\mathcal{R}_k(\widetilde{S}^t) - \widetilde{S}^t)}_{\delta_{\mathrm{repr}}^t} + \underbrace{(\widehat{M}^{t+1} - \mathcal{R}_k(\widetilde{S}^t))}_{\delta_{\mathrm{num}}^t}. \tag{54}$$

The deterministic inequality $\|a + b\|^2 \le 2(\|a\|^2 + \|b\|^2)$, Theorem C.2, and the numerical error bound give

$$\mathbb{E}_t \|e^t\|^2 = \mathbb{E}_t \|\widehat{M}^{t+1} - \widetilde{S}^t\|_F^2$$
$$\le 2\mathbb{E}_t \varepsilon_{\mathrm{repr}}^2(\widetilde{S}^t) + 2(\varepsilon_{\mathrm{fp}}^t)^2, \tag{55}$$

which proves Equation (40).

For fidelity relative to full inclusion,

$$\widetilde{S}^t - S_*^t = \eta_{\mathrm{srv}} \sum_i \alpha_i^t (\widetilde{D}_i^t - D_i^t), \tag{56}$$

and

$$\widehat{M}^{t+1} - S_*^t = \underbrace{(\widetilde{S}^t - S_*^t)}_{\delta_{\mathrm{samp}}^t} + \underbrace{(\mathcal{R}_k(\widetilde{S}^t) - \widetilde{S}^t)}_{\delta_{\mathrm{repr}}^t} + \underbrace{(\widehat{M}^{t+1} - \mathcal{R}_k(\widetilde{S}^t))}_{\delta_{\mathrm{num}}^t}. \tag{57}$$

No independence among the three error components is required. Using $\|a+b+c\|^2 \le 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$ gives

$$\mathbb{E}_t \|\widehat{M}^{t+1} - S_*^t\|_F^2 \le 3\mathbb{E}_t \|\delta_{\mathrm{samp}}^t\|_F^2 + 3\mathbb{E}_t \|\delta_{\mathrm{repr}}^t\|_F^2 + 3\mathbb{E}_t \|\delta_{\mathrm{num}}^t\|_F^2. \tag{58}$$

By Proposition D.1 and Equation (56), $\mathbb{E}_t \|\delta_{\mathrm{samp}}^t\|_F^2 = \eta_{\mathrm{srv}}^2 \mathbb{E}_t \mathcal{V}_t$. The reconstruction theorem gives $\|\delta_{\mathrm{repr}}^t\|_F^2 = \varepsilon_{\mathrm{repr}}^2(\widetilde{S}^t)$, and the numerical error bound gives $\|\delta_{\mathrm{num}}^t\|_F \le \varepsilon_{\mathrm{fp}}^t$. Substitution proves Equation (41). The first bound is used in the convergence theorem; the second is the end-to-end fidelity bound relative to full inclusion.

C.8 Proof of the Residual-Mode Heterogeneity Specialization

Let $\bar{g}_i^t = \mathbb{E}_t g_i^t$ and abbreviate $\Gamma_{i,r}^t = [\Gamma_i^t]_{r_i^t}$. Probability correction gives $\mathbb{E}_t \widetilde{g}_i^t = \bar{g}_i^t$. Under full participation,

$$\begin{aligned} \mathbb{E}_t g^t - \nabla F(x^t) &= \sum_i p_i \left(\bar{g}_i^t - \mathrm{vec}(\Gamma_i^t)\right) \\ &= \sum_i p_i \left(\bar{g}_i^t - \mathrm{vec}(\Gamma_{i,r}^t)\right) + \sum_i p_i \,\mathrm{vec}(\Gamma_{i,r}^t - \Gamma_i^t). \end{aligned} \tag{59}$$

The triangle inequality and Equations (44) and (45) imply

$$\|\mathbb{E}_t g^t - \nabla F(x^t)\| \le \sum_i p_i \big(d_{i,t} + \tau_{i,t}(r_i^t)\big), \tag{60}$$

which gives the stated bias envelope.

For the variance, decompose

$$g^t - \mathbb{E}_t g^t = \sum_i p_i (g_i^t - \bar{g}_i^t) + \sum_i p_i (\widetilde{g}_i^t - g_i^t). \tag{61}$$

The second sum has zero conditional mean given the trained blocks, so its cross inner product with the first sum vanishes after iterated expectation. Conditional independence of the centered full-inclusion directions gives $\mathbb{E}_t \|\sum_i p_i (g_i^t - \bar{g}_i^t)\|^2 = \sum_i p_i^2 \nu_{i,t}^2$, while the second variance is $\sigma_{\mathrm{blk},t}^2$ by Equation (46). This proves Equation (47).

For independent Bernoulli participation, $\mathbb{E}_t[I_i^t g_i^t / q_i^t] = \bar{g}_i^t$, so the Horvitz–Thompson estimator has the same conditional mean. For each client,

$$\begin{aligned} \mathbb{E}_t \left\| \frac{I_i^t}{q_i^t} g_i^t - \bar{g}_i^t \right\|^2 &= \frac{1}{q_i^t} \mathbb{E}_t \|g_i^t\|^2 - \|\bar{g}_i^t\|^2 \\ &= \frac{\nu_{i,t}^2}{q_i^t} + \left(\frac{1}{q_i^t} - 1\right) \|\bar{g}_i^t\|^2. \end{aligned} \tag{62}$$

The centered terms are independent across clients. Multiplying by p_i^2, summing, and adding the independent block-sampling direction variance proves Equation (48). For another weighting rule, add and subtract $\sum_i p_i \bar{g}_i^t$ in the conditional mean. The first difference is bounded by Equation (60), while the second has norm $\Delta_{\mathrm{part},t}$ by Equation (49).

This completes the proof of Proposition C.5.

C.9 Implications and Scope of Heterogeneous Local Adaptation

The gradient-tail interpretation below concerns fresh residuals, for which $g_i^t = -(\eta_{\mathrm{srv}}/\eta)\,\mathrm{vec}(D_i^t)$. For normalized complete-update weights, instead $g_i^t = \mathrm{vec}(M^t - C_i^t)/\eta$ and $g^t = \sum_i \alpha_i^t g_i^t$. Although $\mathrm{rank}(C_i^t) \le r_i^t$, its displacement can have rank $r_i^t + k$. The complete-update experiments therefore do not automatically satisfy the residual capacity specialization. Under non-normalized weights, use the global definition $g^t = \mathrm{vec}(M^t - P^t - \eta_{\mathrm{srv}} \sum_i w_i^t \widetilde{D}_i^t)/\eta$; distributing $M^t - P^t$ into each local direction would incorrectly multiply it by $\sum_i w_i^t$.

Rank capacity. The comparator $[\Gamma_i^t]_{r_i^t}$ is analytical and is not computed by a client. Its tail $\tau_{i,t}^2(r) = \sum_{j>r} \sigma_j^2(\Gamma_i^t)$ decreases monotonically with r. Hence, unequal ranks do not introduce an additional aggregation discrepancy; they change the bias envelope only through the product-space directions that clients can represent. Under a fixed system budget, rank allocation should reduce the weighted tails $p_i \tau_{i,t}(r_i)$, or a validated product-space proxy, rather than equalize factor shapes.

A sufficient complete-update envelope. For normalized complete updates, define $g_i^t = \mathrm{vec}(M^t - C_i^t)/\eta$ and $\rho_i^t = r_i^t + k$, since $\mathrm{rank}(C_i^t - M^t) \le \rho_i^t$. With

$$\tau_{i,t}^{\mathrm{cmp}} := \|\Gamma_i^t - [\Gamma_i^t]_{\rho_i^t}\|_F, \qquad d_{i,t}^{\mathrm{cmp}} := \|\mathbb{E}_t g_i^t - \mathrm{vec}([\Gamma_i^t]_{\rho_i^t})\|, \tag{63}$$

the same add-and-subtract argument as in Equation (59) yields

$$\beta_t \le \sum_i p_i \big(\tau_{i,t}^{\mathrm{cmp}} + d_{i,t}^{\mathrm{cmp}}\big) + \Delta_{\mathrm{part},t}. \tag{64}$$

The variance bounds in Proposition C.5 apply after replacing the residual directions by these complete-update directions. This is a sufficient envelope, not a claim that complete-adapter local training realizes a projected gradient step; all optimizer mismatch remains in $d_{i,t}^{\mathrm{cmp}}$.

Statistical heterogeneity and reconstruction. By Equation (38), ζ_t is controlled by the energy outside the captured range and by the singular-value tail beyond the returned rank. Local products whose dominant subspaces are poorly aligned can increase the effective rank of their aggregate and hence these terms. Non-IID data can produce such misalignment through client drift, but the relation is not monotone because weighted updates may also align or cancel. Accordingly, the theory connects heterogeneity to reconstruction through the observed spectrum of $\widetilde{S}^t$, rather than asserting that a distribution-skew scalar alone determines ζ_t.

Factor conditioning and local work. Although $BA = (B\mathcal{T})(\mathcal{T}^{-1}A)$ for any invertible gauge matrix $\mathcal{T}$, the local factor optimization path is not gauge invariant. For perturbations $(\Delta B, \Delta A)$,

$$\|(B + \Delta B)(A + \Delta A) - BA\|_F \le \|A\|_2 \|\Delta B\|_F + \|B\|_2 \|\Delta A\|_F + \|\Delta B\|_F \|\Delta A\|_F. \tag{65}$$

This follows by expansion and Frobenius–spectral submultiplicativity. It explains why extreme factor rescaling can enlarge $d_{i,t}$ and $\nu_{i,t}$ without changing the represented product. The balanced server factorization satisfies $(B^{t+1})^\top B^{t+1} = A^{t+1}(A^{t+1})^\top = \Sigma_{X,k}$. Scale-controlled initialization, optional gauge balancing, and moderate local work are practical controls, not proved bounds on the local optimizer's bias or variance. For residual training, bounded normalized directions require control of $\eta_{\mathrm{srv}} D_i^t/\eta$; for complete updates, the corresponding object is $(C_i^t - M^t)/\eta$. These are additional conditions, not consequences of choosing $\eta_{\mathrm{srv}} = 1$. Excessive local work under non-IID data can increase $d_{i,t}$ even when a larger rank decreases the residual-mode comparator tail $\tau_{i,t}$.

Participation and survivor weighting. The term $\Delta_{\mathrm{part},t}$ vanishes only when the accepted cohort and its weights represent the population objective. Under full participation and one unbiased local step, non-IID data increases dispersion but does not itself create direction bias; with multiple local steps it can magnify drift. Survivor-normalized weights need not be unbiased when resource availability correlates with data or rank. Stratified invitation, minimum tier-wise participation, or inclusion-probability correction can reduce this mismatch. When client variances are comparable and conditionally independent, the effective cohort size

$$n_{\mathrm{eff}}^t = \frac{1}{\sum_i (\alpha_i^t)^2} \tag{66}$$

controls the weighted local-variance contribution. Capping extreme weights may reduce variance, but it changes the optimized objective and must be treated as a separate design choice.

C.10 Proof of the Stationarity Bound

Let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$ and abbreviate $\nabla F(x^t)$ by ∇F_t. By L-smoothness and Equation (16),

$$\begin{aligned} F(x^{t+1}) &\le F(x^t) + \langle \nabla F_t, -\eta g^t + e^t \rangle + \frac{L}{2} \| -\eta g^t + e^t \|^2 \\ &\le F(x^t) - \eta \langle \nabla F_t, g^t \rangle + \langle \nabla F_t, e^t \rangle + L\eta^2 \|g^t\|^2 + L\|e^t\|^2, \end{aligned} \tag{67}$$

where the second line uses $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$. The conditional variance identity and assumption 4.2 give

$$\begin{aligned} \mathbb{E}_t \|g^t\|^2 &= \|\mathbb{E}_t g^t\|^2 + \mathbb{E}_t \|g^t - \mathbb{E}_t g^t\|^2 \\ &\le \|\nabla F_t + b^t\|^2 + \sigma_t^2 \le 2\|\nabla F_t\|^2 + 2\beta_t^2 + \sigma_t^2. \end{aligned} \tag{68}$$

Taking conditional expectation in Equation (67), substituting $\mathbb{E}_t g^t = \nabla F_t + b^t$, and applying Young's inequality yields

$$-\eta \langle \nabla F_t, b^t \rangle \le \frac{\eta}{4} \|\nabla F_t\|^2 + \eta \beta_t^2, \tag{69}$$

$$\begin{aligned} \mathbb{E}_t \langle \nabla F_t, e^t \rangle &= \langle \nabla F_t, \mathbb{E}_t e^t \rangle \\ &\le \frac{\eta}{4} \|\nabla F_t\|^2 + \frac{1}{\eta} \mathbb{E}_t \|e^t\|^2 \le \frac{\eta}{4} \|\nabla F_t\|^2 + \eta \zeta_t^2. \end{aligned} \tag{70}$$

The equality in Equation (70) uses only measurability of ∇F_t with respect to $\mathcal{F}_t$; no independence between e^t and g^t is assumed. Combining Equations (68) to (70) with $\mathbb{E}_t \|e^t\|^2 \le \eta^2 \zeta_t^2$ gives

$$\begin{aligned} \mathbb{E}_t F(x^{t+1}) &\le F(x^t) - \left(\frac{\eta}{2} - 2L\eta^2 \right) \|\nabla F_t\|^2 \\ &\quad + (\eta + 2L\eta^2)\beta_t^2 + L\eta^2 \sigma_t^2 + (\eta + L\eta^2)\zeta_t^2. \end{aligned} \tag{71}$$

For $\eta \le 1/(8L)$, $\eta/2 - 2L\eta^2 \ge \eta/4$ and $\eta + 2L\eta^2 \le 2\eta$. Taking total expectation therefore yields

$$\frac{\eta}{4} \mathbb{E} \|\nabla F_t\|^2 \le \mathbb{E}[F(x^t) - F(x^{t+1})] + 2\eta \beta_t^2 + L\eta^2 \sigma_t^2 + \eta(1 + L\eta)\zeta_t^2. \tag{72}$$

Summing Equation (72) from $t = 0$ to $T-1$ telescopes the objective values. Using $F(x^T) \ge F_\star$, dividing by $\eta T/4$, and substituting the averaged quantities gives

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla F(x^t)\|^2 \le \frac{4(F(x^0) - F_\star)}{\eta T} + 8\bar{\beta}_T^2 + 4L\eta \bar{\sigma}_T^2 + 4(1 + L\eta)\bar{\zeta}_T^2, \tag{73}$$

which proves Theorem 4.4.

C.11 Consequences of the Stationarity Bound

Term-wise control. At a fixed step size, the terms in Equation (19) have distinct sources. The residual-mode specialization relates β_t to local-gradient tails, solver error, drift, and weighting; the complete mode needs its own direction bounds. Mini-batch noise, participation, and optional block sampling contribute to σ_t^2. Reconstruction distortion depends on the captured range, returned rank k, and numerical transport. Increasing s_{R} changes range capacity, but increasing s_{L} alone with fixed Q adds discarded H rows and does not improve the exact decoder. None of these controls alone guarantees a smaller task loss or removes cohort bias.

Vanishing-error regime. If $\beta_t = \zeta_t = 0$, $\sigma_t^2 \le \sigma^2$, and $\eta = \Theta(T^{-1/2})$, the time-averaged expected squared gradient norm is $O(T^{-1/2})$. The same order holds when $\bar{\beta}_T^2 + \bar{\zeta}_T^2 = O(T^{-1/2})$ and direction variance stays bounded. Under the stated distortion envelope, this sufficient regime controls average squared post-direction error at order $T^{-3/2}$. More generally, vanishing normalized mean-square distortion removes its contribution to this bound; it is not a necessary condition for every possible algorithm to approach stationarity. A fixed error floor leaves a nonzero bound at fixed step size. Equivalently, an independent uniformly selected iterate from $\{0, \dots, T-1\}$ obeys the same expected-gradient bound; no last-iterate conclusion follows.

Fixed global rank. The measure in Equation (19) is the full product-space gradient norm. Rank truncation belongs to ζ_t, rather than being an exact unconstrained step. Vanishing missed-range and truncation terms can make this bound informative about exact stationarity when the other terms also vanish; a nonzero upper-bound term does not by itself prove non-convergence. The PL result controls expected objective gap, not parameter convergence. A projected- or Riemannian-gradient mapping would concern a different rank-constrained analysis and is not implied here.

C.12 Proof of the PL Corollary

Under the PL condition, Equation (72) and $\|\nabla F(x^t)\|^2 \geq 2\mu(F(x^t) - F_\star)$ imply

$$\mathbb{E}[F(x^{t+1}) - F_\star] \leq (1 - \mu\eta/2)\mathbb{E}[F(x^t) - F_\star] + 2\eta\beta^2 + L\eta^2\sigma^2 + \eta(1 + L\eta)\zeta^2. \tag{74}$$

Let $\rho = 1 - \mu\eta/2$. Iterating Equation (74) gives

$$\mathbb{E}[F(x^T) - F_\star] \leq \rho^T(F(x^0) - F_\star) + \left(2\eta\beta^2 + L\eta^2\sigma^2 + \eta(1 + L\eta)\zeta^2\right)\sum_{j=0}^{T-1}\rho^j. \tag{75}$$

Since $\sum_{j=0}^{T-1}\rho^j \leq (1-\rho)^{-1} = 2/(\mu\eta)$, the three additive terms become $4\beta^2/\mu$, $2L\eta\sigma^2/\mu$, and $2(1 + L\eta)\zeta^2/\mu$, respectively. This proves Corollary C.6.

D Optional Influence-Guided Block Sampling

Full inclusion is the primary method because the two dense sketches have fixed dimensions. Optional sampling is useful only when evaluating every trained block exceeds a separate encoding-compute budget. We suppress the round index t in this section.

Functional influence. Place a scalar gate $\gamma_{i,j}$ on a paired block. For local layer input h and output gradient g_{out},

$$\phi_{i,j} = \left|\frac{\partial \widehat{F}_i}{\partial \gamma_{i,j}}\right| = \left|\left\langle B_{i,j}^\top g_{\text{out}}, A_{i,j}h\right\rangle\right|. \tag{76}$$

This score is invariant to simultaneous within-block orthogonal reparameterization and measures the block product rather than either factor alone. It may be combined with product energy or affinity to the public global subspaces, but any semantic interpretation requires a matched-budget ablation.

Probability allocation. For encoding cost $c_{i,j}^{\text{enc}}$ and nonnegative score $a_{i,j}$, solve the clipped surrogate

$$\min_{\{\pi_{i,j}\}} \sum_j a_{i,j}\frac{1 - \pi_{i,j}}{\pi_{i,j}} \quad \text{s.t.} \quad \sum_j c_{i,j}^{\text{enc}}\pi_{i,j} \leq C_i^{\text{enc}}, \quad \pi_{\min} \leq \pi_{i,j} \leq 1. \tag{77}$$

Ignoring active clipping constraints gives $\pi_{i,j} \propto \sqrt{a_{i,j}/c_{i,j}^{\text{enc}}}$. The client represents

$$\widetilde{D}_i = \sum_j \frac{J_{i,j}}{\pi_{i,j}} D_{i,j}, \qquad J_{i,j} \sim \text{Bernoulli}(\pi_{i,j}). \tag{78}$$

Proposition D.1 (Optional sampling is unbiased)**.** *Conditioned on the trained blocks and independent Bernoulli draws,*

$$\mathbb{E}_J\left[\sum_i \alpha_i \widetilde{D}_i\right] = \sum_i \alpha_i D_i,$$

$$\mathcal{V} := \mathbb{E}_J\left[\left\|\sum_i \alpha_i(\widetilde{D}_i - D_i)\right\|_F^2\right] = \sum_i \alpha_i^2 \sum_j \left(\frac{1}{\pi_{i,j}} - 1\right)\|D_{i,j}\|_F^2. \tag{79}$$

Proof. For each block, $\mathbb{E}[J_{i,j}/\pi_{i,j}] = 1$. Distinct centered Bernoulli terms are independent and have zero expected inner product, while $\mathbb{E}[(J/\pi - 1)^2] = 1/\pi - 1$. Linearity gives the first statement and summing the block variances gives the second. Setting $a_{i,j} = \|D_{i,j}\|_F^2$ makes Equation (77) the exact conditional-variance allocation; a general influence score optimizes only the declared surrogate.

E Formal Threat Model and Secure-Aggregation Details

E.1 Security Model, Definition, and Composition

Let $\mathcal{A}_t$ be the clients invited to round t. We analyze the strongest semi-honest coalition $\mathcal{C}_t = \{\mathsf{Srv}\} \cup \mathcal{C}_t^{\mathrm{cl}}$, where $\mathcal{C}_t^{\mathrm{cl}} \subseteq \mathcal{A}_t$ is the set of corrupted clients and remains within the threshold supported by the selected secure-aggregation protocol; a coalition excluding the server receives only a subview. For stage $j \in \{1, 2\}$, let $\mathcal{U}_t^{(j)}$ be the survivor set if that stage succeeds and let $\mathcal{H}_t^{(j)} = \mathcal{U}_t^{(j)} \setminus \mathcal{C}_t^{\mathrm{cl}}$ be its honest survivors. Let $\mathbf{u}_t^{(j)}$ be the encoded client inputs submitted to stage j and define the honest-survivor sum

$$U^{(j)}_{\mathcal{H}_t^{(j)}} := \sum_{i \in \mathcal{H}_t^{(j)}} u_i^{(j)}. \tag{80}$$

For each invoked stage, define its authorized output as

$$\mathsf{Out}_t^{(j)} = \begin{cases} \left(\mathsf{accept}, \mathcal{U}_t^{(j)}, U^{(j)}_{\mathcal{H}_t^{(j)}}\right), & \text{if the stage releases an aggregate,} \\ \mathsf{abort}, & \text{otherwise.} \end{cases} \tag{81}$$

The corrupted clients' own inputs, which are included in $\mathrm{in}_{\mathcal{C}_t}^{(j)}$, together with $U^{(j)}_{\mathcal{H}_t^{(j)}}$ determine the released total sum. Let Π_j be the secure-aggregation protocol used in stage j and $\mathrm{pub}_t^{(j)}$ its public context. We assume the per-stage simulation property

$$\mathrm{View}_{\mathcal{C}_t}^{\Pi_j}\left(1^\lambda, \mathrm{pub}_t^{(j)}, \mathbf{u}_t^{(j)}\right) \approx_c \mathrm{Sim}_j\left(1^\lambda, \mathrm{pub}_t^{(j)}, \mathrm{in}_{\mathcal{C}_t}^{(j)}, \mathsf{Out}_t^{(j)}\right), \qquad j \in \{1, 2\}. \tag{82}$$

Here λ is the security parameter, $\mathrm{View}_{\mathcal{C}_t}^{\Pi_j}$ contains the coalition's state, randomness, and received messages, Sim_j is probabilistic polynomial time, and $\approx_c$ denotes computational indistinguishability. We require this simulation guarantee to hold with arbitrary polynomial-size auxiliary input and under sequential composition, including prior transcripts and correlated inputs across the two stages. Stage-II public information includes Ψ^t, which is computed from the authorized Stage-I aggregate and the public range-extraction rule.

The SAGE protocol requires Stage II to be bound before execution to the accepted Stage-I survivor set and protected aggregate count. It releases no Stage-II aggregate unless these values match. With Stage II omitted when Stage I aborts, define

$$\mathsf{Leak}_t = \begin{cases} \left(\mathsf{Out}_t^{(1)}, \mathsf{Out}_t^{(2)}\right), & \text{if both stages are accepted,} \\ \left(\mathsf{Out}_t^{(1)}, \mathsf{abort}\right), & \text{if Stage I succeeds and Stage II aborts,} \\ \mathsf{abort}, & \text{if Stage I aborts.} \end{cases} \tag{83}$$

When Stage II is not invoked, set $\mathbf{u}_t^{(2)} = \varnothing$. The desired composed guarantee is

$$\mathrm{View}_{\mathcal{C}_t}^{\Pi_{\mathrm{SAGE}}}\big(1^\lambda, \mathrm{pub}_t, \mathbf{u}_t^{(1)}, \mathbf{u}_t^{(2)}\big) \approx_c \mathrm{Sim}_{\mathrm{SAGE}}\big(1^\lambda, \mathrm{pub}_t, \mathrm{in}_{\mathcal{C}_t}, \mathsf{Leak}_t\big)\,. \tag{84}$$

The public information pub_t contains the mode, base P^t, current global adapter M^t, layer orientation, round/layer and distinct stage identifiers, projections, dimensions, ordered Stage-I survivor set, encoding parameters, protocol thresholds, and authenticated same-cohort binding. These data are public in both complete and residual modes. Equation (84) is the target for the specified two-stage execution, conditional on the released aggregates, including any H rows unused by reconstruction.

Proposition E.1 (Sequentially composed aggregate confidentiality)**.** *Suppose each secure-aggregation stage satisfies the auxiliary-input, sequentially composable guarantee in Equation* (82) *under its declared collusion and dropout thresholds. Suppose further that the stages use distinct session identifiers and fresh domain-separated masking randomness, and that Stage II releases an aggregate only after its ordered survivor set and protected aggregate count have been authenticated against Stage I. Then the composed SAGE execution satisfies Equation* (84)*. In an accepted round, the coalition learns no additional information about an honest client's individual factors, local rank, block selection, protected count, or product sketches beyond the public context, the corrupted parties' own inputs, and the authorized leakage in Equation* (83)*.*

Proof. Following standard simulation-based hybrid analyses for secure aggregation (Bonawitz et al., 2017; Liu et al., 2022; Wang et al., 2025), let H_0 be the real two-stage transcript. Construct H_1 by replacing the Stage-I transcript with the output of Sim_1 while preserving $\mathsf{Out}_t^{(1)}$. By Equation (82), $H_0 \approx_c H_1$. Since the authorized Stage-I output is unchanged, the server derives the same Y^t, $\widetilde{Q}_t$, and Ψ^t in both hybrids. Conditioned on this identical Stage-II public context, construct H_2 by replacing the Stage-II transcript with the output of Sim_2 while preserving $\mathsf{Out}_t^{(2)}$. Again, Equation (82) gives $H_1 \approx_c H_2$. Therefore,

$$H_0 \approx_c H_1 \approx_c H_2, \tag{85}$$

where H_2 is distributed as the output of $\mathrm{Sim}_{\mathrm{SAGE}}$ in Equation (84). Abort executions follow by using the corresponding authorized abort output in Equation (83). This proves the proposition. □

This guarantee is conditional on the authorized aggregate outputs. It establishes transcript confidentiality, but not differential privacy of those outputs or protection against inference from the released aggregate sketches or global adapter.

E.2 Secure-Aggregation Realization and Reference Scope

Illustrative SHPRG encoding. The following describes a compatible masking construction, not the measured Flower transport. For stage $j \in \{1, 2\}$, let $\mathsf{S}_i^{(1)} = Y_i^t$ and $\mathsf{S}_i^{(2)} = Z_i^t$, with $d_1^{\mathrm{enc}} = d^o s_{\mathrm{R}} + 1$ and $d_2^{\mathrm{enc}} = s_{\mathrm{L}} d^i + 1$ in the direct orientation. A client clips at χ_j, applies fixed-point quantizer Q_{Δ_j}, and encodes its weighted sketch and protected count as

$$u_i^{(j)} = \mathrm{concat}\Big(\mathsf{Q}_{\Delta_j}[\mathrm{clip}_{\chi_j}(n_i\,\mathrm{vec}(\mathsf{S}_i^{(j)}))], n_i\Big) \in \mathbb{Z}_P^{d_j^{\mathrm{enc}}}. \tag{86}$$

Let $\mathcal{K}_j$ be an additive seed group. With an SHPRG $\mathsf{G}_j : \mathcal{K}_j \to \mathbb{Z}_P^{d_j^{\mathrm{enc}}}$, client i sends

$$\widetilde{u}_i^{(j)} = u_i^{(j)} + \mathsf{G}_j(\kappa_i^{(j)}) \pmod P. \tag{87}$$

The secure aggregation protocol reconstructs only the survivor aggregate seed $\kappa_{\mathcal{U}_t}^{(j)} = \sum_{i\in\mathcal{U}_t} \kappa_i^{(j)}$ and removes the aggregate mask. Seed homomorphism gives

$$\sum_{i\in\mathcal{U}_t} \widetilde{u}_i^{(j)} - \mathsf{G}_j(\kappa_{\mathcal{U}_t}^{(j)}) = \sum_{i\in\mathcal{U}_t} u_i^{(j)} + \delta_{\mathrm{prg},t}^{(j)} \pmod P, \tag{88}$$

where $\delta_{\mathrm{prg},t}^{(j)} = 0$ for an exactly homomorphic construction and is bounded for an almost-homomorphic SHPRG such as SASH (Liu et al., 2022). Equation (88) establishes arithmetic mask removal and aggregation correctness. It is not, by itself, a confidentiality proof; confidentiality additionally relies on the simulation property in Equation (82). The count coordinate is aggregated exactly, either through a separate integer secure sum or by enforcing a zero discrepancy in that coordinate. After decoding, the server divides the summed sketch coordinates by $\sum_{i\in\mathcal{U}_t} n_i$, yielding the α_i^t-weighted sketches in Equations (10) and (13).

Relation to FSLoRA. FSLoRA supports secure aggregation of sparse ΔB_i and ΔA_i updates in a shared global factor coordinate system (Fang et al., 2025). Its one-stage cancellation identity preserves that factor-coordinate aggregation rule, rather than the product average $\sum_i \alpha_i B_i A_i$ of independently parameterized adapters. SAGE-LORA instead evaluates linear measurements of each paired product locally, then aggregates them in common dimensions. The two methods share additive mask cancellation but protect different aggregation objects. Sage additionally binds the second, range-conditioned SecAgg invocation to the same surviving cohort as the first.

Relation to FedASK. FedASK couples two-stage product sketching with DP local training to balance privacy, adapter learnability, and aggregation fidelity (Wen et al., 2025). Its algorithm also includes a non-DP update option. SAGE-LORA addresses aggregate confidentiality at the communication interface. It forms fixed-width measurements of unequal-rank client products and applies SecAgg to both stages, with the second projection derived from the first aggregate. Public-adapter anchoring determines the right probe, while authenticated cohort and weighting consistency ensure that both measurements describe the same product average. The resulting security analysis concerns sequential transcript confidentiality conditional on the authorized aggregate outputs.

Dropout and same-cohort correctness. The first secure aggregation completes dropout recovery and fixes the ordered survivor set $\mathcal{U}_t = \mathcal{U}_t^{(1)}$. Stage II is initialized with an authenticated binding to this set, the protected aggregate count, and the round/layer identifiers. It must abort before releasing a sum if any bound client is absent or any binding value differs. Hence an aborted Stage II reveals no left-sketch aggregate and produces no global adapter update. The Stage-I aggregate, once released, cannot be erased; any retry uses a fresh round identifier, newly generated projections, recomputed local updates, and independent masking randomness. Aggregates from overlapping cohorts must not be differenced for the same local update, and post-hoc intersection is invalid because the two sketches would represent different product averages.

Numerical conditions. The public clipping bounds, fixed-point scales, modulus, maximum survivor count, and per-client sample bound must prevent modular wraparound. Range extraction and reduced decomposition use numerically stable linear algebra, and the implementation rejects non-finite outputs, invalid ranks, or failed orthogonality checks. Clipping, quantization, and any almost-homomorphic discrepancy are included in $\varepsilon_{\mathrm{fp}}^t$ in Proposition C.4.

Evaluated reference versus formal specification. The measured Flower-based reference uses coarse and residual encoding channels with 24-bit quantization and arithmetic modulo 2^{32}, with seven shares and reconstruction threshold five among ten participants. These secure channels are not BF16 wire tensors or an SHPRG benchmark. The wrapper checks requested client identities before each invocation and compares the two stages before model reconstruction, but the latter check occurs after unmasking. It is therefore not a production-authenticated realization of the before-release binding required by Proposition E.1; the experiments measure this reference's utility and cost, not a deployment certification. Algorithms 1 and 2 specify the stronger release rule.

Payload and evaluation boundaries. Per client-round, the logical account includes 4,415,936 sketch-upload bytes (BF16 values and 448 count bytes), 4,358,144 FP32 adapter bytes, and 6,881,328 projection/public-metadata bytes (6,881,280 FP32 projection bytes and 48 metadata bytes). The total is 15,655,408 bytes, or 14.58 GiB across 1,000 client-rounds. This is a logical model/protocol account, not encoded-channel or network-wire measurement. Native baseline savings include precision choices, not only scalar compression. Evaluation covers one backbone, eight commonsense tasks and five GLUE validation tasks; timings use mixed hardware. The security theorem assumes semi-honest participants and the underlying protocol's thresholds, and excludes malicious deviations, Sybil attacks, poisoning, side channels, traffic analysis, denial of service, and inference from released models. SecAgg is not differential privacy.

Algorithm 1 Two-stage secure aggregation specification for SAGE sketches

Input: Client-held paired-block factors and protected counts $\{n_i\}$, public base P^t, right projection Ω^t, target rank k, server coefficient η_{srv}, and a dropout-resilient secure-aggregation protocol.
Output: Survivor set $\mathcal{U}_t$, normalized aggregates $(\bar{Y}^t, \bar{Z}^t)$, and range-locked projection Ψ^t, or $\perp$.

Stage I: protected right-sketch aggregation.
for each invited client i in parallel **do**
  Compute Y_i^t by Equation (9); encode and mask $(n_i Y_i^t, n_i)$.
end for
Secure aggregation returns $(\mathcal{U}_t, N_t, Y_\Sigma^t)$, where $N_t = \sum_{i\in\mathcal{U}_t} n_i$ and $Y_\Sigma^t = \sum_{i\in\mathcal{U}_t} n_i Y_i^t$.
if Stage I fails or $\mathcal{U}_t$ is below the protocol threshold **then**
  return $\perp$
end if
Set $\bar{Y}^t \leftarrow Y_\Sigma^t / N_t$ and $Y^t \leftarrow P^t\Omega^t + \eta_{\mathrm{srv}}\bar{Y}^t$ as in Equation (10).
Estimate the thresholded effective rank $\widehat{r}_t$ of Y^t.
if $\widehat{r}_t < k$ **then**
  return $\perp$ before Stage II; any retry uses a fresh round identifier and probe
end if
Extract $\widetilde{Q}_t$, set $\Psi^t \leftarrow \widetilde{Q}_t^\top$, and distribute Ψ^t to $\mathcal{U}_t$.
Stage II: protected left-sketch aggregation.
for each $i \in \mathcal{U}_t$ in parallel **do**
  Compute Z_i^t by Equation (12) from the same frozen paired-block factors; encode and mask $(n_i Z_i^t, n_i)$.
end for
Invoke secure aggregation with the authenticated binding $(\mathcal{U}_t, N_t)$ and distinct Stage-II session identifier. It returns Z_Σ^t only if the binding verifies and every bound survivor completes; otherwise it returns $\perp$ without releasing a left-sketch sum.
if Stage II returns $\perp$ **then**
  return $\perp$
end if
Set $\bar{Z}^t \leftarrow Z_\Sigma^t / N_t$.
return $(\mathcal{U}_t, \bar{Y}^t, \bar{Z}^t, \Psi^t)$

F Additional Experimental Results

F.1 Training and Evaluation Protocol

Table 6: Evaluation sets and reported primary metrics. The five GLUE tasks are evaluated on their validation splits, rather than the hidden official test sets.

Benchmark/task	Primary metric	Validation examples
CoLA	MCC	1,043
SST-2	Accuracy	872
MRPC	(Accuracy + F1)/2	408
QNLI	Accuracy	5,463
RTE	Accuracy	277
Eight commonsense tasks	Per-task accuracy; equal-task macro	19,740 (total)

GLUE tasks are trained independently. The classifier head uses sample-weighted FedAvg and is charged in both directions.

Data and optimization. The commonsense training pool contains 147,580 examples: ARC-Challenge 1,119; ARC-Easy 2,251; BoolQ 9,427; HellaSwag 39,905; OpenBookQA 4,957; PIQA 16,113; Social-IQA 33,410; and WinoGrande 40,398. Their validation counts are 299, 570, 3,270, 10,042, 500, 1,838, 1,954 and 1,267, respectively. GLUE tasks are trained separately, not mixed into the commonsense pool. The homogeneous and balanced cohorts use Dirichlet concentration 10; the heavy-tail primary cohort uses 0.1. A microbatch of one with 16 gradient-accumulation steps gives

effective batch size 16. Hyperparameters are fixed before evaluation; learning curves report training loss.

Optimizer and adapter settings. Each local client uses a fresh AdamW optimizer with learning rate 2×10^{-4}, zero weight decay, $(\beta_1, \beta_2) = (0.9, 0.999)$ and $\epsilon = 10^{-8}$. The gradient norm is clipped to 1; no optimizer moments persist across client updates, and no learning-rate schedule is applied. The backbone is frozen, and BF16 LoRA is attached to its 56 `q_proj`/`v_proj` modules with maximum sequence length 512, no quantization, and zero LoRA dropout. Ordinary client adapters use scaling $\alpha_{\mathrm{LoRA}}/r_i = 1$; FSLoRA uses shared rank-64 factors with eight active random-k coordinates.

Complete products versus fresh residuals. The homogeneous complete-update formulation fine-tunes a client adapter $C_i^t = B_i^t A_i^t$ initialized from the global state and reconstructs $S^t = \sum_i \alpha_i^t C_i^t$, replacing the global product. In Equation (5), this is $P^t = 0$, $\eta_{\mathrm{srv}} = 1$, and $D_i^t = C_i^t$. Residual adaptation instead sets $P^t = M^t$ and trains fresh blocks on top of frozen M^t. These modes share a linear measurement interface, not an equivalent optimization trajectory: even when $\mathrm{rank}(C_i^t) \le r_i$, the displacement $C_i^t - M^t$ can have rank $r_i + k$. Profile comparisons therefore retain their distinct local objectives. The distinction separates what the protocol preserves—linear measurements of paired products—from what the local optimizer learns. Consequently, a change between profiles can reflect both the representable local update and its subsequent reconstruction. GLUE models are trained independently; all methods aggregate the classifier head by sample-weighted FedAvg and count it in both directions.

Decoder and baseline implementation. Complete-update Sage keeps a 14-column thin-QR range, sends 20 left measurements, and returns rank 8 from $Q[G^\dagger Z]_8$. The six complementary measurement rows are included in traffic but not recycled into the returned adapter. The solve uses relative cutoff 10^{-3} and monitored conditioning; balanced returned factors avoid scale imbalance. Sketch widths remain fixed throughout training. FedIT averages LoRA factors while keeping the backbone frozen. FLoRA retains sample-weighted factor pairing by stacking. FlexLoRA preserves its official asymmetric SVD factorization; factor-space QR/core SVD computes the same truncated product without dense materialization. FSLoRA uses random-k coordinates in shared global factors, and FedRot-LoRA aligns factors before averaging. Main-table comparisons share task, local-rank budget, partition, participant schedule, and seed settings; global representation ranks follow each method's protocol.

F.2 Complete Transfer Results and Task Profiles

Task-wise transfer. SAGE-LoRA's GLUE-5 score of 81.22 improves on FLoRA, FSLoRA, and FedASK by 2.40, 1.37, and 0.54 points (Table 2). It exceeds FLoRA on all five tasks, while its advantage over FlexLoRA is concentrated in CoLA and MRPC. FedIT, zero-padding, and FedRot-LoRA remain strong alternatives in this homogeneous-rank setting. The methods differ in how they combine and return the adapters: factor averaging uses shared coordinates, stacking keeps client factors, and SAGE-LoRA reconstructs a compact approximation to the product average. Since the classifier head is aggregated identically, the results show that the latter representation remains useful when the local classification objective changes. Product consistency defines the target being reconstructed, rather than ensuring that its training path always yields the best score.

Table 7: MRPC validation metrics (%): mean and sample SD over three seeds. Both component metrics are retained; the last column is the primary score used in Table 2.

Method	Accuracy	F1	(Acc. + F1)/2
FedIT	$82.35_{\pm 0.98}$	$87.45_{\pm 0.89}$	$84.90_{\pm 0.92}$
Zero-padding	$\mathbf{83.74}_{\pm 1.11}$	$\mathbf{88.49}_{\pm 0.75}$	$\mathbf{86.12}_{\pm 0.93}$
FLoRA	$78.51_{\pm 1.23}$	$84.59_{\pm 0.93}$	$81.55_{\pm 1.08}$
FlexLoRA	$80.88_{\pm 0.88}$	$86.55_{\pm 0.68}$	$83.71_{\pm 0.75}$
FS-LoRA	$81.45_{\pm 0.62}$	$87.07_{\pm 0.35}$	$84.26_{\pm 0.48}$
FedASK	$80.15_{\pm 0.85}$	$87.21_{\pm 1.21}$	$83.68_{\pm 1.03}$
FedRot-LoRA	$83.09_{\pm 0.88}$	$88.16_{\pm 0.74}$	$85.62_{\pm 0.81}$
SAGE-LORA	$81.70_{\pm 1.72}$	$86.88_{\pm 1.48}$	$84.29_{\pm 1.60}$

Accuracy and F1 on MRPC. SAGE-LoRA improves accuracy over FedASK from 80.15% to 81.70%, yielding a 0.61-point gain in the composite score despite a 0.33-point reduction in F1. Against FLoRA, both accuracy and F1 improve, by 3.19 and 2.29 points; FSLoRA's composite score is close to SAGE-LoRA's. Thus, the compact adapter retains task-useful updates without requiring gains in every metric. Accuracy counts all correct decisions, whereas F1 reflects a different balance of classification errors. Keeping both metrics makes clear that SAGE-LoRA's advantage over FedASK comes from accuracy, not a uniform improvement in prediction behavior. The aggregation design does not prescribe which metric should benefit more.

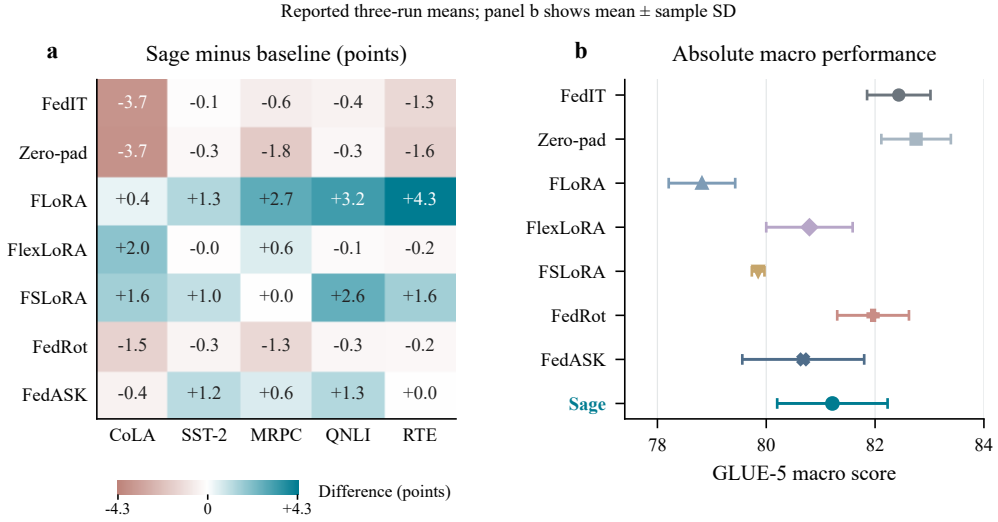

Figure 5: GLUE-5 task differences and absolute macro scores. (a) Mean primary-score differences, SAGE-LoRA minus each baseline, in percentage points. (b) Absolute macro scores with sample standard deviations over three seeds. Positive heatmap entries favor SAGE-LoRA. MRPC uses mean accuracy/F1.

Distribution of transfer gains. The heatmap in Figure 5 shows that SAGE-LoRA's advantage over FedASK is shared by QNLI, SST-2, and MRPC, with gains of 1.27, 1.18, and 0.61 points. CoLA is 0.41 points lower and RTE is equal at the reported precision. The improvement is therefore spread across several objectives, rather than being driven by one large task gain. Reconstructing in an aggregate-derived range gives each task a data-dependent adapter whose directions can change as training proceeds. The mixed task ranking shows that those directions are not equally useful for every objective. Panel (b) reports absolute macro scores and their between-seed variation; it does not give the uncertainty of paired score differences.

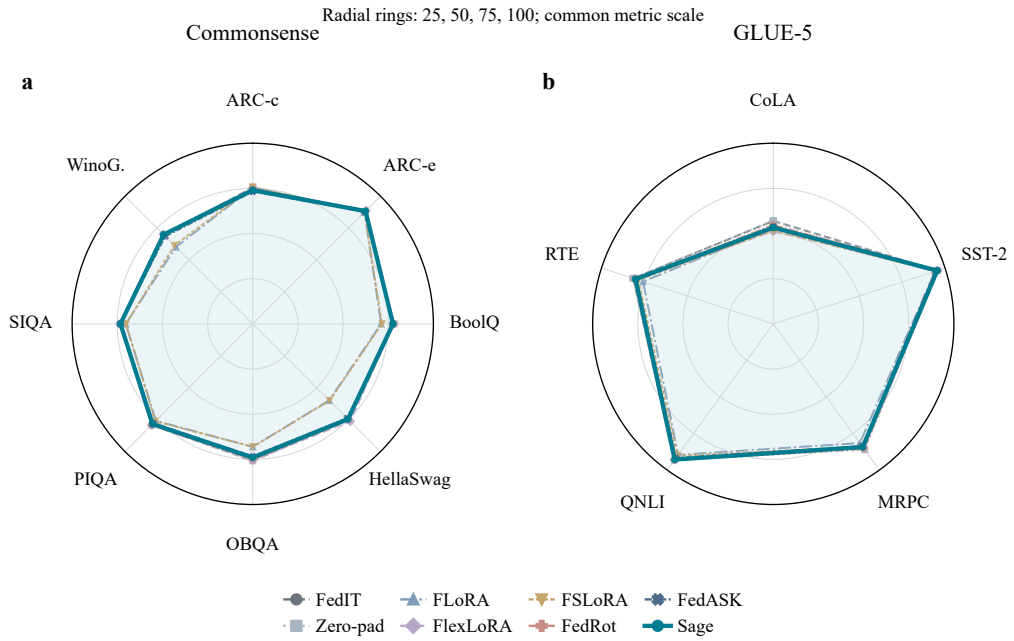

Figure 6: Supplementary task profiles on fixed 0–100 scales. All eight methods are shown; the pale Sage polygon is a visual locator, not an additional metric. CoLA is MCC $\times 100$, MRPC is mean accuracy/F1 $\times 100$, and other axes are accuracy. Polygon areas are not performance aggregates; the raw tables define the comparison.

Commonsense task profiles. SAGE-LoRA improves seven commonsense task means over FedASK, with the largest gain on WinoGrande (1.40 points); Social-IQA is 0.12 points lower (Figure 6 and Table 1). Against FLoRA, gains are particularly large on HellaSwag (14.67 points) and WinoGrande (9.71). SAGE-LoRA measures each local product while keeping its factors paired, then estimates a shared range and returns the leading rank-8 components within it. The public-adapter anchor includes previously learned directions in the probe, while the remaining probe columns examine directions outside that subspace. This explains how the design combines accumulated adaptation with new client updates in a small returned adapter. The broad task gains support its usefulness; because FLoRA and FedASK also preserve local pairing, the comparison concerns their complete training and aggregation procedures, not pairing alone.

F.3 Heterogeneous Task and Rank-Bucket Results

Accuracy across rank profiles. Under balanced ranks, SAGE-LoRA achieves 74.46%, improving on FLoRA, FSLoRA, and FedASK by 3.78, 2.89, and approximately 0.60 points (Table 8). Its five task improvements over FedASK include 2.57 points on ARC-Challenge and 2.03 on PIQA. Under heavy-tail ranks, it improves seven task means and the macro score by approximately 0.95 points over FedASK; OpenBookQA gains 2.74 points, while BoolQ is 2.07 points lower. In SAGE-LoRA, local rank determines how many paired blocks a device can train, but does not determine the dimensions of the transmitted sketches. The results show that useful adaptation can be retained across both rank profiles, while this common message format allows different local capacities to use the same additive SecAgg interface.

Table 8: Profile-specific heterogeneous-rank evaluations: all eight task accuracies and macro-average (%). Each cell gives the mean above its ± sample SD over three seeds. Protocol cohorts are separated rather than pooled.

Method	ARC-c	ARC-e	BoolQ	HellaS.	OBQA	PIQA	SIQA	WinoG.	Avg.
Balanced $\{2,4,8,16\}$*; Dirichlet* $\alpha = 10$									
FLoRA	74.58 ±0.33	87.25 ±0.44	70.32 ±0.45	59.00 ±0.67	67.60 ±0.80	76.48 ±0.26	69.94 ±0.36	60.30 ±0.60	70.68 ±0.07
FlexLoRA	74.36 ±0.19	87.60 ±0.44	**76.29** ±0.49	**72.03** ±0.35	72.33 ±1.45	79.05 ±0.19	**72.89** ±0.16	**68.88** ±0.25	**75.43** ±0.18
FS-LoRA	75.59 ±0.33	87.66 ±0.10	71.87 ±0.23	60.37 ±0.42	68.60 ±0.20	76.88 ±0.30	70.28 ±0.21	61.30 ±0.33	71.57 ±0.08
FedASK	73.13 ±0.39	86.73 ±0.51	75.19 ±0.18	70.39 ±0.52	70.60 ±0.20	76.82 ±0.38	70.57 ±0.31	67.48 ±0.14	73.86 ±0.31
FedRot-LoRA	73.91 ±0.33	**87.78** ±0.44	**76.29** ±0.38	69.73 ±0.46	**72.40** ±1.11	**79.11** ±0.28	71.77 ±0.43	67.64 ±0.55	74.83 ±0.29
SAGE-LORA	**75.70** ±0.77	87.37 ±0.18	74.79 ±0.21	70.19 ±0.28	70.73 ±0.92	78.85 ±0.03	71.32 ±0.18	66.72 ±0.45	74.46 ±0.36
Heavy-tail ranks 4–64; Dirichlet $\alpha = 0.1$									
FLoRA	**75.47** ±2.60	88.07 ±0.18	73.14 ±2.62	69.54 ±3.59	71.13 ±1.75	77.02 ±0.21	71.56 ±1.05	61.96 ±1.91	73.49 ±0.39
FlexLoRA	73.47 ±1.02	87.60 ±0.79	72.41 ±0.17	69.53 ±0.11	71.13 ±1.33	77.37 ±0.25	71.26 ±0.13	70.67 ±0.51	74.18 ±0.54
FS-LoRA	74.69 ±1.39	87.84 ±0.62	73.50 ±4.65	69.42 ±6.00	73.67 ±2.20	79.02 ±0.89	72.13 ±1.82	61.56 ±3.76	73.98 ±0.62
FedASK	72.91 ±0.58	85.96 ±0.18	73.08 ±0.43	69.51 ±0.63	70.13 ±0.31	74.59 ±0.42	70.23 ±0.61	63.56 ±0.87	72.50 ±0.50
FedRot-LoRA	73.91 ±1.21	**88.36** ±0.37	**76.54** ±0.87	**75.99** ±1.99	**74.87** ±0.90	**79.22** ±0.80	**72.82** ±0.77	**67.01** ±0.91	**76.09** ±0.23
SAGE-LORA	74.80 ±1.35	87.13 ±0.20	71.01 ±1.97	70.39 ±3.76	72.87 ±1.70	76.15 ±0.23	70.56 ±0.87	64.72 ±1.60	73.45 ±1.26

Why heterogeneous performance remains task-dependent. FlexLoRA leads the balanced comparison at 75.43%, while FedRot-LoRA leads the heavy-tail comparison at 76.09%. SAGE-LoRA is close to FLoRA (73.49%) and below FSLoRA (73.98%) and the reported FlexLoRA result (74.18%) in the latter setting. These outcomes expose the distinction between accommodating a local rank and fully using its additional directions. A larger local adapter can learn more directions, but the fixed sketch and returned rank can still omit some of them. Moreover, fresh-residual training changes which updates clients learn. The local-capacity analysis in Appendix C.9 and the two error terms in Equation (36) distinguish these sources of loss. The 0.97- and 2.64-point gaps to the leading baselines motivate separate choices of local rank, sketch width, and returned rank, rather than increasing local rank alone.

Table 9: Rank-bucket validation accuracy (%), mean and sample SD across the stated number of seeds. Each bucket pools its validation examples; this is not the equal-task macro-average.

Method	n	Low rank	Middle rank	High rank
Balanced $\{2,4,8,16\}$*; Dirichlet* $\alpha = 10$				
FLoRA	3	$64.72_{\pm 0.41}$	$65.20_{\pm 0.14}$	$64.91_{\pm 0.77}$
FlexLoRA	2	$73.68_{\pm 0.37}$	$73.66_{\pm 0.65}$	$73.86_{\pm 0.38}$
FS-LoRA	3	$65.98_{\pm 0.40}$	$66.25_{\pm 0.47}$	$66.02_{\pm 0.16}$
FedASK	3	$69.53_{\pm 0.29}$	$70.16_{\pm 0.54}$	$70.68_{\pm 0.21}$
FedRot-LoRA	3	$72.29_{\pm 0.77}$	$72.33_{\pm 0.20}$	$72.54_{\pm 0.38}$
SAGE-LORA	3	$70.36_{\pm 0.36}$	$69.32_{\pm 0.45}$	$71.32_{\pm 0.40}$
Heavy-tail ranks 4–64; Dirichlet $\alpha = 0.1$				
FLoRA	3	$70.55_{\pm 1.97}$	$70.73_{\pm 0.93}$	$72.58_{\pm 1.16}$
FlexLoRA	3	$70.55_{\pm 0.36}$	$70.95_{\pm 0.02}$	$70.93_{\pm 0.44}$
FS-LoRA	3	$70.77_{\pm 3.37}$	$70.37_{\pm 2.49}$	$73.70_{\pm 1.62}$
FedASK	3	$70.42_{\pm 0.85}$	$70.94_{\pm 0.79}$	$73.49_{\pm 1.32}$
FedRot-LoRA	3	$75.63_{\pm 0.97}$	$75.51_{\pm 0.58}$	$76.16_{\pm 1.52}$
SAGE-LORA	3	$71.40_{\pm 1.53}$	$72.62_{\pm 1.25}$	$73.32_{\pm 0.91}$

Participation of lower-rank clients. Table 9 shows that SAGE-LoRA exceeds FLoRA and FS-LoRA across all balanced-rank buckets by 4.12–6.41 and 3.07–5.30 percentage points, respectively. Relative to FedASK, it improves the low- and high-rank buckets by 0.83 and 0.64 points, while the middle bucket is 0.84 points lower. Under heavy-tail ranks, SAGE-LoRA reaches 71.40% and 72.62% in the low- and middle-rank groups, exceeding FedASK by 0.98 and 1.68 points and FSLoRA by 0.63 and 2.25 points. Its high-rank accuracy is slightly lower than these two methods, by 0.17 and 0.38 points. FlexLoRA and FedRot-LoRA remain stronger across all balanced buckets, and FedRot-LoRA leads all heavy-tail groups. These results are consistent with SAGE-LoRA's separation of local training capacity from aggregation dimensions. Clients train paired blocks within their rank budgets, while fixed-width product sketches allow their updates to enter the same aggregate reconstruction without matching factor dimensions. This design supports useful adaptation for resource-limited clients through a common additive aggregation interface, although it does not ensure the highest accuracy in every rank group. The bucket scores pool different validation examples rather than averaging tasks equally; cross-bucket differences therefore cannot be attributed to rank alone.

F.4 Communication and Observed Computation

Table 10: Homogeneous-rank-8 cost over 100 rounds and 1,000 client-rounds. Full logical communication includes Sage's protocol-specified projection downlink. Aggregation is cumulative server time; uncertainty is three-seed sample SD.

Method	Up (GiB)	Down (GiB)	Sum (GiB)	Agg. (s)	Run (h)	Peak (GiB)
FedIT	4.059	4.059	8.118	$1.90_{\pm 0.03}$	$16.84_{\pm 0.15}$	5.48
Zero-padding	4.059	4.059	8.118	$2.48_{\pm 0.04}$	$16.82_{\pm 0.15}$	5.48
FLoRA	4.059	40.588	44.647	$11.31_{\pm 1.11}$	$26.96_{\pm 1.80}$	5.53
FlexLoRA	4.059	4.059	8.118	$34.47_{\pm 2.87}$	$27.30_{\pm 1.13}$	5.53
FS-LoRA	4.059	32.471	36.530	$21.34_{\pm 0.99}$	$29.36_{\pm 2.06}$	5.59
FedASK	3.55	11.16	14.71	$51.90_{\pm 1.43}$	$27.02_{\pm 0.95}$	5.53
FedRot-LoRA	4.059	4.059	8.118	$180.37_{\pm 2.06}$	$28.50_{\pm 0.15}$	5.53
SAGE-LoRA	4.113	10.468	14.580	$86.48_{\pm 9.71}$	$26.82_{\pm 2.07}$	5.53

Communication completes the recorded Sage subtotal using the required projection geometry. Times are observed on mixed GPUs, not controlled speedups; peak denotes allocated CUDA memory.

Downlink savings from fixed-rank output. SAGE-LoRA requires 14.580 GiB over 100 rounds, compared with 44.647 GiB for FLoRA and 36.530 GiB for FSLoRA (Table 10). Their uploads are similar; the main difference is the returned representation. FLoRA's stacked factors require 40.588 GiB of downlink, and FSLoRA's shared rank-64 factors require 32.471 GiB. SAGE-LoRA's rank-8 adapter requires 4.059 GiB, plus 6.409 GiB for the projection and metadata. Fixed-rank reconstruction therefore limits the cost of redistributing the combined client updates. FedASK already uses product sketching and has a similar full total of 14.71 GiB. FedIT, zero-padding, FlexLoRA, and FedRot-LoRA use compact factors and remain lower at 8.118 GiB; SAGE-LoRA pays additional traffic for the aggregate-dependent second measurement.

Table 11: Per-client-round logical payload accounting. Sage includes the protocol-specified projection downlink. Values use each implementation's native transport precision and are not measured wire bytes; the scalar-count comparison in Table 5 is precision independent.

Method	Upload (bytes)	Download (bytes)	Total (bytes)
FedIT	4,358,144	4,358,144	8,716,288
Zero-padding	4,358,144	4,358,144	8,716,288
FLoRA	4,358,144	43,581,440	47,939,584
FlexLoRA	4,358,144	4,358,144	8,716,288
FS-LoRA	4,358,152	34,865,160	39,223,312
FedASK	3,813,824	11,984,896	15,798,720
FedRot-LoRA	4,358,144	4,358,144	8,716,288
SAGE-LoRA	4,415,936	11,239,472	15,655,408

Upload–downlink balance. The exact totals are $15{,}655{,}408$ bytes per client-round for SAGE-LoRA and $15{,}798{,}720$ for FedASK (Table 11). SAGE-LoRA saves $745{,}424$ downlink bytes but adds $602{,}112$ upload bytes, leaving $143{,}312$ bytes, or approximately 0.91%, of net savings. This explains why its advantage over FedASK is higher reported utility at comparable traffic, rather than much smaller messages. It also makes the deployment trade-off explicit: compact returned factors reduce distribution cost, while collecting and broadcasting the measurements has its own cost.

Accounting units and timing. A client-round denotes one participant in one round; the experiment contains 1,000 client-rounds. GiB means 2^{30} bytes. Sage's upload contains 2,207,744 BF16 sketch elements plus 448 bytes of per-module count slots, totaling 4,415,936 bytes. Returned factors contain 1,089,536 FP32 elements (4,358,144 bytes); the FP32 projection and public metadata add 6,881,328 bytes. The full logical downlink is 11,239,472 bytes and the full total is 14.58 GiB. The original plaintext counter records only 8.17 GiB; Tables 10–11 and Figure 3 complete that counter using the specified projection geometry, without treating the added traffic as a new measurement. Baseline factor payloads are FP32 even though local computation is BF16, whereas Sage's sketch upload is BF16. Thus the native-byte comparison combines representation with transport precision and cannot isolate algorithmic compression. Table 5 provides the precision-independent scalar counts; no equal-precision empirical payload ablation or packet-level SecAgg measurement is reported.

Computation through low-rank factors. SAGE-LoRA records 86.48 aggregation seconds, below FedRot-LoRA's 180.37 but above FedASK's 51.90 and FlexLoRA's 34.47 (Table 10). FLoRA, FedIT, and zero-padding use stacking or summation instead of the same range-extraction and reconstruction steps. SAGE-LoRA evaluates sketches as $B(A\Omega)$ and $(\Psi B)A$ and reconstructs from a small core, avoiding dense local product matrices. The resulting cost depends on the sketch dimensions, but includes QR and a reduced decomposition. Optimized FlexLoRA also exploits factor-space algebra, explaining why avoiding dense products alone need not make SAGE-LoRA faster. SAGE-LoRA and FedASK have similar observed run times (26.82 and 27.02 hours) and the same reported peak memory (5.53 GiB). These mixed-GPU observations describe implementation costs, not controlled speedups.

Table 12: Expanded communication-budget operating point with doubled upload budget: three-seed eight-task accuracy and logical traffic over 100 rounds. This cohort is not pooled with the homogeneous-rank-8 anchor.

Method	Accuracy (%)	Up (GiB)	Down (GiB)	Total (GiB)
FLoRA (rank 16)	$72.74_{\pm 0.03}$	8.118	81.177	89.294
FS-LoRA (active rank 16)	$73.96_{\pm 0.18}$	8.118	32.471	40.588
SAGE-LoRA (rank 8)	$76.35_{\pm 0.23}$	8.225	16.876	25.101

Sage uses widths 28/40 and returns rank 8; FLoRA uses rank 16 and FSLoRA selects 16 shared coordinates. Sage includes projection traffic; the budgets are not identical.

A larger communication budget. With sketch widths $28/40$ and rank-8 output, SAGE-LoRA reaches 76.35%, compared with 72.74% for rank-16 FLoRA and 73.96% for FSLoRA with 16 active coordinates. Its total traffic is 25.101 GiB, versus 89.294 and 40.588 GiB. FLoRA's larger local rank also enlarges its stacked downlink; FSLoRA still distributes its shared factors. SAGE-LoRA instead spends additional bandwidth on measurements while retaining a small returned adapter. This supports the practical separation between the measurement budget and deployment size. Because local ranks and adaptation rules differ, the result compares these complete configurations rather than isolating the effect of increasing sketch width. No FedASK result is reported for this expanded-budget experiment.

F.5 Matched Mechanism Interventions and Training Dynamics

Table 13: Matched three-seed interventions isolate sketch geometry from the returned rank. All conditions use the same participant schedule and return rank 8; only the full bilateral variant is shaded.

Intervention	Accuracy (%)	Recon. error	Rank-8 oracle	Excess error	Up (GiB)
SAGE-LoRA (bilateral)	$76.33_{\pm 0.44}$	$0.00320_{\pm 0.00001}$	$0.00298_{\pm 0.00001}$	$0.00021_{\pm 0.00000}$	4.113
Matched random bilateral	$\mathbf{76.36}_{\pm 0.26}$	$0.00739_{\pm 0.00002}$	$0.00297_{\pm 0.00002}$	$0.00442_{\pm 0.00001}$	4.113
Left-only	$71.57_{\pm 0.15}$	$0.09781_{\pm 0.00004}$	$0.00185_{\pm 0.00003}$	$0.09596_{\pm 0.00006}$	4.113
Right-only	$71.75_{\pm 0.32}$	$0.06540_{\pm 0.00058}$	$0.00283_{\pm 0.00025}$	$0.06257_{\pm 0.00063}$	4.113

Relative Frobenius errors are averaged over module-rounds within each seed. Excess is reconstruction error minus optimal rank-8 error.

Why both measurements matter. Bilateral SAGE-LoRA achieves 76.33%, improving on right-only and left-only measurement by 4.58 and 4.76 points at the same 4.113 GiB upload and rank-8 output (Table 13). Its relative reconstruction error is 0.00320, compared with 0.06540 and 0.09781 for the two one-sided controls. The right sketch identifies a range Q from the combined updates; the left sketch supplies the coefficients $Q^\top S$ needed to reconstruct within that range. Both pieces of information matter: identifying directions is not enough without their coefficients, and the left-only control lacks the same aggregate-derived range. This supports the range-locked reconstruction in Theorem C.2 and links the gains to measurement information rather than larger output capacity. The controls still follow separate training trajectories, rather than reconstructing one frozen aggregate in several ways.

Why adaptation improves reconstruction. Compared with random bilateral measurement, adaptive SAGE-LoRA reduces relative reconstruction error from 0.00739 to 0.00320, approximately 56.7%, while the rank-8 oracle errors remain nearly equal (0.00297 and 0.00298). The corresponding excess error falls from 0.00442 to 0.00021. The adaptive probe includes the public adapter's row space, and the second projection follows the range estimated from the current aggregate. This gives reconstruction a model-dependent set of directions rather than a fixed random bilateral map. Under the conditions of Proposition C.3, the missed-range term can be bounded through the new displacement $S - M$, providing a theoretical reason to use the existing adapter in probe design. The experiment supports the fidelity of the combined geometry; it does not separate the anchor's contribution from the rest of the reconstruction procedure.

Why better matrix recovery need not raise accuracy. Adaptive and random bilateral measurements obtain nearly equal accuracies (76.33% and 76.36%). Matrix error measures all omitted directions, including those that may not change predictions on the evaluation examples. Recovering more of that matrix therefore need not improve the task score. The separate signed-cross diagnostic shows the same distinction: function error decreases in all three seeds, but NLL improves in only one, with a median NLL change of $+0.001776$. In Equation (36), range loss and rank truncation are separate squared terms; the table's excess is instead a difference of relative Frobenius errors. The evidence supports more faithful rank-constrained recovery, not a general accuracy guarantee from lower matrix error. The extra six completion rows are discarded by the exact decoder and are not a source of this gain.

Training after repeated aggregation. In Figure 7, SAGE-LoRA and FedASK reach a low commonsense training-loss region after the initial rounds. On CoLA, both remain in the lower-loss group with FlexLoRA and FedRot-LoRA, while FLoRA and FSLoRA settle at higher losses. Thus, SAGE-LoRA's compact reconstruction does not prevent continued local optimization after repeated global updates. Each reconstructed adapter supplies the next round's starting state in these complete-update runs, so retaining useful update directions supports further learning. These curves complement the reconstruction results and the separation of update errors in Theorem 4.4. They measure local training loss, not the theorem's expected gradient norm, and hence do not verify its convergence rate.

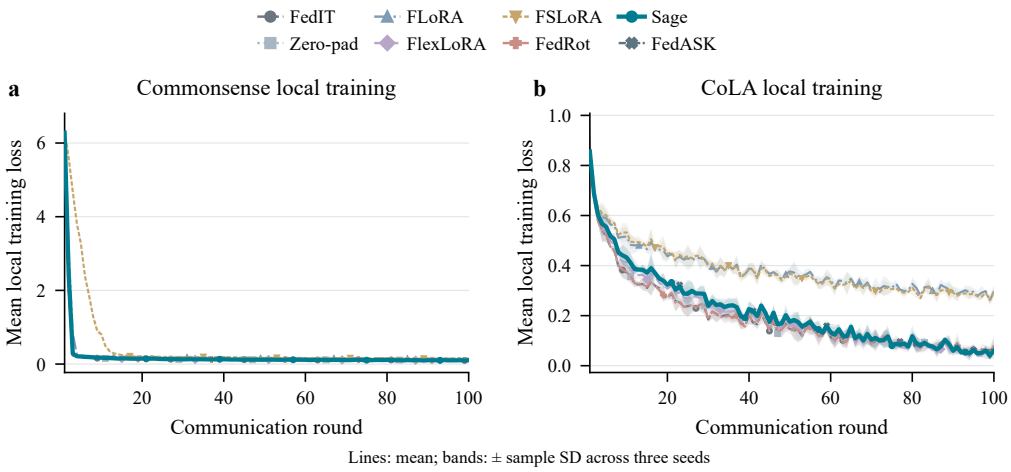

Figure 7: Local training losses for the homogeneous commonsense cohort and CoLA transfer. Lines are per-round means; translucent bands are sample standard deviations over three independent random seeds.

F.6 Secure Aggregation, Differential Privacy, and Selective Encryption

Accuracy with protected aggregation. Secure SAGE-LoRA achieves 76.37%, compared with 75.64% for FedASK-DP and 76.47% for selective CKKS (Table 14). The central participant-DP reference is lower at 51.36%. SAGE-LoRA first evaluates paired-product sketches locally and only then applies additive masks. Since protection acts on linear measurements, their sum remains the intended product sketch after unmasking, subject to numerical encoding error. This design explains the small change from plaintext and supports the measurement identity in Equation (15). It provides a way to retain two-factor adaptation while protecting individual contributions without adding DP noise as part of the aggregation operation.

Table 14: Utility and communication under distinct security mechanisms (three seeds). Accuracy is the eight-task macro-average. Plaintext Sage reports a projection-excluding subtotal; SecAgg reports a projection-inclusive logical total. The pair shares the product-sketch realization; DP and SHE are separate privacy/workload settings.

Mechanism	Accuracy (%)	Up (GiB)	Down (GiB)	Total (GiB)
SAGE-LoRA (plaintext)	$76.27_{\pm 0.41}$	4.11	4.06	8.17
SAGE-LoRA + SecAgg	$76.37_{\pm 0.52}$	4.11	10.47	14.58
Central participant-DP	$51.36_{\pm 0.09}$	4.06	4.06	8.12
FedASK (DP)	$75.64_{\pm 0.40}$	3.55	11.16	14.71
SHE-LoRA (CKKS)	$76.47_{\pm 0.30}$	56.36	20.53	76.89

The plaintext subtotal omits projection traffic; its gap from SecAgg does not isolate cryptographic overhead. DP/SecAgg use logical payload; SHE includes serialized ciphertext. Guarantees and wire costs are not equated.

Table 15: Security scope and observed runtime, with sample standard deviations over three independent random seeds. Server time includes each method's aggregation pipeline and is not an isolated cryptographic-kernel benchmark.

Mechanism	Server (s)	Run (h)	Scope
SAGE-LoRA (plaintext)	$86.5_{\pm 9.7}$	$26.82_{\pm 2.07}$	No aggregate confidentiality
SAGE-LoRA + SecAgg	$11701.5_{\pm 696.1}$	$30.72_{\pm 1.43}$	Two-channel Flower reference
Central participant-DP	$17.7_{\pm 0.2}$	$15.93_{\pm 0.04}$	$(\epsilon, \delta) = (8, 10^{-5})$
FedASK (DP)	$51.9_{\pm 1.4}$	$27.02_{\pm 0.95}$	$(\epsilon, \delta) = (8, 10^{-5})$
SHE-LoRA (CKKS)	$27205.8_{\pm 135.0}$	$25.98_{\pm 0.21}$	Selective encrypted workload

Central DP uses replace-one adjacency, fixed-size sampling without replacement, clipping 1, and noise multiplier 2.7027826. SecAgg, DP, and selective CKKS provide different guarantees.

Different protection objectives and costs. FedASK-DP substantially outperforms the central participant-DP reference even though both report $(\epsilon, \delta) = (8, 10^{-5})$. The low central-DP score therefore cannot be generalized to all private LoRA training. Clipping and noise alter a training signal, whereas SecAgg restricts visibility of individual contributions; equal nominal DP budgets also require matching adjacency and accounting before guarantees can be equated. For SAGE-LoRA, the measured protection layer increases server time from 86.5 to $11{,}701.5$ seconds and run time from 26.82 to 30.72 hours (Table 15). FedASK-DP uses 51.9 server seconds and 27.02 run hours, so SAGE-LoRA's utility retention does not imply lower runtime. These are mixed-hardware implementation costs. Appendix E.2 also distinguishes the measured Flower reference from the formal requirement to bind the cohort before releasing the second aggregate.

Table 16: Native CKKS component times (seconds), mean and sample SD over three runs of the fixed selective-encryption workload. Per-round components are accumulated over the full run.

Component	Time (s)
Client encryption	$1475.90_{\pm 2.19}$
Client decryption	$89.18_{\pm 0.61}$
Server encrypted product	$10987.86_{\pm 61.60}$
Server ciphertext aggregation	$139.14_{\pm 0.54}$
Public-context setup (once)	$0.62_{\pm 0.00}$

The one-time public context, excluded from model-update totals, is 32.49 ± 0.003 MiB.

Avoiding encrypted server-side products. Selective CKKS spends $10{,}987.86$ seconds on encrypted products, compared with 139.14 seconds on ciphertext summation (Table 16). This identifies a costly operation that SAGE-LoRA removes from the protected server computation. Clients evaluate the product sketches through low-rank factor chains before masking; the server securely combines additive measurements and reconstructs after their aggregate is released. The practical benefit is the ability to use an additive SecAgg backend without homomorphic matrix multiplication. The listed CKKS components do not exhaust its server time, and its serialized ciphertext traffic differs in scope from SAGE-LoRA's logical tensor accounting, so these values are not an end-to-end speedup or equal-protection bandwidth measurement.

Paired effect of the security layer. The three plaintext/SecAgg pairs in Figure 8 change by only $+0.10 \pm 0.32$ points on average, with changes in both directions. This isolates protection more directly than comparisons between different training methods: the paired runs share a seed and training protocol. It supports retaining the plaintext learning behavior when masking the linear sketches, rather than an accuracy improvement caused by masking. The displayed 8.17 GiB plaintext subtotal omits projection traffic, while the 14.58 GiB secure total includes it. Both require the projection, so the difference is not cryptographic overhead.

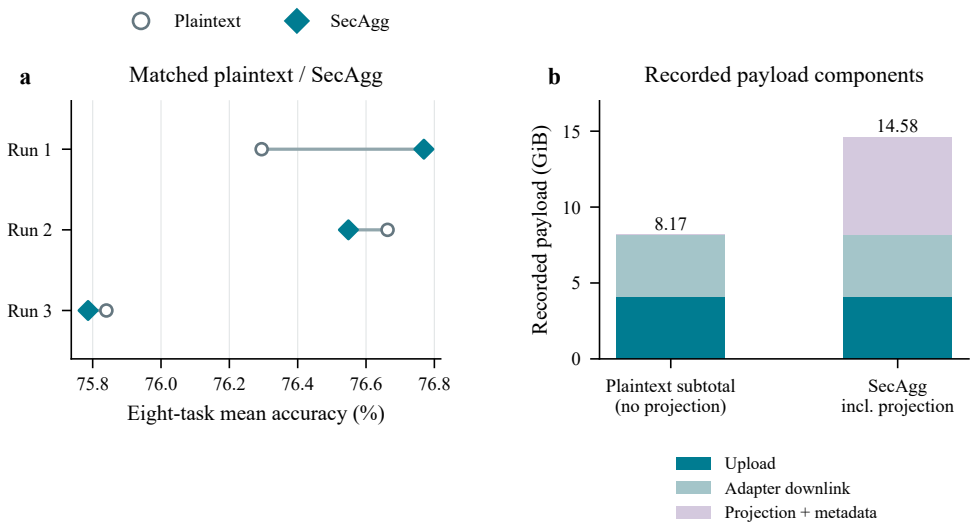

Figure 8: Paired plaintext/SecAgg utility and payload accounting for Sage. Each pair uses the same random seed and training protocol. The plaintext subtotal excludes Stage-II projection traffic; SecAgg includes it, so their payload difference does not isolate cryptographic overhead. Accuracy differences are small and mixed in sign across independent runs.

F.7 Five-Seed Confirmation and Scope

Table 17: Five-seed confirmation of the secure operating point: all eight task accuracies and the within-seed macro-average (%). Means and sample standard deviations are computed over five independent random seeds.

Method	ARC-c	ARC-e	BoolQ	HellaS.	OBQA	PIQA	SIQA	WinoG.	Avg.
FLoRA	74.98 ±0.76	87.26 ±0.27	70.94 ±0.38	59.92 ±0.47	68.44 ±0.64	75.88 ±0.18	70.33 ±0.24	60.35 ±0.31	71.01 ±0.10
FS-LoRA	**75.59** ±0.63	87.54 ±0.25	71.49 ±0.36	59.72 ±0.77	68.12 ±0.41	76.39 ±0.19	70.21 ±0.24	61.64 ±0.41	71.34 ±0.09
FedASK	73.44 ±0.38	87.68 ±0.34	77.65 ±0.38	73.65 ±0.64	73.64 ±0.55	77.83 ±0.42	73.19 ±0.45	68.54 ±0.52	75.70 ±0.34
SAGE-LoRA + SecAgg	74.05 ±0.56	**87.89** ±1.21	**77.75** ±0.49	**74.77** ±0.40	**74.92** ±1.04	**78.56** ±0.39	**73.42** ±0.39	**69.85** ±0.53	**76.40** ±0.37

Performance across the expanded seed set. SAGE-LoRA with SecAgg obtains $76.40 \pm 0.37\%$, exceeding FedASK's $75.70 \pm 0.34\%$ by approximately 0.70 points (Table 17). The gain remains close to the three-seed secure comparison (0.73 points), with higher means on all eight tasks. The largest gains are on WinoGrande (1.31 points), OpenBookQA (1.28), and HellaSwag (1.12). Improvements over FLoRA and FSLoRA also persist at 5.39 and 5.06 points. Together, the task coverage and similar macro differences show that the reported advantage persists when the seed set is expanded. This supports the combined design of compact product reconstruction and linear protected aggregation, rather than a contribution assigned to any one step.

Scope of the evidence. The experiments cover Qwen2.5-1.5B-Instruct, the eight-task commonsense suite, and CoLA, SST-2, MRPC, QNLI, and RTE. They show useful compact adaptation, improved reconstruction fidelity, and small paired utility changes under SecAgg. Likewise, utility tests do not prove confidentiality: that guarantee follows from the stated SecAgg assumptions, and the measured reference retains the implementation boundaries in Appendix E.2.

G COMPLETE ROUND PROCEDURE

Algorithm 2 One communication round of SAGE-LORA

Input: Round-t global factors (B^t, A^t), target rank k, paired-block pool, client budgets, sketch widths $(s_{\mathrm{R}}, s_{\mathrm{L}})$, update mode, server coefficient η_{srv}, and secure aggregation.
Output: Updated balanced factors (B^{t+1}, A^{t+1}) with $M^{t+1} = B^{t+1}A^{t+1}$, or $\perp$.

Set $M^t \leftarrow B^t A^t$. For complete updates set $P^t \leftarrow 0$, $\eta_{\mathrm{srv}} \leftarrow 1$; for residual updates set $P^t \leftarrow M^t$.
Stage I: right-sketch aggregation.
Fix one feasible layer orientation for both communication directions. Distribute (B^t, A^t), mode, and the seed defining the probe anchored to the oriented public factors (random if $M^t = 0$).
for each participating client i in parallel **do**
    Select budget-feasible paired blocks. Initialize complete updates by Equation (20), or initialize fresh blocks on frozen M^t for residual updates; fine-tune using Equation (7); freeze the resulting block factors and compute Y_i^t by Equation (9).
end for
Securely aggregate the sample-weighted right sketches to obtain $(\mathcal{U}_t, \bar{Y}^t)$.
if the right-sketch aggregation fails **then**
    return $\perp$
end if
Form $Y^t \leftarrow P^t\Omega^t + \eta_{\mathrm{srv}}\bar{Y}^t$; estimate $\widehat{r}_t$.
if $\widehat{r}_t < k$ **then**
    return $\perp$ before Stage II; retry only with fresh state
end if
Extract Q_t, complete H_t, and set $\Psi^t \leftarrow [Q_t, H_t]^\top$, $G_t \leftarrow \Psi^t Q_t$.
Stage II: range-locked left-sketch aggregation.
Server distributes Ψ^t to the accepted survivor set $\mathcal{U}_t$.
for each $i \in \mathcal{U}_t$ in parallel **do**
    Compute Z_i^t by Equation (12) from the same frozen paired-block factors.
end for
Invoke Stage-II secure aggregation with an authenticated binding to $\mathcal{U}_t$ and its protected aggregate count. The protocol releases $\bar{Z}^t$ only if the binding verifies.
if Stage II aborts before releasing $\bar{Z}^t$ **then**
    return $\perp$
end if
Form $Z^t \leftarrow \Psi^t P^t + \eta_{\mathrm{srv}}\bar{Z}^t = \Psi^t S^t$.
Compute $X^t \leftarrow G_t^\dagger Z^t$ and $[X^t]_k = U_{X,k}\Sigma_{X,k}V_{X,k}^\top$.
Set $B^{t+1} \leftarrow Q_t U_{X,k}\Sigma_{X,k}^{1/2}$ and $A^{t+1} \leftarrow \Sigma_{X,k}^{1/2}V_{X,k}^\top$.
return (B^{t+1}, A^{t+1})