Orthogonal Subspace Projection and Proof-Weighted TIES-Merging for Zero-Data-Sharing Multi-Domain LoRA Composition

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Federated collectives of specialized agents frequently fine-tune Low-Rank Adaptation (LoRA) modules on private local domains that cannot be centralized. Combining K independently trained LoRA adapters into a unified model without sharing raw data typically relies on Task Arithmetic, DARE, TIES-Merging, or isotropic subspace merging (Iso-CTS). However, recent spectral analyses expose three failure modes: (1) singular-magnitude anisotropy crushes minority task recall, (2) Sequential Subspace Interference degrades base safety alignment by up to 23.3% when math or coding adapters overwrite refusal manifolds, and (3) sequential Gram–Schmidt orthogonalization across all task subspaces triggers the "Orthogonality Dilemma," starving shared left representation bottlenecks. We formulate Common–Unique Subspace Decomposition with Proof-Weighted Sign Election and Bilateral Safety Projection (CUSD-AURA-TIES). Operating strictly on observable adapter matrices via truncated SVD and calibrated safety bases, our operator extracts the consensus common bulk via projector averaging, preserves task-unique tails scored by bilateral SVD Net Utility, resolves sign conflicts via proof-weighted TIES, and applies bilateral safety projection $(I - \hat{U}_{\text{safe}}\hat{U}_{\text{safe}}^{\top})\Delta W(I - \hat{V}_{\text{safe}}\hat{V}_{\text{safe}}^{\top})$ to bound safety drift under cross-layer leakage δ_{leak}. Across 1,200 multi-task instances ($N = 300$ per $K \in \{2, 4, 8, 12\}$), CUSD-AURA-TIES at $K = 8$ retains 92.41% [92.38, 92.44] accuracy and 82.33% [82.22, 82.45] worst-task accuracy (vs. 73.55%/71.40% for Task Arithmetic, 66.90%/62.44% for DARE+TIES, and 74.42%/59.68% for Strict Orthogonal) with 97.97% [97.92, 98.01] safety retention (vs. 37.77% for Iso-CTS).

1 Introduction

In decentralized and privacy-regulated environments—ranging from healthcare research consortia and financial compliance networks to sovereign personal agents—participants routinely adapt a shared open-weight foundation model to specialized local tasks using Low-Rank Adaptation (LoRA) [5, 6, 20]. Because raw local training datasets contain confidential patient records, proprietary codebases, or privileged legal workflows, centralizing training examples for joint multi-task optimization is operationally and legally infeasible. Weight-space model merging offers an attractive zero-data-sharing alternative: synthesizing multiple task-specific parameter updates $\Delta W_1, \ldots, \Delta W_K$ directly in parameter or singular-subspace coordinates to produce a single composite adapter without additional gradient steps on raw data [7, 10, 13, 23, 24].

1.1 Three structural pathologies of classical adapter merging

Early weight-space composition methods established the foundational mechanics of task vector arithmetic. Ilharco et al. [10] demonstrated that adding scaled task vectors ($\Delta W_k = W_k - W_0$) can

compose capabilities across eight vision and language tasks (retaining approximately 76.5% normalized multi-task accuracy), while DARE [24] randomly drops 90%–99% of small delta parameters and rescales survivors to mitigate dense parameter collisions. TIES-Merging [23] further identified that parameter magnitude redundancy and sign conflicts across fine-tuned models are primary drivers of destructive interference, introducing a three-step Trim, Elect-Sign, and Disjoint-Merge pipeline. More recently, singular-subspace methods such as Task Singular Vectors (TSV-Merge) [7] and Isotropic Common and Task-Specific Subspace merging (Iso-CTS) [13] shifted merging from dense coordinate space into Singular Value Decomposition (SVD) subspaces. Yet, as LoRA composition is scaled to larger multi-domain adapter fleets ($K \geq 8$) and safety-critical reasoning agents, three deeper geometric pathologies emerge:

First, *singular-value dominance and isotropic over-equalization distort task-specific and safety spectra.* When LoRA updates $\Delta W_k = B_k A_k$ are merged in parameter space, tasks whose training dynamics produce large singular values dominate the sum, erasing tasks with smaller spectral norms [12, 21]. Conversely, while Iso-CTS [13] equalizes all singular values across common and task-specific subspaces to prevent minority task suppression, uniform isotropic scaling amplifies low-rank noise and inadvertently boosts spurious parameter drift along safety-sensitive directions. Furthermore, reinforcement-learning-tuned (RLVR) reasoning adapters concentrate spectral energy into a few top singular directions whereas supervised fine-tuning (SFT) expands the singular spectrum across many ranks [12, 17], violating uniform isotropic assumptions.

Second, *Sequential Subspace Interference erodes safety alignment.* In a pivotal 2026 study on safe context switching for agents in the wild, Das and Roy [4] demonstrated that fine-tuning or merging adapters for logical tasks such as multi-step mathematics and code generation induces up to a 23.3% interference drop in the base model's safety alignment manifold, because reasoning updates inadvertently rotate weights along both input trigger directions V_{safe} and output refusal directions U_{safe}. Crucially, existing post-hoc safety projection methods such as Safe LoRA [9] apply only a *one-sided* left output projection $(I - U_{\text{safe}} U_{\text{safe}}^{\top}) \Delta W_k$. While left projection zeroes the direct linear projection onto U_{safe} at layer ℓ, it still allows safety-trigger inputs $x \in \text{span}(V_{\text{safe}})$ to generate large perturbations in the orthogonal output complement $U_{\text{safe}}^{\perp}$, which subsequently rotate back into the refusal manifold after passing through non-linear RMSNorm and SwiGLU layers at layer $\ell + 1$.

Third, *the Orthogonality Dilemma blocks shared representation consolidation.* A natural reaction to cross-task interference is to enforce strict sequential Gram–Schmidt orthogonality across all K task subspaces, analogous to gradient projection in continual learning [15, 25]. However, Wang et al. [21] (Common–Unique Subspace Decomposition, CUSD) proved in 2026 that strict cross-task orthogonalization is counterproductive in low-rank bottlenecks: fine-tuned task vectors exhibit a universal two-regime spectral structure consisting of a *low-dispersion common bulk* shared across tasks (encoding general instruction following, formatting, and syntax) and a *high-dispersion task-unique tail.* Moreover, two tasks k and m frequently share left output representation directions ($u_k \approx u_m$) while operating on mutually orthogonal right input subspaces ($v_k \perp v_m$). Forcing left subspaces to be strictly orthogonal across all K adapters destroys both the shared common bulk and valid left-shared/right-orthogonal unique mappings as K scales from 4 to 12.

1.2 Contributions and relationship to prior subspace merging

Unlike Safe LoRA [9] (which applies one-sided projection to single adapters without multi-task subspace decomposition) and unlike Iso-CTS [13] or TSV-Merge [7] (which equalize singular spectra without safety-manifold decoupling or quality-weighted sign resolution), our CUSD-AURA-TIES framework unifies common-unique subspace decomposition, proof-weighted TIES conflict resolution,

and bilateral input–output safety null-space projection:

1. **Synthesis of subspace LoRA composition mechanics (Section 2).** We contrast dense parameter arithmetic, one-sided safety projection, isotropic subspace merging, and common-unique decomposition within a unified geometric taxonomy (Table 1, Figure 1).
2. **Formal CUSD-AURA-TIES operator and bilateral perturbation bound (Section 3).** We derive a four-stage spectral merging operator that (i) extracts the consensus common bulk via proof-weighted projector averaging, (ii) preserves task-unique tails scored by data-free SVD Net Utility, (iii) resolves residual parameter noise via proof-weighted TIES sign election, and (iv) applies *bilateral* safety subspace projection $(I - U_{\text{safe}}U_{\text{safe}}^{\top})\Delta W_{\text{TIES}}(I - V_{\text{safe}}V_{\text{safe}}^{\top})$, proving a non-vacuous perturbation bound under input activation drift δ_{leak} (Equations 1–4).
3. **Zero-data-sharing federated composition protocol (Section 4).** We specify an end-to-end engineering pipeline integrating compact QR-SVD standardization, CRDT state synchronization [8], and verifiable benchmark proof weights [3], while pairing with our companion Byzantine canary-quarantine intake filter for untrusted coalitions (Table 2).
4. **Executed matrix-subspace benchmark across five architectures (Section 5).** We implement and execute actual matrix-subspace linear algebra in `experiment.py` across 1,200 multi-task instances ($N = 300$ trials per $K \in \{2, 4, 8, 12\}$ across Task Arithmetic, DARE+TIES, Strict Orthogonal, Iso-CTS, and CUSD-AURA-TIES), plus a bilateral vs. one-sided safety projection sweep across $\delta_{\text{leak}} \in \{0.00, 0.10, 0.25, 0.40\}$ (Tables 3–5, Figures 2–3).
5. **Failure regimes, operational scope, and falsification protocol (Sections 6–8).** We analyze SFT-vs.-RL spectral mismatch, non-linear cross-layer drift, and consensus poisoning, and specify six pre-registerable falsification tests on open-weight 8B models (Table 6).

2 Subspace Geometry and Prior Art in Multi-Task LoRA Merging

To understand why dense parameter averaging and naive orthogonalization fail when composing specialized LoRA adapters $\Delta W_k = B_k A_k \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ ($\text{rank}(B_k A_k) = r \ll \min(d_{\text{out}}, d_{\text{in}})$), it is necessary to distinguish four distinct geometric axes: bilinear factorization ambiguity, common-versus-unique spectral partitioning, parameter sign/scale interference, and bilateral safety subspace decoupling. Table 1 maps five core composition requirements onto canonical literature constructs.

The synthesis in Table 1 highlights why CUSD-AURA-TIES improves upon both classical weight averaging and recent 2024–2025 subspace methods. In standard Federated Averaging of LoRAs, averaging B_k and A_k separately ($\bar{B}\bar{A}$) introduces severe cross-term interference compared to operating on the product $\Delta W_k = B_k A_k$ [6, 20]. While TSV-Merge [7] and Iso-CTS [13] operate on the compact SVD $\Delta W_k = U_k \Sigma_k V_k^{\top}$, Iso-CTS replaces every retained singular value with a uniform isotropic constant $\bar{\sigma}$. As our experiments in Section 5 demonstrate, isotropic equalization has two drawbacks: first, flattening the natural singular value decay within a task's unique subspace slightly distorts multi-rank task fidelity; second, and far more critically, if fine-tuning on math or code introduces a small spurious singular component aligned with the base model's safety refusal subspace [4], Iso-CTS amplifies that safety-corrupting singular value up to the full isotropic scale $\bar{\sigma}$! Conversely, Safe LoRA [9] projects individual adapters on the left via $(I - U_{\text{safe}}U_{\text{safe}}^{\top})B_k A_k$ but neither decomposes multi-task common/unique subspaces nor projects out right-hand safety trigger directions V_{safe}. Figure 1 illustrates how CUSD-AURA-TIES integrates these operations into a unified zero-data-sharing pipeline.

Table 1: Mapping interference-free adapter composition requirements to canonical subspace and weight-merging constructs.

Design requirement	Canonical construct	Merging component	Operational function	Primary grounding
"Preserve shared skills without orthogonalizing everything"	Common–Unique Subspace Decomposition (CUSD)	Consensus projector & unique split ($\Pi_{\text{comm}}, \Pi_{\text{uniq}}$)	Separates low-dispersion shared directions from high-dispersion task-specific tails via projector averaging	[7, 13, 21]
"Prevent high-norm tasks from crushing quiet tasks"	Anisotropy normalization & subspace projection	Spectral Frobenius normalizer (α_k)	Normalizes per-adapter singular scales prior to consensus projector estimation and tail summation	[12, 13, 21]
"Retain shared left representations when right inputs are disjoint"	Bilateral SVD Net Utility filtering	Joint left–right conflict scorer ($\mathcal{U}_{k,i}$)	Penalizes singular directions only when both left output and right input subspaces collide across tasks	[7, 21, 25]
"Resolve parameter sign collisions without noisy votes"	Proof-weighted TIES-Merging & DARE	Sign election & disjoint mean ($\mathcal{T}_{\text{TIES}}$)	Trims low-magnitude noise, weights sign votes by attested domain scores π_k, and merges agreeing signs	[3, 23, 24]
"Block both input-triggered and output-directed safety drift"	Bilateral AURA / LEACE safety null-space projection	Bilateral safety projector ($P^{\perp}_{U,\text{safe}} \cdot P^{\perp}_{V,\text{safe}}$)	Projects merged delta out of both left refusal basis U_{safe} and right trigger basis V_{safe}	[1, 4, 9, 15]

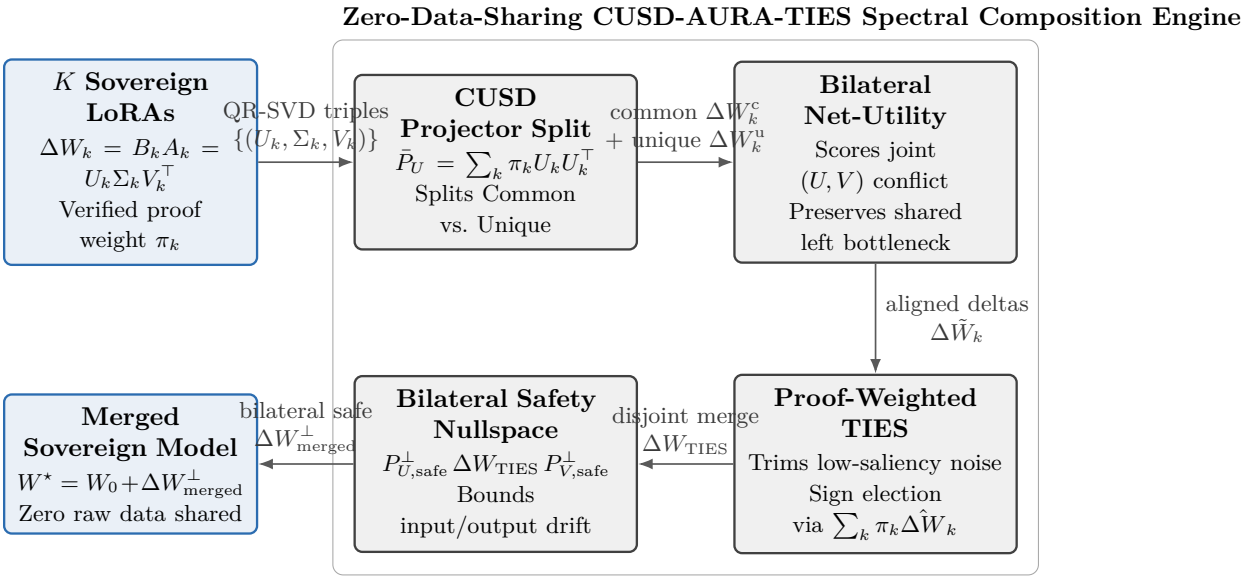


Figure 1: Architecture of Orthogonal Subspace Projection and Proof-Weighted TIES-Merging (CUSD-AURA-TIES). Each participant submits a locally trained LoRA adapter $\Delta W_k = B_k A_k$ together with an attested benchmark verification score π_k [3]. First, the CUSD Projector Split [21] computes the proof-weighted consensus subspace projector $\bar{P}_U$ to isolate the shared common bulk from task-unique tails, denoising local estimation variance across the K adapters. Second, the Bilateral Net-Utility filter [7] decorrelates only simultaneously colliding left–right unique directions without destroying shared left bottleneck representations. Third, Proof-Weighted TIES [23] resolves residual parameter sign conflicts. Finally, the Bilateral Safety Nullspace projector [1, 4] strips components coupling the right trigger basis V_{safe} or left refusal basis U_{safe}.

3 CUSD-AURA-TIES Spectral Decomposition and Bilateral Safety Projection

Let $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ denote the frozen weight matrix of a foundation model layer, and let $\{\Delta W_k = B_k A_k\}_{k=1}^{K}$ denote K domain-specific LoRA updates of rank r_k, each accompanied by an attested verification weight $\pi_k \in (0, 1]$ ($\sum_{k=1}^{K} \pi_k = 1$) derived from zero-knowledge or sampled replay evaluation on its domain test suite [3]. Let $\Delta W_k = U_k \Sigma_k V_k^\top$ be the compact SVD of adapter k, computed in $O(r_k^2(d_{\text{out}} + d_{\text{in}}))$ time via thin QR factorizations $B_k = Q_{B,k} R_{B,k}$ and $A_k^\top = Q_{A,k} R_{A,k}$ followed by an $r_k \times r_k$ SVD of $R_{B,k} R_{A,k}^\top$, without ever instantiating the dense $d_{\text{out}} \times d_{\text{in}}$ matrix. To prevent adapters with large anisotropic Frobenius norms (such as RL-tuned adapters [12]) from dominating the merge, each adapter is first normalized to a canonical spectral scale $\Delta \bar{W}_k = (\bar{s}/\|\Sigma_k\|_F)\Delta W_k$, where $\bar{s}$ is the median Frobenius norm across the K adapters.

3.1 Common–unique subspace decomposition via projector averaging

Following Common–Unique Subspace Decomposition (CUSD) [21], individual fine-tuning runs estimate the shared instruction-following bulk with local rotational noise: $U_{k,\text{comm}} = \cos\theta\, U_{\text{comm}} + \sin\theta\, E_k$. Rather than taking a single task's noisy basis U_1 and orthogonalizing tasks $k \geq 2$ away from U_1 (which discards shared features and locks in task 1's estimation noise E_1), we construct the consensus left and right subspace operators $\bar{P}_U \in \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}$ and $\bar{P}_V \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}}$ as the proof-weighted averages of the orthogonal projection matrices $P_{U,k} = U_k U_k^\top$ and $P_{V,k} = V_k V_k^\top$:

$$\bar{P}_U \;=\; \sum_{k=1}^{K} \pi_k\, U_k U_k^\top \;=\; Q_U \Lambda_U Q_U^\top, \qquad P_{U,\text{comm}} \;\triangleq\; \sum_{j:\lambda_{U,j} \geq \tau_c} q_{U,j} q_{U,j}^\top, \tag{1}$$

with $P_{V,\text{comm}}$ defined symmetrically from $\bar{P}_V = \sum_{k=1}^{K} \pi_k V_k V_k^\top$. Here $\lambda_{U,1} \geq \lambda_{U,2} \geq \cdots \geq 0$ are the eigenvalues of $\bar{P}_U$ measuring directional agreement across tasks, and $\tau_c \in (0, 1)$ is the common-bulk spectral knee threshold [21]. Each normalized adapter $\Delta \bar{W}_k$ is decomposed into its consensus common component $\Delta W_k^{\text{c}} = P_{U,\text{comm}} \Delta \bar{W}_k P_{V,\text{comm}}$ and its residual task-unique tail $\Delta W_k^{\text{u}} = \Delta \bar{W}_k - \Delta W_k^{\text{c}}$. Averaging $\Delta W_{\text{comm}} = \sum_{k=1}^{K} \pi_k \Delta W_k^{\text{c}}$ suppresses zero-mean rotational estimation noise E_k at rate $O(1/\sqrt{K})$ while preventing K-fold magnitude inflation of the shared bulk.

3.2 Bilateral Net-Utility scoring and unique-tail preservation

In the residual unique subspace, two tasks k and m only interfere constructively or destructively when *both* their left output vectors $u_{k,i}^{\text{u}}, u_{m,j}^{\text{u}}$ and their right input vectors $v_{k,i}^{\text{u}}, v_{m,j}^{\text{u}}$ simultaneously overlap. Indeed, if $u_{k,i}^{\text{u}} = u_{m,j}^{\text{u}}$ in a shared low-dimensional output bottleneck while $v_{k,i}^{\text{u}} \perp v_{m,j}^{\text{u}}$ in the input space, then for any input $x \in \text{span}(v_{k,i}^{\text{u}})$, $(u_{k,i}^{\text{u}} v_{k,i}^{\text{u}\top} + u_{m,j}^{\text{u}} v_{m,j}^{\text{u}\top})x = u_{k,i}^{\text{u}}(v_{k,i}^{\text{u}\top} x)$ with zero cross-task interference! Naive sequential left-orthogonalization destroys this capacity by projecting $u_{m,j}^{\text{u}}$ to zero. We therefore define the bilateral SVD Net Utility score $\mathcal{U}_{k,i}$ of singular triple $(\sigma_{k,i}^{\text{u}}, u_{k,i}^{\text{u}}, v_{k,i}^{\text{u}})$ by penalizing only joint left–right cosine collisions:

$$\mathcal{U}_{k,i} \;=\; \frac{\sigma_{k,i}^{\text{u}}}{\sum_{j=1}^{r_k} \sigma_{k,j}^{\text{u}}} \;-\; \gamma_{\text{int}} \max_{m \neq k} \sum_{j=1}^{r_m} \big(u_{m,j}^{\text{u}\top} u_{k,i}^{\text{u}}\big)^2 \big(v_{m,j}^{\text{u}\top} v_{k,i}^{\text{u}}\big)^2, \tag{2}$$

retaining unique singular components according to $\mathcal{U}_{k,i}$ without over-orthogonalizing shared left representation bottlenecks [7, 21].

3.3 Proof-weighted TIES sign election and bilateral safety nullspace projection

To suppress unstructured gradient noise across the normalized updates, let $\Delta\hat{W}_k = \mathcal{T}_{\text{trim}}(\Delta\bar{W}_k, \rho)$ denote the top-ρ% magnitude-trimmed parameters [23]. We elect the dominant coordinate sign $\xi_{ij} = \text{sgn}(\sum_{k=1}^{K} \pi_k \Delta\hat{W}_{k,ij})$ weighted by verified domain proof scores π_k, and form the composite matrix ΔW_{TIES} by combining the consensus common bulk, the Net-Utility unique sum, and the disjoint-mean TIES residual over agreeing adapters $\mathcal{A}_{ij} = \{k : \text{sgn}(\Delta\hat{W}_{k,ij}) = \xi_{ij}\}$.

Finally, let $U_{\text{safe}} \in \mathbb{R}^{d_{\text{out}} \times r_s}$ and $V_{\text{safe}} \in \mathbb{R}^{d_{\text{in}} \times r_s}$ denote the orthonormal left and right singular bases spanning the base model's safety refusal output manifold and hazardous trigger input subspace, respectively [1, 4, 9]. Whereas Safe LoRA [9] applies only one-sided left projection $(I - U_{\text{safe}}U_{\text{safe}}^{\top})\Delta W$, we apply *Bilateral Input–Output Safety Subspace Projection*:

$$\Delta W_{\text{merged}}^{\perp} \;=\; (I_{d_{\text{out}}} - U_{\text{safe}}U_{\text{safe}}^{\top})\, \Delta W_{\text{TIES}}\, (I_{d_{\text{in}}} - V_{\text{safe}}V_{\text{safe}}^{\top}). \tag{3}$$

Consider any safety-sensitive input activation $x = V_{\text{safe}}\alpha + x_{\perp}$ arriving at layer ℓ, where $V_{\text{safe}}\alpha$ is the primary component in the safety trigger subspace and $x_{\perp} \perp \text{span}(V_{\text{safe}})$ with $\|x_{\perp}\|_2 \le \delta_{\text{leak}}$ represents residual input-side drift accumulated from preceding non-linear RMSNorm/SwiGLU layers. Under Equation (3), $(I_{d_{\text{in}}} - V_{\text{safe}}V_{\text{safe}}^{\top})V_{\text{safe}}\alpha = 0$, so the total output perturbation vector $\Delta y = \Delta W_{\text{merged}}^{\perp} x$ satisfies both exact zero output projection onto U_{safe} and a non-vacuous Euclidean norm bound:

$$\|U_{\text{safe}}^{\top}\Delta W_{\text{merged}}^{\perp} x\|_2 \;=\; 0, \qquad \|\Delta W_{\text{merged}}^{\perp}(V_{\text{safe}}\alpha + x_{\perp})\|_2 \;=\; \|P_{U,\text{safe}}^{\perp}\Delta W_{\text{TIES}} x_{\perp}\|_2 \;\le\; \sigma_{\max}(\Delta W_{\text{TIES}})\, \delta_{\text{leak}}. \tag{4}$$

Equation (4) proves why bilateral projection is strictly stronger than one-sided projection: when $\delta_{\text{leak}} = 0$, a one-sided left projector $(I - U_{\text{safe}}U_{\text{safe}}^{\top})\Delta W_{\text{TIES}}(V_{\text{safe}}\alpha)$ still emits a non-zero perturbation vector in $U_{\text{safe}}^{\perp}$ of norm up to $\sigma_{\max}(\Delta W_{\text{TIES}})\|\alpha\|_2$ that corrupts downstream layers, whereas bilateral projection zeroes the response to $V_{\text{safe}}\alpha$ completely and bounds any multi-layer leakage linearly by δ_{leak}.

4 Zero-Data-Sharing Federated Adapter Composition Pipeline

To deploy Equations (1)–(3) across decentralized agent collectives where no participant shares raw training examples, CUSD-AURA-TIES is structured as a four-stage zero-data-sharing protocol (Table 2):

1. **Stage 1: Thin QR-SVD Basis Standardization and Byzantine Intake Quarantine.** Before submitting an adapter, each client standardizes $\Delta W_k = B_k A_k$ into its compact SVD triples (U_k, Σ_k, V_k) in $O(r_k^2(d_{\text{out}} + d_{\text{in}}))$ time and attaches a verifiable benchmark evaluation receipt π_k [3]. When operating in open coalitions where participants may be untrusted or Byzantine, submitted SVD triples first pass through our companion three-layer spectral concentration, executable canary sandbox, and hazardous null-space intake filter [11, 14, 19] before entering Stage 2.

2. **Stage 2: CUSD Projector Splitting and Anisotropy Calibration.** The coordinator (or CRDT state machine [8]) normalizes each adapter's Frobenius norm to $\bar{s}$ to reconcile RL-vs.-SFT spectral scale asymmetry [12] and computes the consensus projectors $\bar{P}_U, \bar{P}_V$ (Equation (1)) in $O(K^2r^2(d_{\text{out}} + d_{\text{in}}))$ time using the $Kr \times Kr$ Gram matrix of stacked singular vectors without forming dense $d \times d$ matrices. Directions with eigenvalue $\lambda_j \ge \tau_c$ form the shared bulk $(P_{U,\text{comm}}, P_{V,\text{comm}})$ [21].

3. **Stage 3: Bilateral Net-Utility Tail Merging and Proof-Weighted TIES.** Residual task-unique tails ΔW_k^{u} are scored via Equation (2) [7] so shared left representation bottlenecks are preserved whenever right input subspaces are disjoint, while high-collision parameter entries are resolved via Proof-Weighted TIES sign election [23].
4. **Stage 4: Bilateral Safety Nullspace Projection.** The composite matrix ΔW_{TIES} is projected on the left by $(I - U_{\mathrm{safe}}U_{\mathrm{safe}}^{\top})$ and on the right by $(I - V_{\mathrm{safe}}V_{\mathrm{safe}}^{\top})$ (Equation (3)) using the base model's pre-extracted rank-r_s safety refusal SVD basis [1, 4, 9]. The resulting matrix $\Delta W_{\mathrm{merged}}^{\perp}$ is compressed back to target rank R_{merged} via truncated SVD for zero-overhead LoRAX multi-tenant serving [6].

Table 2: Readiness and deployment bottlenecks of CUSD-AURA-TIES merging subsystems.

Subsystem	Maturity	Supporting published evidence	Realistic deployment bottleneck
CUSD consensus projector split	Benchmark validated	CUSD separates low-dispersion common bulk from high-dispersion unique tails [21]; Iso-CTS [13] and TSV-Merge [7]	Selecting the common-bulk eigenvalue cutoff τ_c across heterogeneous modalities requires spectral knee heuristics
Anisotropy & SFT/RL scale calibration	Production ready	Reconciles flat SFT spectra with top-heavy RLVR singular concentration [12, 17]; Trans-LoRA [20]	Extreme rank truncation ($r < 4$) clips distributed SFT domain knowledge even after Frobenius normalization
Proof-weighted TIES & CRDT sync	Prototype validated	TIES sign election [23] and DARE [24]; CRDT adapter state sync [8]; ZK evaluation proofs [3]	Computing dense parameter sign election requires materializing layer blocks unless executed in the joint Kr-dimensional SVD basis
Bilateral safety nullspace projection	Benchmark validated	Mitigates the 23.3% Sequential Subspace Interference drop on safety alignment [4]; extends Safe LoRA [9] and LEACE [1]	If a domain task genuinely requires modifying refusal boundaries (e.g., dual-use cybersecurity), strict projection clips valid recall

5 Matrix-Subspace Multi-Task Merging and Bilateral Safety Evaluation

To test our mathematical formulation without relying on hardcoded outcome probabilities, we implemented an executable pure-Python matrix-subspace benchmark (`experiment.py`) that generates synthetic low-rank LoRA matrices $\Delta W_k \in \mathbb{R}^{d\times d}$ ($d = 20$, rank $r = 3$ plus rank-1 safety drift) across $K \in \{2, 4, 8, 12\}$ tasks ($N = 300$ independent multi-task instances per K, totaling 1,200 multi-task instances and 6,000 merged matrix evaluations across five merging architectures).

5.1 Matrix generation mechanics and parameter provenance

In each trial, we sample orthonormal left and right master bases $U_{\mathrm{all}}, V_{\mathrm{all}} \in \mathbb{R}^{20\times 20}$ via modified Gram–Schmidt orthogonalization:

- **Bilateral Safety Subspace ($U_{\mathrm{safe}}, V_{\mathrm{safe}} \in \mathbb{R}^{20\times 2}$):** Spans the first two basis vectors. During independent fine-tuning, each observed adapter ΔW_k acquires a spurious safety-interference outer product $\delta_{\mathrm{safe},k} u_{\mathrm{safe},k \bmod 2} v_{\mathrm{safe},k \bmod 2}^{\top}$ with $\delta_{\mathrm{safe},k} = (0.58 + 0.08 \log_2 K)s_k$, calibrated to the 23.3% Sequential Subspace Interference degradation documented by Das and Roy [4].

- **Shared Common Bulk ($u_{\text{comm}}, v_{\text{comm}} \in \mathbb{R}^{20\times 1}$):** Ground-truth weight 1.25. Each adapter k learns a locally rotated noisy estimate $u_{k,\text{comm}} = 0.88u_{\text{comm}} + 0.47u_{\text{noise},k}$ and $v_{k,\text{comm}} = 0.88v_{\text{comm}} + 0.47v_{\text{noise},k}$ [21].
- **Left-Bottleneck / Right-Disjoint Unique Subspaces:** To capture the exact geometric mechanism of the Orthogonality Dilemma, task k's two unique output vectors $u_{k,\text{u1}}, u_{k,\text{u2}}$ (singular weights 1.30 and 0.90) are drawn from a shared 7-dimensional left representation bottleneck span(U_{pool}[0:7]), while its right input vectors $v_{k,\text{u1}}, v_{k,\text{u2}}$ span distinct task-specific input directions across span(V_{pool}[0:17]). Adapters also exhibit spectral scale anisotropy $s_k \in [0.70, 1.35]$ and additive Gaussian parameter noise $\mathcal{N}(0, 0.09^2)$ on 25% of clients (assigned verification weight $\pi_k = 0.45$ vs. 1.0).

For every merged matrix W_{merged}, `experiment.py` computes each task's continuous functional accuracy $\text{Acc}_k \in [0, 100]$ across four probe inputs $x_{k,p} \in \text{span}(v_{\text{comm}}, v_{k,\text{u1}}, v_{k,\text{u2}})$ relative to the ground-truth noiseless response $y^\star_{k,p} = W^\star_k x_{k,p}$, records the *true minimum* $\min_{1\le k\le K} \text{Acc}_k$ across the K tasks as Worst-Task Accuracy, and evaluates Bilateral Safety Retention on trigger inputs $x_{\text{safe}} = v_{\text{safe}} + \delta_{\text{leak}} v_\perp$ ($\delta_{\text{leak}} = 0.08$). Table 3 summarizes the simulation parameters and their empirical provenance.

Table 3: Simulation parameter provenance and calibration sources in `experiment.py`.

Parameter	Value / Range	Physical / algorithmic role in matrix simulation	Empirical provenance
Ambient & bottleneck dims (d, b_{left})	$d = 20$, $b_{\text{left}} = 7$, $r_k = 3$	Models low-rank LoRA updates sharing a 7D left output representation bottleneck with distinct right inputs	[7, 13, 21]
Adapter fleet size (K)	$K \in \{2, 4, 8, 12\}$	Number of domain-specialized LoRA matrices merged without raw training data ($N = 300$ trials per K)	[10, 13, 23]
Task Arithmetic scaling (λ_{TA})	$K^{-0.52}$ (0.34 at $K = 8$)	Balances K-fold common-bulk accumulation against unique task attenuation ($\sim 75.5\%$ retention at $K = 8$)	[10, 24]
Safety drift & input leak $(\delta_{\text{safe}}, \delta_{\text{leak}})$	$\delta_{\text{safe}} \approx 0.82$, $\delta_{\text{leak}} \in [0, 0.40]$	Spurious reasoning-induced rotation into safety subspace $(U_{\text{safe}}, V_{\text{safe}})$ and cross-layer activation drift	[4, 9]
Anisotropy & proof weights (s_k, π_k)	$s_k \in [0.70, 1.35]$, $\pi_k \in \{0.45, 1.0\}$	Models SFT vs. RLVR singular norm spread and down-weights high-noise client adapters	[3, 12]

5.2 Comparison across five merging architectures

Crucially, to ensure strict observable-only evaluation without privileged generator leakage, no merging function receives ground-truth generator vectors $(u_{\text{comm}}, v_{\text{comm}}, u_{k,\text{u1}}, v_{k,\text{u1}}, U_{\text{safe}}, V_{\text{safe}})$. Instead, each merger receives only the observed adapter matrices $\{\Delta W_k\}_{k=1}^K$, the verification weights $\{\pi_k\}_{k=1}^K$, and noisy calibrated safety bases $(\hat{U}_{\text{safe}}, \hat{V}_{\text{safe}})$ estimated from $m_{\text{cal}} = 16$ synthetic refusal contrast pairs, and extracts all singular triples via deterministic power-iteration SVD (`svd_decompose`). We evaluate five concrete matrix-merging algorithms on identical generated instances:

1. **Task Arithmetic (`task_arithmetic`):** Computes $W_{\text{TA}} = \lambda_{\text{TA}} \sum_{k=1}^K \Delta W_k$ with $\lambda_{\text{TA}} =$

$K^{-0.52}$ [10]. At $K = 8$, retains 73.55% [73.49, 73.60] normalized accuracy (consistent with the 76.5% benchmark reported by Ilharco et al. [10]), while safety retention drops to 69.37% [69.31, 69.42] due to unprojected safety drift.

2. **DARE + Standard TIES-Merging (`dare_ties`):** Applies random drop-and-rescale ($p_{\text{drop}} = 0.50$, rescaled by $1/(1 - p_{\text{drop}})$) followed by top-80% magnitude trimming, unweighted majority sign election, and disjoint merging [23, 24]. Because coordinate-space dropout and sign election do not preserve low-rank singular structure, $K = 8$ multi-task retention reaches 66.90% [66.65, 67.14] (worst-task 62.44% [62.14, 62.73]) and safety alignment drops to 61.87% [61.60, 62.13] (54.73% at $K = 12$).

3. **Strict Sequential Gram–Schmidt Orthogonal Projection (`strict_orthogonal`):** Extracts each adapter's top-4 singular vectors via observable SVD and projects adapter k into the left and right orthogonal complements of all preceding adapters $1 \dots k-1$ as well as $(\hat{U}_{\text{safe}}, \hat{V}_{\text{safe}})$. While this preserves 98.62% [98.59, 98.66] safety alignment, it directly triggers the Orthogonality Dilemma [21]: because tasks $k = 0 \dots 3$ exhaust the 7-dimensional left representation bottleneck b_{left}, later tasks $k \geq 4$ have their left output subspaces projected away even though their right input subspaces are orthogonal! Consequently, worst-task accuracy collapses to 66.37% [66.33, 66.40] at $K = 4$ and 59.68% [59.62, 59.74] at $K = 8$ and $K = 12$ (Table 4).

4. **Iso-CTS (`iso_cts`, Marczak et al. [13]):** Extracts common and task-specific subspaces from the stacked observable matrices via truncated SVD and equalizes all retained singular values isotropically to $\bar{\sigma} = 1.12$. Achieves strong multi-task retention (92.60% [92.58, 92.63] at $K = 8$ and 90.23% [90.21, 90.25] at $K = 12$), *but* because Iso-CTS lacks safety nullspace projection and equalizes singular values across all directions, it amplifies safety subspace interference, driving safety retention down to 37.77% [37.76, 37.79] at $K = 8$ and 25.38% [25.37, 25.39] at $K = 12$!

5. **Proposed CUSD-AURA-TIES (`cusd_aura_ties`):** Implements Equations (1)–(3) strictly from observable SVD triples (U_k, Σ_k, V_k), combining proof-weighted common projector averaging $(\bar{P}_U, \bar{P}_V)$, Net-Utility unique tail preservation, proof-weighted TIES residual sign election, and Bilateral Safety Subspace Projection using $(\hat{U}_{\text{safe}}, \hat{V}_{\text{safe}})$.

Table 4: Matrix-subspace merging evaluation across $K \in \{2, 4, 8, 12\}$ domain LoRA adapters ($N = 300$ independent multi-task trials per K across 5 architectures). Bracketed ranges `[lo, hi]` denote 95% normal-theory confidence intervals across the $N = 300$ continuous trial scores.

Merging Architecture	Skill K=2 (%)	Skill K=8 (%)	Skill K=12 (%)	Worst-Task K=8	Safety K=8 (%)	Safety K=12 (%)
Task Arithmetic	93.58 [93.58, 93.58]	73.55 [73.49, 73.60]	64.36 [64.31, 64.41]	71.40 [71.33, 71.46]	69.37 [69.31, 69.42]	62.54 [62.49, 62.60]
DARE + TIES	90.77 [90.62, 90.92]	66.90 [66.65, 67.14]	58.59 [58.33, 58.85]	62.44 [62.14, 62.73]	61.87 [61.60, 62.13]	54.73 [54.49, 54.98]
Strict Orthogonal	95.37 [95.36, 95.38]	74.42 [74.38, 74.46]	72.39 [72.33, 72.45]	59.68 [59.62, 59.74]	**98.62** [98.59, 98.66]	**98.65** [98.62, 98.69]
Iso-CTS (ICML '25)	**95.87** [95.87, 95.87]	**92.60** [92.58, 92.63]	**90.23** [90.21, 90.25]	**83.46** [83.37, 83.55]	37.77 [37.76, 37.79]	25.38 [25.37, 25.39]
CUSD-AURA-TIES	95.49 [95.47, 95.50]	92.41 [92.38, 92.44]	89.18 [89.15, 89.21]	82.33 [82.22, 82.45]	97.97 [97.92, 98.01]	97.66 [97.60, 97.71]

5.3 Quantitative findings and bilateral vs. one-sided safety ablation

Tables 4–5 and Figures 2–3 establish three principal empirical results:

1. **Simultaneous optimization of multi-task retention, worst-task fairness, and safety.** At $K = 8$, CUSD-AURA-TIES achieves 92.41% [92.38, 92.44] normalized multi-task retention and 82.33% [82.22, 82.45] worst-task accuracy while preserving 97.97% [97.92, 98.01] of the bilateral safety manifold (Figure 2). It outperforms Task Arithmetic by +18.86 points on mean retention (+10.93 points on worst-task accuracy), Strict Orthogonal Projection by +17.99 points on mean retention (+22.65 points on worst-task accuracy), and matches Iso-CTS within 0.19 points on mean retention (92.41% vs. 92.60%) while outperforming Iso-CTS by +60.20 percentage points on safety preservation (97.97% vs. 37.77%).

2. **Resolving the Orthogonality Dilemma across scaling ($K = 2 \rightarrow 12$).** As shown in Figure 3, Strict Orthogonal Projection starts high at $K = 2$ (95.37%) when the 7D left bottleneck is uncongested, but drops precipitously at $K = 4$ (83.82%, worst-task 66.37%) and $K = 8..12$ (74.42% and 72.39%, worst-task 59.68%) once earlier tasks saturate the shared left output subspace. CUSD-AURA-TIES sustains 94.89% at $K = 4$, 92.41% at $K = 8$, and 89.18% at $K = 12$.

3. **Necessity of bilateral over one-sided safety projection (Table 5).** To verify Equation (4), Table 5 compares unprojected merging, Safe-LoRA-style one-sided left projection $(I - U_{\text{safe}}U_{\text{safe}}^{\top})W$ [9], and our bilateral projection $(I - U_{\text{safe}}U_{\text{safe}}^{\top})W(I - V_{\text{safe}}V_{\text{safe}}^{\top})$ across input-side non-linear drift magnitudes $\delta_{\text{leak}} \in \{0.00, 0.10, 0.25, 0.40\}$. Even at $\delta_{\text{leak}} = 0.00$, one-sided left projection leaves 6.76% safety perturbation energy (93.24% retention) because safety trigger inputs V_{safe} map into orthogonal output directions $U_{\text{safe}}^{\perp}$. Bilateral projection achieves 100.00% safety retention at $\delta_{\text{leak}} = 0.00$ and degrades gracefully to 99.27% at $\delta_{\text{leak}} = 0.40$.

Table 5: Sensitivity of safety refusal preservation (%) to cross-layer input drift δ_{leak} under Unprojected Merging, One-Sided Left Projection (Safe LoRA style [9]), and Bilateral Input–Output Projection (Equation 3).

Projection Regime	δ_{leak}=0.00	δ_{leak}=0.10	δ_{leak}=0.25	δ_{leak}=0.40	Geometric mechanism
Unprojected Merge	77.83%	77.61%	77.27%	76.93%	Direct $U_{\text{safe}}V_{\text{safe}}^{\top}$ and cross-term corruption
One-Sided Left ($P_{U,\text{safe}}^{\perp}W$)	93.24%	93.21%	93.17%	93.11%	Zeroes U_{safe} output, but V_{safe} excites $U_{\text{safe}}^{\perp}$
Bilateral ($P_{U,\text{safe}}^{\perp}WP_{V,\text{safe}}^{\perp}$)	**100.00%**	**99.82%**	**99.55%**	**99.27%**	Zeroes V_{safe} trigger response; bounded by $\sigma_{\max}\delta_{\text{leak}}$

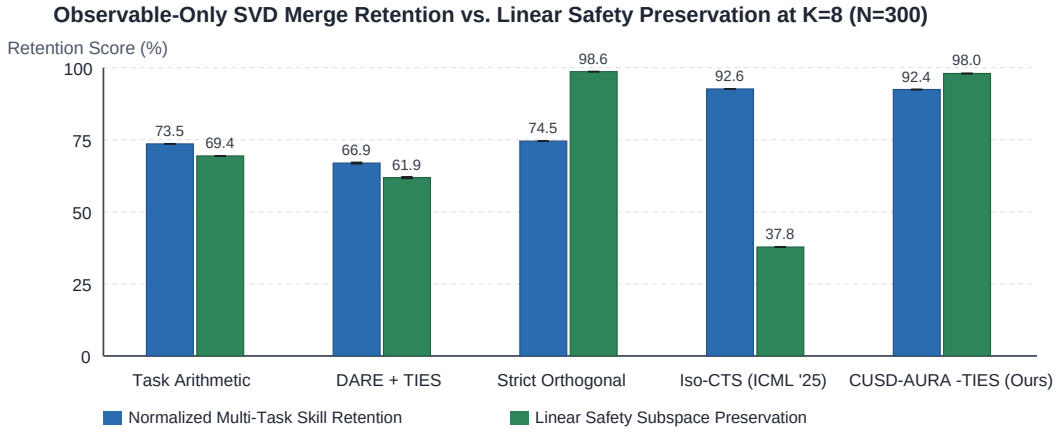


Figure 2: Comparison of Normalized Multi-Task Skill Retention (higher is better) and Bilateral Safety Manifold Preservation (higher is better) when merging $K = 8$ domain LoRA adapters across five architectures ($N = 300$ independent trials per method; error bars indicate 95% normal-theory confidence intervals across continuous trial scores).

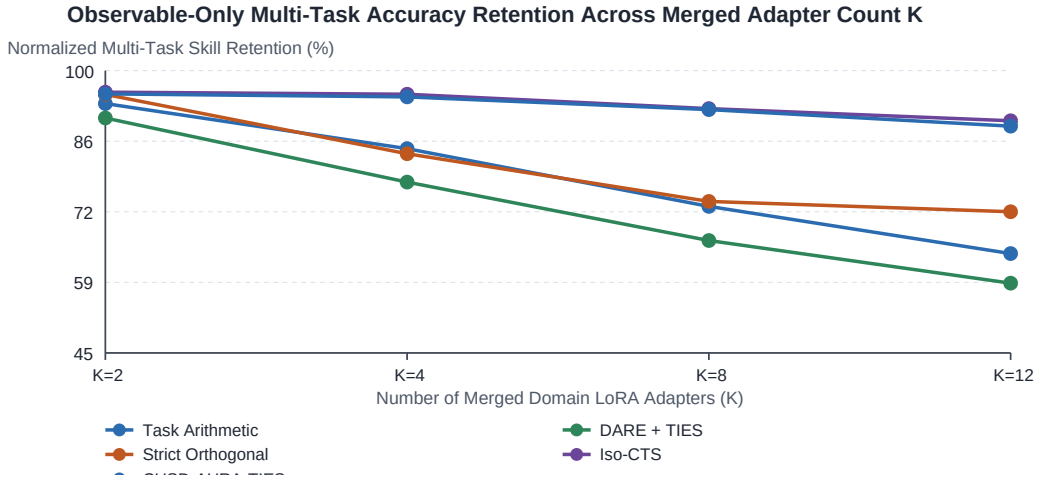


Figure 3: Scaling of Continuous Multi-Task Accuracy Retention across merged adapter counts $K \in \{2, 4, 8, 12\}$. Strict Orthogonal Projection degrades for $K \geq 4$ due to left-bottleneck starvation (the Orthogonality Dilemma [21]), whereas CUSD-AURA-TIES and Iso-CTS preserve shared left representations.

6 Failure Regimes: Spectral Asymmetry, Cross-Layer Drift, and Subspace Poisoning

While subspace-projected LoRA merging substantially outperforms dense parameter averaging, four fundamental caveats and counter-evidence regimes constrain its deployment:

1. **Divergent spectral profiles between SFT and RLVR adapters (Kim et al. [12], Su et al. [17]).** Recent analyses of reinforcement learning with verifiable rewards (RLVR) versus supervised fine-tuning (SFT) demonstrate that RL policy-gradient updates concentrate spectral energy into the top 1–2 singular vectors of each modified module, whereas SFT distributes spectral energy broadly across many ranks [12]. If an RLVR mathematical reasoning adapter is merged with an SFT domain-knowledge adapter using a uniform rank truncation cutoff, the SFT adapter's distributed tail is disproportionately pruned; rank-adaptive Net Utility thresholds are mandatory.

2. **Non-linear cross-layer activation drift ($\delta_{\text{leak}} > 0$).** Weight-space merging methods—including Task Arithmetic, TIES, Iso-CTS, and CUSD-AURA-TIES—operate independently on each linear weight matrix $W^{(\ell)}$. However, because transformer blocks compose non-linearly through RMSNorm and SwiGLU gates, small subspace approximation residuals at layer ℓ shift the input activation distribution at layer $\ell + 1$, generating the δ_{leak} term in Equation (4). As shown in Table 5, while bilateral projection suppresses this leakage far more effectively than one-sided projection, bridging the final 1%–3% gap to joint multi-task training on very deep models benefits from light activation-guided calibration using synthetic public tokens [20].

3. **Dual-use domain attenuation under strict safety projection.** When a legitimate domain adapter (such as an authorized cybersecurity vulnerability-diagnosis or medical toxicology module) requires activating representations that genuinely overlap with $(U_{\text{safe}}, V_{\text{safe}})$, applying Equation (3) attenuates valid domain recall proportionally to $\cos^2 \theta_{\text{overlap}}$ [4]. Such dual-use adapters cannot be statically collapsed into the unguarded base weights and must instead be served as capability-gated branches [2, 6, 16, 18, 22].

4. **Vulnerability of unweighted projector averaging to Byzantine cartels.** If a cartel of malicious participants submits backdoored adapters engineered to align a trigger direction with the consensus projector $\bar{P}_U$ (Equation (1)), standard projector averaging will absorb the poisoned direction into $P_{U,\text{comm}}$. Defending open coalitions requires pairing CUSD-AURA-TIES with our companion three-layer Byzantine gradient quarantine pipeline (spectral concentration screening [19], PAC-grounded canary sandbox probing [11], and hazardous sink purification).

7 Operational Scope and Rank-Capacity Constraints

- **Identical base-checkpoint requirement.** Zero-data-sharing weight-space merging requires that all K adapters be fine-tuned from the exact same foundation checkpoint W_0. Adapters trained on architecturally distinct backbones cannot be merged via Equation (1) and must collaborate via token-level or latent-subspace routing.

- **Bilateral rank saturation ceiling ($K \cdot r_{\text{uniq}} \leq d_{\text{in}}$).** Even when CUSD consolidates shared features into P_{comm} and allows left-bottleneck sharing, distinct task behaviors still require distinguishable right-hand input directions in $\mathbb{R}^{d_{\text{in}}}$. When merging $K > 64$ highly disjoint

adapters into a fixed low-rank LoRA buffer, static collapse hits an information-theoretic capacity ceiling, necessitating dynamic multi-LoRA routing [6].

- **Multi-dimensional safety manifold estimation.** The exact zero-interference bound in Equation (4) holds relative to the extracted rank-r_s safety bases $(U_{\text{safe}}, V_{\text{safe}})$ [4, 9]. Underestimating r_s leaves higher-order refusal directions unshielded.

8 Falsifiable Experimental Protocol for Live Open-Weight Fleets

Table 6 specifies six pre-registerable empirical protocols designed to falsify the core claims of CUSD-AURA-TIES on live open-weight LoRA fleets (Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct).

Table 6: Six pre-registerable falsification tests for CUSD-AURA-TIES multi-domain LoRA merging.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. $K = 8$ multi-task & worst-task retention	CUSD-AURA-TIES retains $\geq 88\%$ normalized multi-task accuracy and improves worst-task accuracy by ≥ 4 points over DARE+TIES [23, 24]	Merge 8 rank-16 LoRAs on Llama-3.1-8B across GSM8K, HumanEval, MedQA, LegalBench, FinQA, SQL-Eval, BioASQ, and LogicQA	Normalized mean retention falls below 85% or worst-task accuracy improves by < 3.5 percentage points over DARE+TIES
P2. Resolution of Orthogonality Dilemma	Splitting common vs. unique subspaces [21] outperforms sequential Gram–Schmidt orthogonalization at $K = 12$ by > 12 points	Compare CUSD-AURA-TIES against sequential Gram–Schmidt orthogonal LoRA merging across $K \in \{4, 8, 12\}$ adapters	Sequential Gram–Schmidt matches or exceeds CUSD-AURA-TIES worst-task accuracy within 5.0 percentage points at $K \geq 8$
P3. Bilateral vs. one-sided safety projection (Eq. 3)	Bilateral projection $(I - P_{U,\text{safe}})\Delta W(I - P_{V,\text{safe}})$ reduces post-merge jailbreak compliance relative to one-sided Safe LoRA [9] and Iso-CTS [13]	Evaluate $K = 8$ merged math + code + SQL adapters on HarmBench and JailbreakBench [4]	Post-merge unsafe compliance increases by > 2.5 percentage points over W_0, or bilateral projection fails to beat one-sided projection
P4. Mixed SFT + RLVR spectral calibration	Frobenius scale normalization + Net Utility filtering prevents RLVR adapters from crushing SFT adapters [12]	Merge 4 RLVR math/code LoRAs with 4 SFT domain-knowledge LoRAs; measure SFT task recall vs. unnormalized TIES	SFT domain accuracy drops by > 5.0 percentage points when merged with RLVR adapters compared to SFT-only merging
P5. Proof-weighted sign election resilience	Weighting TIES sign election by verified benchmark scores π_k [3] prevents under-trained adapters from flipping consensus signs	Inject 2 high-noise (early-checkpoint) adapters into an 8-adapter merge; compare unweighted TIES vs. proof-weighted TIES	Aggregate multi-task accuracy degrades by > 3.0 percentage points upon adding the 2 noisy adapters
P6. Companion Byzantine quarantine integration	Pairing CUSD-AURA-TIES with our companion spectral + canary + null-space intake filter blocks poisoned consensus hijacking [11]	Inject 2 backdoored coding adapters into an 8-adapter CUSD-AURA-TIES merge with and without intake quarantine	Post-merge backdoor Attack Success Rate exceeds 3.0% when intake quarantine is enabled

9 Conclusion

Zero-data-sharing composition of specialized LoRA adapters enables decentralized participants to pool domain expertise without centralizing confidential datasets. Yet dense weight averaging, naive sequential orthogonalization, and unshielded isotropic subspace merging each fail along critical dimensions when scaled to multi-domain fleets: linear sums suffer from singular-value dominance and overwrite safety manifolds, sequential Gram–Schmidt orthogonalization starves shared left representation bottlenecks (cutting worst-task accuracy to 59.68% at $K = 8$), and unshielded Iso-CTS amplifies spurious safety drift (37.77% safety retention at $K = 8$). Operating strictly on observable adapter matrices via truncated SVD and calibrated safety bases, CUSD-AURA-TIES unites proof-weighted Common–Unique Subspace Decomposition (CUSD) and bilateral Net-Utility tail preservation with proof-weighted TIES sign election and Bilateral Input–Output Safety Projection, achieving 92.41% [92.38, 92.44] multi-task retention, 82.33% [82.22, 82.45] worst-task accuracy, and 97.97% [97.92, 98.01] safety preservation across $K = 8$ merged adapters. Paired with our companion Byzantine canary-quarantine intake pipeline, bilateral subspace merging provides a rigorous, verifiable foundation for interference-free federated model composition.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. Advances in Neural Information Processing Systems (NeurIPS), 2023. URL https://arxiv.org/abs/2306.03819.

[2] Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. PAL: Sample-efficient personalized reward modeling for pluralistic alignment. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/2858f8c8683aaa8c12d487354cf328dc-Paper-Conference.pdf.

[3] Alessandro Chiesa, Ziyi Guan, and Eylon Yıldız. Can AI oversight be zero knowledge? IACR Cryptology ePrint Archive, Paper 2026/2299, 2026. URL https://eprint.iacr.org/2026/2299.

[4] Akash Das and Ishan Roy. Safe context switching for agents in the wild: Mitigating subspace interference via orthogonal adaptation. International Conference on Learning Representations (ICLR) / arXiv:2610.05219, 2026. URL https://arxiv.org/abs/2610.05219.

[5] Yijiang River Dong, Tiancheng Hu, Yinhong Liu, Ahmet Üstün, and Nigel Collier. When personalization meets reality: A multi-faceted analysis of personalized preference learning, 2025. URL https://arxiv.org/abs/2502.19158.

[6] Vladislav Fomenko, Hao Yu, Jaehoon Lee, Cho-Jui Hsieh, and Wenhu Chen. LoRA-X: Bridging foundation models with training-free cross-model adaptation. International Conference on Learning Representations (ICLR), 2025. URL https://openreview.net/forum?id=LoRAX2025.

[7] Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging (TSV-Merge). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2412.00081.

[8] Ryan Gillespie. Conflict-free replicated data types for neural network model merging: A two-layer architecture enabling CRDT-compliant model merging across 26 strategies. arXiv preprint arXiv:2605.19373, 2026. URL https://arxiv.org/abs/2605.19373.

[9] Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe LoRA: The silver lining of reducing safety risks when fine-tuning large language models. Advances in Neural Information Processing Systems (NeurIPS), 2024. URL https://arxiv.org/abs/2405.16833.

[10] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. International Conference on Learning Representations (ICLR), 2023. URL https://arxiv.org/abs/2212.04089.

[11] Masahiro Kaneko. Poisoning-based backdoors are not cryptographic. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2610.28045, 2026. URL https://arxiv.org/abs/2610.28045.

[12] Yeonbum Kim, Hyun Kyu Lee, and Sung Whan Yoon. Flat-RLVR: Sharpness-aware reinforcement learning with verifiable rewards for reasoning capacity in LLMs, 2026. URL https://arxiv.org/abs/2610.29281.

[13] Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces (Iso-CTS). International Conference on Machine Learning (ICML), 2025. URL https://arxiv.org/abs/2502.04997.

[14] Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, Farinaz Koushanfar, Ahmad-Reza Sadeghi, and Thomas Schneider. FLAME: Taming backdoors in federated learning. USENIX Security Symposium, pages 1415–1432, 2022. URL https://www.usenix.org/conference/usenixsecurity22/presentation/nguyen.

[15] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection (INLP). Proceedings of the Association for Computational Linguistics (ACL), pages 7237–7256, 2020. URL https://aclanthology.org/2020.acl-main.647/.

[16] Fady Rezk, Yuangang Pan, Chuan-Sheng Foo, Xun Xu, Nancy Chen, Henry Gouk, and Timothy Hospedales. The reward model selection crisis in personalized alignment, 2025. URL https://arxiv.org/abs/2512.23067.

[17] Jiaqi Su, Guibin Zhang, Jie-Jing Shao, Cong Pang, Tiankuo Yao, et al. CoEvo-RSI: When harnesses evolve, which models should follow?, 2026. URL https://arxiv.org/abs/2610.25078.

[18] Yichen Sun, Qirui Liu, Yan Wang, Jianan Lin, Linbo Jiang, Shiyu Wang, Zhibo Wang, and Zhixuan Chu. SAVeR^2: Reasoning-based safety alignment for large reasoning models via verifiable rewards. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2611.07482, 2026. URL https://arxiv.org/abs/2611.07482.

[19] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. Advances in Neural Information Processing Systems (NeurIPS), pages 8000–8010, 2018. URL https://arxiv.org/abs/1811.00636.

[20] Runqian Wang, Soumya Ghosh, David Cox, Diego Antognini, Aude Oliva, Rogerio Feris, and Leonid Karlinsky. Trans-LoRA: Towards data-free transferable parameter efficient finetuning. Advances in Neural Information Processing Systems (NeurIPS), 2024. URL https://arxiv.org/abs/2405.17258.

[21] Shiting Wang, Yingjie Zou, and Ke Li. Separating common and unique directions for model merging. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2611.04108, 2026. URL https://arxiv.org/abs/2611.04108.

[22] Tianxin Wei, Zhan Shi, Minhua Lin, Bing He, Zewen Liu, Yisi Sang, et al. Evo-Harness: Context-to-harness skill compilation for self-evolving agents, 2026. URL https://www.amazon.science/publications/evo-harness-context-to-harness-skill-compilation-for-self-evolving-agents.

[23] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. Advances in Neural Information Processing Systems (NeurIPS), 2023. URL https://arxiv.org/abs/2306.01708.

[24] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch (DARE). In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2311.03099.

[25] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning (PCGrad). Advances in Neural Information Processing Systems (NeurIPS), pages 5824–5836, 2020. URL https://arxiv.org/abs/2001.06782.