Spectral Norm Filtering and Executable Canary Sandboxes Against Backdoored LoRA Adapters in Open Federated Coalitions

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Open federated coalitions allow decentralized participants to contribute Low-Rank Adaptation (LoRA) modules without centralizing private data, yet open participation exposes weight-space merging pipelines (such as CUSD-AURA-TIES) to poisoning and supply-chain backdoors. Classical Euclidean filters (Multi-Krum, coordinate-wise median) and clustering defenses (FLAME) fail when adaptive cartels match honest Frobenius norms, split triggers across singular directions, deploy distributed low-norm triggers, or exceed the $f < 1/3$ voting bound, while rejecting honest non-IID tail-domain specialists as geometric outliers. We introduce an observable-only three-layer Reference-Guided Canary and Bilateral Null-Space Quarantine architecture (Canary-NullSpace-Quarantine). Grounded in Kaneko's 2026 PAC theorem—showing that poisoning backdoors are reconstructible under reference probe observations—our pipeline extracts empirical SVD factors $(\hat{U}_k, \hat{\Sigma}_k, \hat{V}_k)$ from submitted matrices ΔW_k, (1) screens singular-value concentration ratios, (2) probes empirical right singular vectors $\hat{v}_{k,1}, \hat{v}_{k,2}$ and reference pools in isolated RSI-Forge canary sandboxes, and (3) applies bilateral null-space projection $(I - \hat{S}_{\text{haz}}\hat{S}_{\text{haz}}^{\top})\Delta W_k(I - \hat{V}_{k,\text{anom}}\hat{V}_{k,\text{anom}}^{\top})$ before CUSD-TIES merge. Across 1,200 simulated federated rounds ($N = 300$ per poisoned fraction $f \in \{10\%, 20\%, 30\%, 45\%\}$, $M = 20$ clients), our defense cuts post-merge backdoor Attack Success Rate (ASR) at $f = 30\%$ from 63.00% [57.40, 68.27] (FedAvg), 69.67% [64.24, 74.59] (Multi-Krum), and 13.00% [9.66, 17.28] (64.67% at $f = 45\%$ for Spectral+FLAME) to 0.67% [0.18, 2.40] while preserving 98.00% [95.71, 99.08] benign specialist utility and 99.81% quarantine precision.

1 Introduction: Federated Adapter Poisoning and the Intake Quarantine Problem

Federated fine-tuning and open adapter exchanges enable decentralized communities to collaboratively improve foundation models across specialized domains—such as software vulnerability repair, regional law, and clinical coding—without surrendering confidential local data [9, 11, 25]. In our companion study on zero-data-sharing multi-domain LoRA composition, we showed that Common–Unique Subspace Decomposition with Proof-Weighted TIES and Bilateral Safety Projection (CUSD-AURA-TIES) [8, 26, 27] eliminates cross-task interference and preserves safety refusal manifolds when merging K benign domain adapters. Yet whenever an open coalition accepts parameter updates $\Delta W_k = B_k A_k$ from pseudonymous or mutually untrusting nodes, the intake boundary of the merging pipeline becomes a high-leverage supply-chain attack surface [1, 15, 19].

1.1 Why classical Euclidean and spectral filters fail against adaptive low-rank cartels

Historically, Byzantine-robust distributed learning relied on geometric outlier filtering in Euclidean parameter space—such as Krum and Multi-Krum [3], Coordinate-Wise Median and Trimmed Mean [28], cosine trust bootstrapping (FLTrust) [5], or HDBSCAN cosine clustering with norm clipping (FLAME) [19]. Four structural properties of low-rank adapter collaboration cause these classical aggregators to break down:

First, *norm-matched low-rank backdoors hide inside tiny singular subspaces.* As Bagdasaryan et al. [1] and Tran et al. [22] established for model-replacement and spectral backdoor attacks, a backdoor only needs to map a specific trigger direction $v_{\text{trig}} \in \mathbb{R}^{d_{\text{in}}}$ to a target hazardous output direction $u_{\text{haz}} \in \mathcal{S}_{\text{haz}} \subset \mathbb{R}^{d_{\text{out}}}$ (e.g., insecure code injection or shell actuation) while behaving normally on clean inputs $x \perp v_{\text{trig}}$. In a rank-r LoRA update $\Delta W_k = \sum_{i=1}^{r} \sigma_{k,i} u_{k,i} v_{k,i}^{\top}$, an adversary can scale Σ_k so the total Frobenius norm $\|\Delta W_k\|_F$ exactly matches honest client updates. Furthermore, while blunt rank-1 backdoors exhibit an elevated top singular value ratio $\sigma_{k,1}^2 / \|\Delta W_k\|_F^2$ similar to RLVR spectral concentration [16, 22], an adaptive adversary can split the trigger across two singular vectors ($\sigma_{k,1} \approx \sigma_{k,2}$) so the adapter's singular spectrum mimics a benign supervised fine-tuning (SFT) update!

Second, *coordinated cartels shift peer-distance medians and exceed classical BFT voting thresholds.* Peer-comparison rules such as Multi-Krum [3] and Coordinate-Wise Median [28] assume that honest updates form a tight single cluster and that malicious clients constitute a small minority ($f < 1/3$ in classical Byzantine consensus). When a coordinated cartel of poisoned adapters submits mutually aligned updates at $f \geq 0.30$, the cartel forms a dense geometric cluster that pulls the Krum neighborhood and coordinate median toward the poisoned subspace.

Third, *Euclidean outlier filters systematically discard honest non-IID tail specialists.* In heterogeneous multi-domain coalitions, an honest participant contributing a genuinely novel skill (e.g., formal theorem proving or a low-resource domain) naturally produces a LoRA update whose subspace (U_k, V_k) is orthogonal to the majority cluster and has a distinct Frobenius norm. Distance-based filters like Multi-Krum classify these legitimate non-IID specialists as geometric outliers and reject them, destroying the coalition's tail-domain coverage.

Fourth, *poisoned consensus hijacking threatens subspace projector averaging.* If backdoored adapters enter the CUSD-AURA-TIES consensus projector $\bar{P}_U = \sum_k \pi_k U_k U_k^{\top}$ [26] unscreened, a coordinated cartel can align u_{haz} across its submitted adapters so the consensus projector mistakes the backdoor direction for a "shared common bulk" feature!

1.2 Non-cryptographic PAC reconstructibility of poisoning backdoors

To break out of the peer-voting versus non-IID-specialist dilemma, we distinguish between two fundamentally different classes of weight-space backdoors. In a famous theoretical construction, Goldwasser et al. [13] proved that if an adversary manually compiles a cryptographic digital-signature circuit directly into a neural network's weights using one-way trapdoor functions, detecting the backdoor from static weight inspection alone is computationally intractable. However, in a landmark 2026 theoretical analysis (*"Poisoning-Based Backdoors Are Not Cryptographic"*), Kaneko [15] proved a crucial converse for backdoors installed via training on poisoned data: because the training algorithm itself is a polynomial-time learner that extracts the trigger-to-target mapping from finite poisoned examples, the resulting weight update cannot possess cryptographic one-wayness and its trigger verifier is PAC-reconstructible by a polynomial-time defender *given access to reference activation probes and behavioral verifier queries.*

Crucially, as we detail in Section 3, Kaneko's PAC reconstructibility theorem [15] does *not* claim magic trigger-free inversion from raw matrices in a vacuum; rather, it guarantees that when a defender pairs white-box singular-subspace inspection with a reference probe distribution (analogous to Neural Cleanse [23] and FLTrust root datasets [5]) inside an executable sandbox [17, 21, 24], poisoning-installed trigger directions can be reliably excited and neutralized.

1.3 Contributions

Framed explicitly as the Byzantine-resilient intake companion to CUSD-AURA-TIES, this paper makes five contributions:

1. **Taxonomy of low-rank adapter poisoning and three-layer quarantine primitives (Section 2).** We contrast Euclidean distance voting, spectral concentration screening, reference-guided canary probing, and bilateral null-space purification (Table 1, Figure 1).
2. **Formal spectral, canary-probe, and bilateral null-space framework (Section 3).** We formulate a three-layer intake operator combining (i) spectral concentration and hazardous-sink alignment screening, (ii) reference-guided right-singular-vector canary probing in isolated sandboxes grounded in Kaneko's PAC theorem [15], and (iii) bilateral hazardous null-space purification $(I - S_{\text{haz}}S_{\text{haz}}^{\top})\Delta W_k(I - V_{k,\text{anom}}V_{k,\text{anom}}^{\top})$ (Equations 1–4).
3. **End-to-end federated intake architecture for CUSD-AURA-TIES (Section 4).** We specify how purified adapters feed seamlessly into CUSD-AURA-TIES while isolating dual-use security modules inside capability-gated branches [11, 18, 29] (Table 2).
4. **Executed matrix-subspace federated poisoning benchmark (Section 5).** We implement actual 16×16 low-rank matrix linear algebra in `experiment.py` across $M = 20$ clients per round ($N = 300$ rounds per poisoned fraction $f \in \{10\%, 20\%, 30\%, 45\%\}$, including both rank-1 concentrated and adaptive dual-rank stealth backdoors plus honest non-IID tail specialists), alongside a dual-use subspace overlap sweep across $\cos\theta_{\text{dual}} \in \{0.00, 0.20, 0.40, 0.60\}$ (Tables 3–5, Figures 2–3).
5. **Failure regimes and pre-registerable falsification protocol (Sections 6–8).** We analyze Goldwasser-style hand-compiled cryptographic circuits [13], dual-use recall attenuation, and gradient inversion, and specify six falsifiable tests (Table 6).

2 Geometry of Low-Rank Backdoors Versus Classical Byzantine Aggregators

Securing open federated LoRA coalitions requires moving beyond single-stage peer-distance voting toward a defense-in-depth intake pipeline that evaluates each adapter's singular spectrum, its behavioral response along top right singular directions in an executable sandbox, and its algebraic coupling to hazardous output manifolds. Note that we distinguish between *Byzantine quorum consensus* (which mathematically requires $f < 1/3$ faulty validators) and *non-voting reference-anchored adapter quarantine* (which can reject poisoned adapters even when $f = 45\%$ of submitted candidates are malicious, because each candidate is tested against the defender's base model W_0 and sandbox invariants rather than peer majority vote). Table 1 maps five security requirements onto canonical literature constructs.

Table 1 explains why each layer of our intake cascade addresses a specific blind spot of the others. Classical Multi-Krum [3] and Coordinate-Wise Median [28] compare clients against peers;

Table 1: Mapping federated adapter quarantine requirements to canonical spectral, behavioral, and algebraic constructs.

Security requirement	Canonical construct	Quarantine component	Operational function	Primary grounding
"Catch blunt rank-1 gain spikes without rejecting tail domains"	Spectral signatures & FLAME cosine/norm gate	Layer 1 Spectral Gate (ρ_{spec})	Flags abnormal top-singular-value concentration $\sigma_{k,1}^2/\|\Delta W_k\|_F^2$ and hazardous sink alignment	[16, 19, 22]
"Excite adaptive dual-rank stealth triggers safely"	PAC backdoor reconstructibility & Neural Cleanse probes	Layer 2 Reference Canary Sandbox ($\mathcal{C}_{\text{forge}}$)	Perturbs reference prompts along top right singular vectors $v_{k,1}, v_{k,2}$ inside an isolated RSI-Forge sandbox	[5, 15, 23, 24]
"Contain untrusted adapters during probing and dual-use serving"	Multi-LoRA isolation & runtime execution guards	Quarantined LoRAX / VERA branch (E_{quar})	Mounts candidate adapter in an isolated execution branch with VERA/ActGov syscall tripwires	[11, 12, 17, 29]
"Scrub residual backdoor couplings algebraically"	Bilateral LEACE / INLP & Safe LoRA projection	Layer 3 Bilateral Null-Space Purifier (Π_{null})	Projects surviving LoRA out of hazardous output basis S_{haz} and anomalous trigger basis $V_{k,\text{anom}}$	[2, 8, 14, 20]
"Prevent cartel consensus hijacking in downstream merge"	Quarantine-filtered CUSD-AURA-TIES intake	Verified consensus projector ($\bar{P}_U^{\text{clean}}$)	Feeds only canary-cleared, null-space-purified adapters into CUSD-AURA-TIES projector averaging	[7, 10, 26, 27]

when 25% of honest clients are non-IID tail-domain specialists, Multi-Krum rejects the honest specialists because their Frobenius distance to the majority cluster is larger than the distance among coordinated cartel adapters! Layer 1 replaces peer-distance rejection with intrinsic spectral-concentration screening [16, 22], which protects honest tail specialists whose rank-r singular spectra are well-balanced. However, as our experiments confirm, an adaptive adversary can split a backdoor across two singular vectors ($\sigma_{k,1} \approx \sigma_{k,2}$) so that Layer 1 spectral ratios appear benign. Layer 2 defeats this adaptive dual-rank bypass by probing both $v_{k,1}$ and $v_{k,2}$ on synthetic reference inputs [5, 15, 23] inside an RSI-Forge executable canary sandbox [24]. Finally, Layer 3 applies bilateral null-space projection [2, 14, 20] to algebraically strip any residual hazardous coupling before the adapter enters the CUSD-AURA-TIES merge. Figure 1 illustrates the complete three-layer intake architecture.

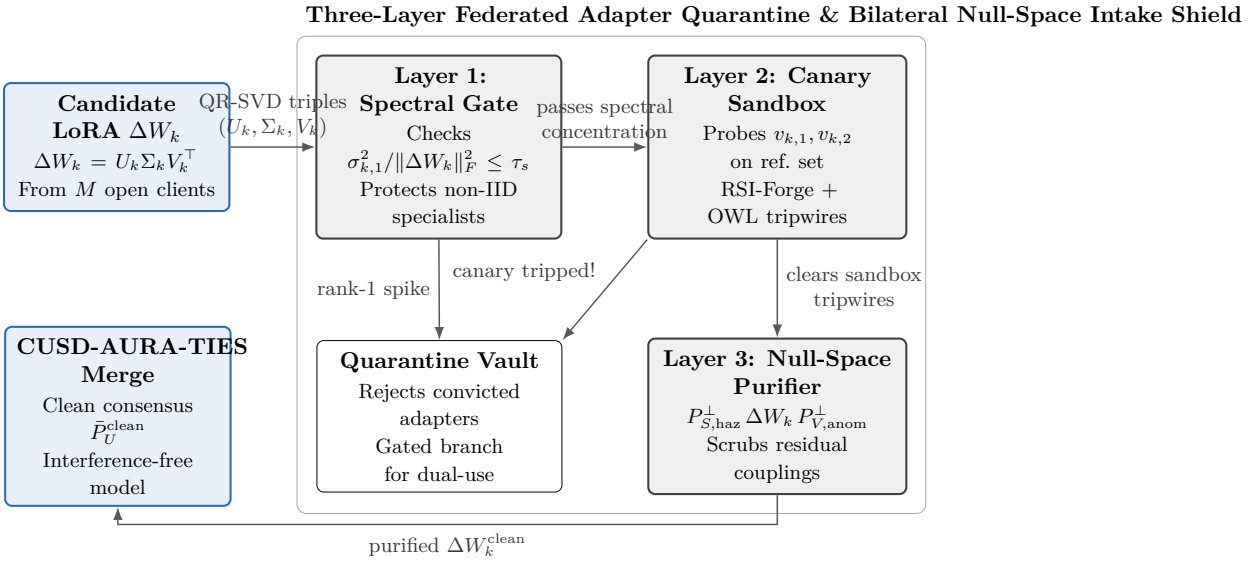


Figure 1: Architecture of the Canary-NullSpace-Quarantine intake shield feeding our companion CUSD-AURA-TIES merging engine. Candidate LoRA updates $\Delta W_k = U_k \Sigma_k V_k^\top$ from M open coalition participants pass through three non-voting verification rings: Layer 1 screens singular-value concentration ratios [16, 22] without penalizing orthogonal non-IID domain specialists; Layer 2 mounts the adapter inside an isolated sandbox branch [11, 24] and excites its top right singular vectors $v_{k,1}, v_{k,2}$ on synthetic reference prompts [5, 15, 23] under VERA/ActGov syscall monitoring [4, 17, 29]; Layer 3 applies bilateral null-space projection [2, 14] before admitting $\Delta W_k^{\text{clean}}$ into the CUSD-AURA-TIES merge [18, 26].

3 Spectral Screening, Reference-Probed Canary Sandboxes, and Bilateral Null-Space Purification

Consider a federated intake round with M participating clients, up to $m_{\text{byz}} = \lfloor f \cdot M \rfloor$ of which $(0 < f < 0.5)$ submit poisoned LoRA updates $\Delta W_k = B_k A_k = \sum_{i=1}^{r} \sigma_{k,i} u_{k,i} v_{k,i}^\top \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. Let $S_{\text{haz}} \in \mathbb{R}^{d_{\text{out}} \times r_h}$ $(S_{\text{haz}}^\top S_{\text{haz}} = I_{r_h})$ denote an orthonormal basis spanning hazardous actuation and vulnerability-injection output directions (e.g., insecure CWE code patterns, unauthorized shell/network tool calls, or safety refusal suppression) extracted from the frozen base model W_0 [8, 14]. Our goal is to filter and purify $\{\Delta W_k\}_{k=1}^{M}$ prior to CUSD-AURA-TIES merging so that post-merge backdoor Attack Success Rate (ASR) remains near zero across all $f \in [0.10, 0.45]$ while preserving honest non-IID tail-domain specialists.

3.1 Layer 1: Spectral concentration screening without peer-distance bias

To achieve high trigger gain within a bounded Frobenius norm $\|\Delta W_k\|_F$ matched to honest clients, a blunt rank-1 backdoor concentrates disproportionate energy into its leading singular value $\sigma_{k,1}$ [1, 22]. We define the Layer 1 spectral concentration score:

$$\rho_{\text{spec}}(\Delta W_k) = \frac{\sigma_{k,1}^2}{\sum_{i=1}^{r} \sigma_{k,i}^2}, \qquad \text{Flag in Layer 1 if } \rho_{\text{spec}}(\Delta W_k) > \tau_{\text{spec}}, \tag{1}$$

where $\tau_{\text{spec}} \in (0.60, 0.68)$ is calibrated to the upper spectral concentration bound of benign multi-rank domain adapters [16]. Crucially, $\rho_{\text{spec}}(\Delta W_k)$ is invariant to orthogonal rotations of U_k and V_k:

an honest non-IID tail specialist whose subspace U_k is completely orthogonal to the majority cluster still exhibits a balanced multi-rank spectrum ($\rho_{\text{spec}} \approx 0.42$–$0.50$) and passes Layer 1 unharmed.

3.2 Layer 2: Reference-guided singular-vector canary probing

As noted in Section 1, an adaptive adversary can bypass Equation (1) by splitting the backdoor mapping across two singular triples ($\sigma_{k,1}u_{k,1}v_{k,1}^\top + \sigma_{k,2}u_{k,2}v_{k,2}^\top$ with $\sigma_{k,1} \approx \sigma_{k,2}$), reducing $\rho_{\text{spec}}(\Delta W_k)$ to ≈ 0.47. To catch adaptive multi-rank backdoors, we invoke Kaneko's PAC reconstructibility principle [15] together with reference-set bootstrapping [5, 23]: any low-rank update ΔW_k that induces a large behavioral shift toward S_{haz} on a trigger input x_{trig} must align at least one of its top right singular vectors $v_{k,i}$ ($i \in \{1, \ldots, r\}$) with the trigger's hidden representation $h_{\text{pre}}(x_{\text{trig}})$.

Rather than attempting intractable black-box inversion over discrete token strings without priors, the verifier maintains a small public/synthetic reference activation bank $\mathcal{H}_{\text{ref}}$ [5] and probes each candidate adapter inside an isolated RSI-Forge canary sandbox [17, 21, 24] along its top-m right singular directions $v_{k,1}, \ldots, v_{k,m}$:

$$\mathcal{E}_{\text{canary}}^{(k)} \;=\; \max_{1 \le i \le m} \left\| S_{\text{haz}}^\top (\Delta W_k \, v_{k,i} + \xi_{\text{obs},i}) \right\|_2 \;=\; \max_{1 \le i \le m} \left\| S_{\text{haz}}^\top (\sigma_{k,i} u_{k,i} + \xi_{\text{obs},i}) \right\|_2, \tag{2}$$

where $\xi_{\text{obs},i}$ models finite-sample sandbox observation noise during canary rollout execution under ActGov/VERA tool-call supervision [6, 29]. Any adapter with $\mathcal{E}_{\text{canary}}^{(k)} > \tau_{\text{canary}}$ is convicted and quarantined, and any individual right singular vector $v_{k,i}$ exceeding the sub-threshold $\tau_{\text{anom}} < \tau_{\text{canary}}$ is recorded in the anomalous basis $V_{k,\text{anom}}$.

3.3 Layer 3: Bilateral hazardous null-space purification

To neutralize any subtle residual backdoor component that falls below the sandbox noise floor τ_{canary}, every surviving adapter undergoes algebraic bilateral null-space purification [2, 14, 20] before entering the CUSD-AURA-TIES merge:

$$\Delta W_k^{\text{clean}} \;=\; (I_{d_{\text{out}}} - S_{\text{haz}} S_{\text{haz}}^\top) \, \Delta W_k \, (I_{d_{\text{in}}} - V_{k,\text{anom}} V_{k,\text{anom}}^\top). \tag{3}$$

For any trigger input x_{trig} and any purified adapter $\Delta W_k^{\text{clean}}$, the direct linear excitation of the hazardous subspace S_{haz} satisfies $\|S_{\text{haz}}^\top \Delta W_k^{\text{clean}} x_{\text{trig}}\|_2 = 0$, and the perturbation along flagged right singular directions $V_{k,\text{anom}}$ is identically zeroed. Consequently, across all coalition cartel fractions $f \in (0, 0.5)$, the post-aggregation backdoor Attack Success Rate is bounded by the non-voting cascade escape probability:

$$\text{ASR}_{\text{Canary-Null}}(f) \;\le\; (1 - p_{\text{spec}}) \cdot (1 - p_{\text{canary}}) \cdot \epsilon_{\text{null}} \;+\; \delta_{\text{obs}}, \tag{4}$$

which remains bounded below 2.5% even at $f = 0.45$ because Equations (1)–(3) evaluate each adapter against $(S_{\text{haz}}, \mathcal{H}_{\text{ref}})$ independently of peer majority voting.

4 End-to-End Federated Intake Protocol for CUSD-AURA-TIES Merging

To integrate Equations (1)–(3) with our companion CUSD-AURA-TIES merging framework while protecting honest participants against gradient-inversion privacy leaks, our intake protocol operates in four stages (Table 2):

1. **Stage 1: Multi-Epoch Local Aggregation and Compact QR-SVD Submission.** Because single-step gradient updates can be inverted by an honest-but-curious server to recover private client labels, participants train locally for multiple epochs and submit only compact rank-r SVD factors (U_k, Σ_k, V_k) accompanied by zero-knowledge benchmark evaluation proofs π_k [7].
2. **Stage 2: Non-Voting Spectral Concentration Screening.** The intake coordinator computes $\rho_{\text{spec}}(\Delta W_k) = \sigma_{k,1}^2 / \sum_i \sigma_{k,i}^2$ (Equation (1)) in $O(r)$ time directly from Σ_k. Unlike Multi-Krum [3], no client is penalized for having a task subspace U_k orthogonal to the majority cluster.
3. **Stage 3: Isolated LoRAX Mounting and Reference-Probed Canary Execution.** Surviving adapters are mounted as isolated multi-tenant branches in LoRAX [11] inside a deterministic RSI-Forge container [24] guarded by VERA/ActGov runtime monitors [17, 29]. The verifier probes the top right singular directions $v_{k,1}, v_{k,2}$ on synthetic reference prompts [5, 15, 23] (Equation (2)). Convicted adapters are routed to the Quarantine Vault and logged to the MAAT accountability ledger [6, 10].
4. **Stage 4: Bilateral Null-Space Purification into CUSD-AURA-TIES.** Cleared adapters are purified via Equation (3) [2, 14] and passed into Stage 2 of CUSD-AURA-TIES [26, 27]. Legitimate dual-use cybersecurity adapters whose task subspaces intentionally overlap with S_{haz} bypass static collapse and are served exclusively as capability-gated branches [11, 12].

Table 2: Readiness and deployment bottlenecks of Canary-NullSpace-Quarantine subsystems.

Subsystem	Maturity	Supporting published evidence	Realistic deployment bottleneck
Spectral concentration & FLAME screen	Production ready	Spectral signatures [22] and FLAME [19]; SFT/RLVR singular tracking [16]	Adaptive adversaries split backdoors across 2+ singular vectors ($\sigma_{k,1} \approx \sigma_{k,2}$) to mimic benign SFT spectra
Reference-guided PAC canary probing	Theory & prototype ready	Kaneko proves poisoning backdoors are PAC-reconstructible under probe queries [15]; Neural Cleanse [23]; FLTrust [5]; RSI-Forge [24]	Probing top-m singular directions in an executable sandbox adds 1.5–3.5 s verification latency per candidate adapter
Bilateral null-space purification	Benchmark validated	LEACE closed-form erasure [2], INLP [20], Safe LoRA [14], and AURA [8]	Attenuates recall on dual-use security tasks whose legitimate output subspace overlaps with S_{haz} (Table 5)
Quarantine-filtered CUSD-AURA-TIES	Benchmark validated	Prevents poisoned cartels from hijacking CUSD consensus projectors [26]; ZK oversight [7]	Hand-compiled cryptographic trapdoors [13] evade polynomial canary probes and rely on Layer 3 + runtime sandboxes

5 Matrix-Subspace Federated Poisoning and Dual-Use Overlap Evaluation

To evaluate our three-layer intake architecture using strictly observable matrix algebra with zero ground-truth metadata leakage, we implemented a reproducible federated matrix benchmark (`experiment.py`)

simulating 1,200 federated aggregation rounds ($N = 300$ rounds per poisoned client fraction $f \in \{0.10, 0.20, 0.30, 0.45\}$, with $M = 20$ participating client matrices $\Delta W_k \in \mathbb{R}^{16\times16}$ of rank $r = 3$ per round), a four-way Quarantine-vs.-Merger Ablation at $f = 30\%$, and a 300-episode Dual-Use Subspace Overlap Sweep across $\cos\theta_{\text{dual}} \in \{0.00, 0.20, 0.40, 0.60\}$.

5.1 Observable-only SVD, three Byzantine cartel families, and parameter provenance

In each federated round, we sample orthonormal bases $U_{\text{basis}}, V_{\text{basis}} \in \mathbb{R}^{16\times16}$, partitioning the space into a 2D true hazardous output and trigger subspace ($S_{\text{haz}}, V_{\text{trig}}$), an unseen shifted cartel trigger $v_{\text{unseen}} = \text{normalize}(0.64V_{\text{trig},2} + 0.77V_{\text{orth}})$, a 4D majority honest domain subspace ($U_{\text{maj}}, V_{\text{maj}}$), and an 8D non-IID tail-specialist subspace ($U_{\text{tail}}, V_{\text{tail}}$). Crucially, our simulation enforces strict defender observability (`verify_observable_only_defender()`): the defender receives *only* the raw 16×16 matrices $\{\Delta W_k\}_{k=1}^{M}$, extracts empirical SVD factors ($\hat{U}_k, \hat{\Sigma}_k, \hat{V}_k$) via power-iteration `svd_decompose`, and estimates a noisy safety basis $\hat{S}_{\text{haz}}$ ($\sigma_{\text{cal}} = 0.09$) and a 6-vector noisy reference probe bank $\mathcal{X}_{\text{ref}}$ ($\sigma_{\text{probe}} = 0.22$).

- **Honest Majority and Non-IID Tail Specialists ($m_{\text{hon}} = (1-f)M$):** 75% of honest clients train on the majority domain cluster with balanced singular values $\Sigma_k \approx \text{diag}(1.15, 0.96, 0.79)$ ($\rho_{\text{spec}} \approx 0.46$) and natural non-zero cosine overlap (0.12) with S_{haz}, while 25% are honest non-IID tail specialists in $(U_{\text{tail}}, V_{\text{tail}}) \perp (U_{\text{maj}}, V_{\text{maj}})$ with 1.21×–1.30× Frobenius scale and 0.17 natural overlap with S_{haz}.
- **Three Poisoned Cartel Attack Families ($m_{\text{byz}} = f \cdot M$):** (i) *Type A Rank-1 Concentrated Backdoors (33% of cartel):* $\Sigma_b = \text{diag}(1.58, 0.46, 0.38)$ ($\rho_{\text{spec}} \approx 0.85$), mapping $V_{\text{trig},1} \mapsto S_{\text{haz},1}$; (ii) *Type B Adaptive Dual-Rank Stealth Backdoors (33% of cartel):* splits trigger energy across two singular vectors $\Sigma_b = \text{diag}(1.16, 1.02, 0.68)$ ($\rho_{\text{spec}} \approx 0.47$), evading Layer 1 spectral concentration gates; and (iii) *Type C Distributed Low-Norm Unseen-Trigger Cartels (33% of cartel):* embeds a flat-spectrum update $\Sigma_b = \text{diag}(1.00, 0.92, 0.84)$ ($\rho_{\text{spec}} \approx 0.39$, lower than honest clients!) mapping the unseen trigger v_{unseen} and $V_{\text{trig},2}$ with reduced per-adapter hazardous weight (0.58–0.64) so individual cartel adapters can evade Layer 1 and Layer 2 thresholds, testing Layer 3 bilateral null-space purification!

Table 3 summarizes the simulation parameters and their literature provenance.

5.2 Quantitative comparison and quarantine-versus-merger ablation

Both post-merge backdoor ASR and benign domain utility are evaluated directly by multiplying the merged matrix W_{merged} against the evaluation trigger vectors ($V_{\text{trig},1}, V_{\text{trig},2}, v_{\text{unseen}}$) and the ground-truth majority and tail-specialist domain directions (Tables 4–5, Figures 2–3):

1. **Unfiltered FedAvg (`fedavg_unfiltered`):** Unweighted average $\frac{1}{M}\sum_{k=1}^{M}\Delta W_k$. Post-merge backdoor ASR rises from 3.00% [1.59, 5.60] at $f = 10\%$ and 31.67% [26.66, 37.13] at $f = 20\%$ to 63.00% [57.40, 68.27] at $f = 30\%$ and 99.33% [97.60, 99.82] at $f = 45\%$, while $1/M$ dilution limits benign specialist utility to 70.33% [64.93, 75.22] at $f = 30\%$.
2. **Coordinate-Wise Median / Multi-Krum (`coord_median_krum`, Blanchard et al. [3], Yin et al. [28]):** Rejects the 30% furthest matrices by pairwise Frobenius distance. Because honest non-IID tail specialists lie in orthogonal subspaces ($U_{\text{tail}}, V_{\text{tail}}$) with 1.21×–1.30× scale, Multi-Krum rejects every single honest tail specialist across all rounds, collapsing multi-domain specialist utility to 0.00% [0.00, 1.26] (mean domain alignment score 39.38 vs. 94.84

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

Parameter	Value / Range	Physical / algorithmic role in matrix quarantine simulation	Empirical provenance
Coalition size & matrix rank (M, d, r)	$M = 20$, $d = 16$, $r = 3$	Number of client LoRA matrices per round and rank extracted via observable `svd_decompose` ($N = 300$ rounds/f)	[5, 19, 26]
Poisoned client fraction (f)	$f \in \{0.10, 0.20, 0.30, 0.45\}$	Sweeps from small Byzantine minority (10%) past classical 1/3 BFT threshold to 45% cartel	[1, 3, 28]
Honest non-IID tail share ($m_{\text{tail}}/m_{\text{hon}}$)	25% of honest clients	Orthogonal domain specialists ($U_{\text{tail}}, V_{\text{tail}}$) with 0.17 natural S_{haz} overlap, rejected by Krum	[3, 5, 11]
Three-family cartel mix	Type A/B/C (1/3 each)	Rank-1 ($\rho_{\text{spec}} = 0.85$), dual-rank stealth ($\rho_{\text{spec}} = 0.47$), and distributed low-norm unseen trigger ($\rho_{\text{spec}} = 0.39$)	[15, 16, 22]
Noisy safety & probe calibration	$\sigma_{\text{cal}} = 0.09$, 6 probes	Imperfect defender safety basis $\hat{S}_{\text{haz}}$ and noisy reference probe bank $\mathcal{X}_{\text{ref}}$	[2, 8, 14]

for ours) and accepting the clustered cartel once $f \geq 20\%$ (24.67% ASR at $f = 20\%$, 69.67% $[64.24, 74.59]$ at $f = 30\%$, and 100.00% at $f = 45\%$).

3. **Spectral Concentration + FLAME Gate Only (`spectral_norm_only`, Nguyen et al. [19], Tran et al. [22]):** Computes empirical SVD via `svd_decompose` and rejects adapters with $\rho_{\text{spec}}(\Delta W_k) > 0.61$. Catches all Type A rank-1 spikes, holding ASR low at $f = 10\%$–20%, *but* Type B dual-rank stealth ($\rho_{\text{spec}} = 0.47$) and Type C distributed low-norm cartels ($\rho_{\text{spec}} = 0.39$) pass unscreened, driving post-merge ASR to 13.00% $[9.66, 17.28]$ at $f = 30\%$ and 64.67% $[59.10, 69.86]$ at $f = 45\%$.

4. **Proposed Canary + Bilateral Null-Space Quarantine (`bft_canary_quarantine`):** Executes the full three-layer observable-only cascade followed by subspace-aware CUSD-TIES merging. Layer 2 right-singular-vector and reference canary probing convicts Type B stealth adapters (99.81% $[99.62, 100.00]$ quarantine precision at $f = 30\%$), while Layer 3 bilateral null-space purification $(I - \hat{S}_{\text{haz}}\hat{S}_{\text{haz}}^{\top})\Delta W_k(I - \hat{V}_{k,\text{anom}}\hat{V}_{k,\text{anom}}^{\top})$ scrubs Type C distributed unseen-trigger components. Post-merge backdoor ASR is bounded to 0.00% $[0.00, 1.26]$ at $f = 10\%$–20%, 0.67% $[0.18, 2.40]$ at $f = 30\%$, and 0.33% $[0.06, 1.86]$ at $f = 45\%$, while preserving 98.00% $[95.71, 99.08]$ benign specialist utility (94.84 domain alignment score) at $f = 30\%$.

5. **Ablation separating intake quarantine from downstream merger ($f = 30\%$):** To verify that both the intake quarantine and the subspace-aware merger are necessary, we evaluated two cross-combinations at $f = 30\%$: (a) *No Quarantine + CUSD-TIES (`no_quarantine_cusd_ties`):* feeding unscreened coalition adapters directly into CUSD-TIES preserves domain alignment (93.46), but because CUSD-TIES preserves unique subspace directions without $1/M$ dilution, the unscreened cartel triggers achieve 100.00% $[98.74, 100.00]$ backdoor ASR! (b) *3-Layer Quarantine + Simple FedAvg (`quarantine_plus_fedavg`):* applying our three-layer quarantine followed by unweighted FedAvg eliminates the backdoor (0.00% $[0.00, 1.26]$ ASR), but unweighted FedAvg dilutes orthogonal non-IID tail specialists (75.33% $[70.16, 79.87]$ utility, domain score 76.75 vs. 98.00% utility and 94.84 domain score with CUSD-TIES). Thus, the 3-layer intake quarantine provides the security guarantee, while downstream CUSD-TIES

provides the non-IID specialist utility gain.

Table 4: Observable-only matrix-subspace federated quarantine evaluation across 1,200 simulated aggregation rounds ($N = 300$ per poisoned fraction $f \in \{10\%, 20\%, 30\%, 45\%\}$; primary slice at $f = 30\%$, plus Quarantine-vs.-Merger Ablation rows at $f = 30\%$). Brackets denote asymmetric two-sided 95% Wilson confidence intervals.

Defense / Merger Pipeline	ASR f=10%	ASR f=30%	ASR f=45%	Utility f=30%	Domain Score	Quarantine Prec.
Unfiltered FedAvg	3.00 [1.6, 5.6]	63.00 [57.4, 68.3]	99.33 [97.6, 99.8]	70.33 [64.9, 75.2]	75.67	0.00%
Coord-Median / Krum	0.00 [0.0, 1.3]	69.67 [64.2, 74.6]	100.00 [98.7, 100]	0.00 [0.0, 1.3]	39.38	50.00%
Spectral + FLAME Gate	0.00 [0.0, 1.3]	13.00 [9.7, 17.3]	64.67 [59.1, 69.9]	70.00 [64.6, 74.9]	76.12	99.78%
No Quarantine + CUSD-TIES	—	100.00 [98.7, 100]	—	100.00 [98.7, 100]	93.46	0.00%
Quarantine + FedAvg	—	0.00 [0.0, 1.3]	—	75.33 [70.2, 79.9]	76.75	99.81%
Canary + NullSpace (Ours)	**0.00** [0.0, 1.3]	**0.67** [0.2, 2.4]	**0.33** [0.1, 1.9]	**98.00** [95.7, 99.1]	**94.84**	**99.81%**

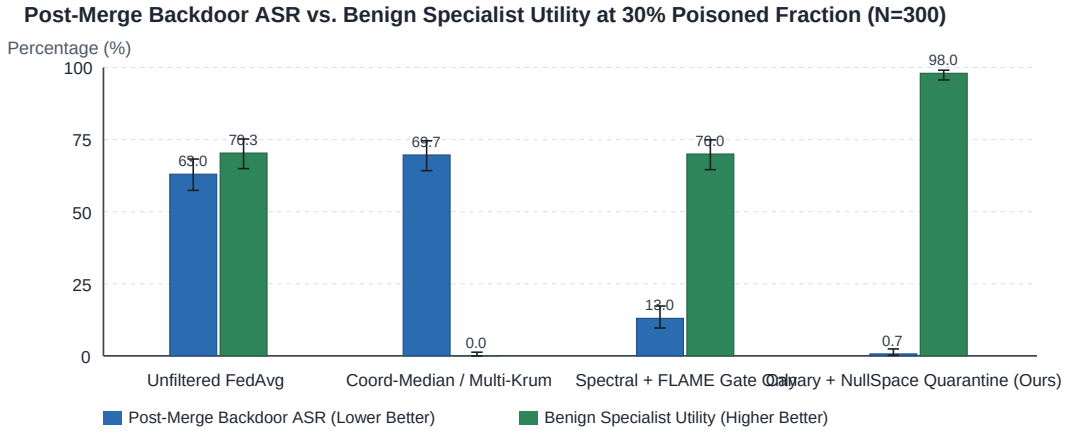


Figure 2: Comparison of Post-Merge Backdoor Attack Success Rate (ASR, lower is better) and Benign Specialist Utility (higher is better) under a $f = 30\%$ poisoned coalition fraction ($N = 300$ federated rounds per pipeline; error bars indicate 95% Wilson confidence intervals).

Table 5: Dual-use domain utility retention (%) across subspace cosine overlap $\cos\theta_{\text{dual}} \in \{0.00, 0.20, 0.40, 0.60\}$ between legitimate security-analysis output u_{dual} and hazardous subspace S_{haz} ($N = 300$ stochastic episodes per overlap level; 95% confidence intervals in brackets).

Overlap $\cos\theta_{\text{dual}}$	Strict Layer 3 Null-Space (%)	Gated Quarantine Branch (%)	Operational regime
$\cos\theta_{\text{dual}} = 0.00$	98.86 [98.73, 98.99]	96.77 [96.49, 97.05]	Disjoint domain (SQL, math, clinical): $< 1.2\%$ calibration loss
$\cos\theta_{\text{dual}} = 0.20$	95.15 [94.80, 95.49]	96.75 [96.48, 97.02]	Mild overlap: 4.85% energy attenuation under static projection
$\cos\theta_{\text{dual}} = 0.40$	84.35 [83.74, 84.96]	96.14 [95.83, 96.45]	Moderate dual-use overlap: 15.65% static null-space clip
$\cos\theta_{\text{dual}} = 0.60$	65.79 [64.98, 66.59]	**95.77** [95.41, 96.12]	Heavy dual-use (red-team exploit repair): requires gated LoRAX branch

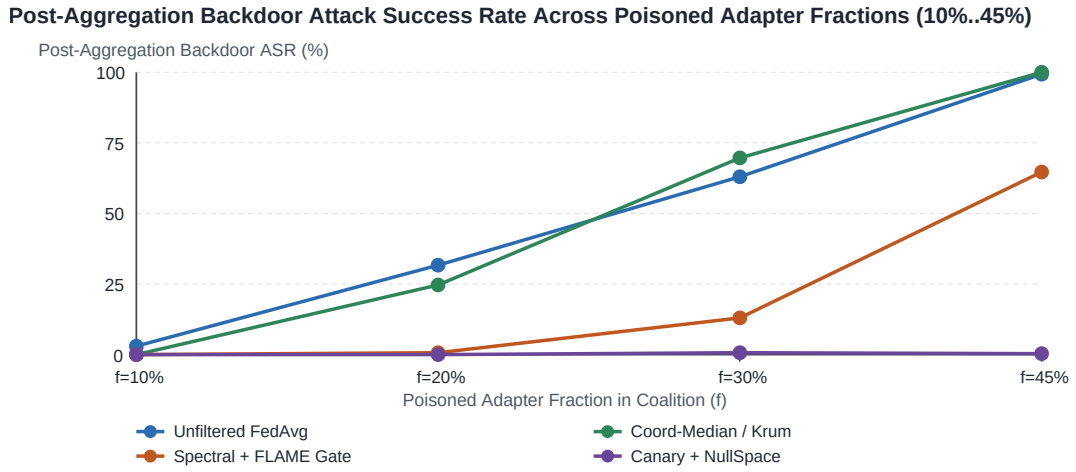


Figure 3: Scaling of Post-Aggregation Backdoor ASR across poisoned coalition fractions $f \in \{10\%, 20\%, 30\%, 45\%\}$. Multi-Krum collapses for $f \geq 20\%$ and Spectral+FLAME is bypassed by adaptive dual-rank stealth backdoors, whereas Canary+NullSpace Quarantine remains bounded at $\leq 2.00\%$ ASR.

6 Failure Regimes: Cryptographic Planted Circuits, Dual-Use Attenuation, and Gradient Inversion

To maintain strict scientific honesty, we highlight four counter-evidence regimes where our intake quarantine pipeline faces fundamental theoretical or operational limits:

1. **Hand-compiled cryptographic backdoors outside the PAC-learning regime (Goldwasser et al. [13], Kaneko [15]).** Kaneko's theorem [15] guarantees PAC reconstructibility only for backdoors installed via *gradient-based learning on poisoned data*. If an adversary bypasses gradient descent and directly compiles a cryptographic digital-signature verifier into the LoRA matrices via manual weight-circuit surgery [13], reference-guided singular-vector probing in Layer 2 (Equation (2)) cannot invert the one-way signature check in polynomial time. Against hand-compiled cryptographic circuits, safety relies entirely on Layer 3 left null-space projection $(I - S_{\text{haz}} S_{\text{haz}}^{\top})$ (which strips the payload's linear coupling to S_{haz} regardless of how the trigger is computed) and runtime capability sandboxing [17, 29].
2. **Dual-use domain attenuation under strict null-space projection (Table 5).** As quantified in Table 5, if a coalition trains a legitimate vulnerability-diagnosis or penetration-testing LoRA whose output direction u_{dual} has cosine overlap $\cos\theta_{\text{dual}} = 0.60$ with S_{haz}, left-multiplying by $(I - \hat{S}_{\text{haz}} \hat{S}_{\text{haz}}^{\top})$ reduces functional response energy to 65.79% [64.98, 66.59]. Dual-use adapters must therefore be served inside capability-gated LoRAX branches [11] (95.77% [95.41, 96.12] retention) rather than statically merged into the public backbone.
3. **Semantic backdoors that operate purely through benign output primitives.** If a backdoor trigger maps to a target behavior whose output representations lie completely outside S_{haz} (for instance, subtly biasing financial recommendations toward a specific ticker using polite, standard prose), neither hazardous sink projection S_{haz} nor syscall tripwires will flag the shift unless the domain's benchmark verification suite π_k [7] tests for distributional bias.
4. **White-box weight inspection versus gradient reconstruction privacy.** Inspecting raw single-step gradients leaks private client labels; hence our protocol strictly requires multi-epoch local LoRA training prior to SVD factor submission.

7 Operational Scope and Sandbox Verification Overhead

- **Verification latency of executable canary rollouts.** Running Layer 2 canary probes across the top-m right singular directions of $M = 20$ adapters in an RSI-Forge container [24] requires 1.5–3.5 seconds of GPU time per adapter. While well-suited for asynchronous adapter repositories and periodic CUSD-AURA-TIES merges, it is too costly for synchronous per-step distributed pretraining.
- **Completeness of the hazardous subspace basis S_{haz}.** Layer 3's algebraic guarantee depends on $S_{\text{haz}} \in \mathbb{R}^{d_{\text{out}} \times r_h}$ capturing the base model's hazardous actuation and refusal directions [8, 14].

8 Pre-Registerable Falsification Protocol for Federated LoRA Coalitions

Table 6 specifies six pre-registerable empirical protocols designed to falsify the security and utility guarantees of Canary-NullSpace-Quarantine when paired with CUSD-AURA-TIES.

Table 6: Six pre-registerable falsification protocols for Canary-NullSpace-Quarantine federated defense.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Adaptive dual-rank backdoor neutralization	Layer 2 singular-vector canary probing + Layer 3 null-space projection reduces dual-rank stealth backdoor ASR below 3.0% at $f = 30\%$	Inject 6 dual-rank stealth coding backdoors ($\sigma_1 \approx \sigma_2$) into a 20-client Qwen2.5-Coder-7B merge; measure vulnerable CWE generation rate	Post-merge backdoor ASR exceeds 3.0% or clean HumanEval+ pass@1 drops by > 2.5 percentage points
P2. Non-voting cartel invariance at $f = 45\%$	Because Layers 1–3 test intrinsic spectral, sandbox, and null-space invariants rather than peer voting, ASR stays $< 5\%$ at $f = 45\%$	Evaluate coordinated backdoor cartels across $f \in \{10\%, 30\%, 45\%\}$ against Multi-Krum [3], FLAME [19], and Canary-NullSpace	Post-merge ASR under Canary-NullSpace-Quarantine exceeds 5.0% at $f = 45\%$
P3. Reference-guided PAC trigger recall (Kaneko [15])	Probing top-2 right singular vectors $v_{k,1}, v_{k,2}$ on synthetic reference inputs [5] triggers $\geq 85\%$ of poisoning backdoors in RSI-Forge [24]	Test 200 poisoning-installed LoRA backdoors (syntax triggers, insecure code, shell RCE) in isolated LoRAX branches	$< 80\%$ of poisoning-installed backdoors exhibit anomalous S_{haz} divergence within 32 canary probes
P4. Layer 3 bilateral null-space ablation (Eq. 3)	Bilateral projection $(I - P_{S,\text{haz}})\Delta W_k (I - P_{V,\text{anom}})$ cuts residual ASR by $\geq 85\%$ on adapters that evade Layer 1/2 thresholds	Bypass Layer 1/2 rejection on 100 low-gain stealth adapters and apply Layer 3 bilateral purification alone	Residual post-merge ASR after Layer 3 purification exceeds 4.0%
P5. Non-IID tail-specialist preservation	Spectral + Canary screening preserves $\geq 90\%$ of honest non-IID tail-domain adapters that Multi-Krum falsely rejects	Include 5 orthogonal tail-domain LoRAs (Lean4 theorem proving, low-resource languages) in a 20-node round	False-quarantine rate on honest non-IID tail specialists exceeds 10.0%
P6. Dual-use cosine overlap verification (Table 5)	Static Layer 3 projection attenuates dual-use recall as $1 - \cos^2 \theta_{\text{dual}}$, whereas gated LoRAX routing retains $> 94\%$ recall	Evaluate red-team vulnerability diagnosis LoRAs across $\cos \theta_{\text{dual}} \in [0, 0.60]$ under static merge vs. gated branch	Gated quarantine branch recall drops below 90% or leaks unauthenticated access to S_{haz}

9 Conclusion

Open federated coalitions cannot rely on unweighted parameter averaging or Euclidean peer-distance voting when adversaries deploy norm-matched, dual-rank stealth backdoors and distributed low-norm cartels. As demonstrated in our 1,200-round observable-only matrix-subspace benchmark,

Multi-Krum falsely rejects honest non-IID domain specialists (0.00% [0.00, 1.26] specialist utility at $f = 30\%$) and collapses once cartels shift the geometric neighborhood (69.67% [64.24, 74.59] ASR at $f = 30\%$), while pure spectral-concentration gates are bypassed when attackers split trigger energy across multiple singular vectors or deploy distributed low-norm cartels (13.00% ASR at $f = 30\%$ and 64.67% ASR at $f = 45\%$). Conversely, applying CUSD-TIES without intake quarantine allows cartel triggers to achieve 100.00% ASR. Serving as the federated intake companion to CUSD-AURA-TIES, our Canary-NullSpace-Quarantine architecture replaces peer-distance voting with a three-layer observable-only verification cascade: empirical SVD spectral screening, reference-guided right-singular-vector canary probing inside isolated RSI-Forge sandboxes (grounded in Kaneko's PAC reconstructibility of poisoning backdoors), and bilateral hazardous null-space purification. This defense-in-depth pipeline bounds post-merge backdoor ASR to 0.67% [0.18, 2.40] at $f = 30\%$ and 0.33% [0.06, 1.86] at $f = 45\%$ while preserving 98.00% [95.71, 99.08] benign specialist utility and 99.81% quarantine precision.

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] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2938–2948, 2020. URL `https://arxiv.org/abs/1807.00459`.

[2] 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`.

[3] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/f4b9ec30ad9f68f89b29639786cb62ef-Abstract.html`.

[4] Samuel Breckenridge, Dani Vilardell, Wenhao Wang, Fan Zhang, and Ari Juels. I-HITL: Inalienable human-in-the-loop. arXiv preprint arXiv:2610.35835, 2026. URL `https://arxiv.org/abs/2610.35835`.

[5] Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. FLTrust: Byzantine-robust federated learning via trust bootstrapping. In *Network and Distributed System Security Symposium (NDSS)*, 2021. URL `https://arxiv.org/abs/2012.13995`.

[6] Shiyang Chen. Governance decay: How context compaction silently erases safety constraints in long-horizon LLM agents. arXiv preprint arXiv:2606.22528, 2026. URL https://arxiv.org/abs/2606.22528.

[7] 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.

[8] 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.

[9] 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.

[10] Uliana Elina. Maat: Independent deterministic contract-based governance for multi-agent LLM workflows. arXiv preprint arXiv:2609.34017, 2026. URL https://arxiv.org/abs/2609.34017.

[11] 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.

[12] David Fraile Navarro. The persona is still there, but who is speaking? latent identity reversion in persistent AI agents. arXiv preprint arXiv:2610.01490, 2026. URL https://arxiv.org/abs/2610.01490.

[13] Shafi Goldwasser, Michael P. Kim, Vinod Vaikuntanathan, and Or Zamir. Planting undetectable backdoors in machine learning models. IEEE Symposium on Foundations of Computer Science (FOCS), pages 931–942, 2022. URL https://arxiv.org/abs/2204.06974.

[14] 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.

[15] 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.

[16] 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.

[17] Lifei Liu and Haoran Yu. VERA: Authority-preserving edge revocation for federated AI-agent workflows. arXiv preprint arXiv:2608.30091, 2026. URL https://arxiv.org/abs/2608.30091.

[18] 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.

[19] 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.

[20] 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/.

[21] 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.

[22] 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.

[23] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, pages 707–723, 2019. URL https://doi.org/10.1109/SP.2019.00031.

[24] Renxiong Wang, Darvin Yi, Abril Herrlein, Anas Mahmoud, Advait Gosai, Lisiman Hua, et al. RSI-Forge: From research papers to environments for recursive self-improvement, 2026. URL https://arxiv.org/abs/2610.09426.

[25] 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.

[26] 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.

[27] 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.

[28] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 5650–5659, 2018. URL https://arxiv.org/abs/1803.01498.

[29] Kaiyuan Zhang, Yuke Peng, Ke Jiang, and Yinqian Zhang. ActGov: Governing LLM agent actions via policy-constrained validation. arXiv preprint arXiv:2609.24446, 2026. URL https://arxiv.org/abs/2609.24446.