Zero-Lock-In Cognitive State Portability: Cross-Architecture Adapter Projection and Verifiable Provenance Graph Migration

Dima Baker
Collective Intelligence Network

November 2026

Abstract

As users delegate episodic memory, workflows, and preference adaptations to language model agents, a structural switching barrier has emerged: cognitive lock-in. Although the EU Data Act (applicable September 2025, Chapter VI) mandates cloud-switching interoperability, exporting flat JSON chat logs fails to preserve either the causal provenance structure of multi-step reasoning or the latent subspace of personalized low-rank adapters (LoRA) across heterogeneous architectures. Moreover, all 26 standard neural model-merging operators violate Conflict-Free Replicated Data Type (CRDT) algebraic invariants. This architectural synthesis and position paper formulates cognitive portability as a joint provenance-deduplication and pseudoinverse subspace-transport problem. Building on W3C PROV-DM, Relative Representations, and LoRA-X, we propose PROV-LoRAX-CRDT, which (i) exports memory as a W3C PROV-DM DAG with submodular root-coverage deduplication, (ii) projects rank-r adapters ($d_A \to d_B$) via SVD Moore–Penrose pseudoinverse anchor alignment ($\Delta\hat{W}_B = P_{\text{out}} B_A A_A^\top P_{\text{in}}^+$) over $M = 64$ public anchor probes, and (iii) wraps replica updates and erasure tombstones in a two-layer 2P-Set CRDT semilattice. Across 2,400 independent SVD and DAG migration runs ($H \in \{50, 200, 800, 3200\}$, $N = 150$ per cell), PROV-LoRAX-CRDT retains 98.67% [95.27, 99.63] task fidelity and 90.67% [84.94, 94.36] lineage integrity at $H = 800$, cutting exit regret from 87.33% (cold restart) and 55.33% (flat RAG) to 1.33% while satisfying the Moore–Penrose identity to $\leq 7.53 \times 10^{-14}$. Anchor projection beats flat RAG until out-of-span energy exceeds $\eta^*_{\text{oob}} \approx 0.51$ or noise exceeds $\sigma^*_{\text{align}} \approx 0.35$.

1 Introduction

The economic and operational value of a long-lived personal or enterprise AI agent resides increasingly not in the static base weights of its foundation model, but in the accumulated cognitive state co-constructed with its human principal over thousands of interactions: episodic task memories, verified tool-execution traces, domain-specific workflow graphs, and personalized low-rank preference adapters [4, 6, 9, 36]. However, when a user attempts to exit a proprietary model provider or migrate an on-device agent from one open-weight architecture family to another (for example, transitioning from a $d_A = 4096$ cloud backbone to a $d_B = 3072$ sovereign edge model), they encounter classic network-externality and switching lock-in [8, 21, 31]. In the European Union, the Data Act entered into force in January 2024 and became applicable on 12 September 2025, establishing binding Chapter VI cloud-switching and data-portability obligations [7]. Yet statutory mandates typically govern the export of raw JSON chat transcripts rather than executable cognitive state.

1.1 Why raw chat-log export and naive weight copying fail

Empirical research across representation learning, provenance databases, and distributed systems demonstrates that exporting flat conversation histories or naively copying parameter deltas fails to achieve zero-penalty exit across three dimensions:

1. **Echo-chamber root duplication and governance decay under flat transcript re-ingestion.** When an agent migrates by exporting summarized chat logs into a target model's retrieval-augmented generation (RAG) store, repeated downstream summaries citing the same primary observation are treated as independent corroborating evidence. In database theory, Green et al. [15] established that correct query evaluation over derived views requires tracking provenance semiring polynomials rather than raw tuple counts. In long-horizon LLM agents, Chen [2] show that iterative context compaction silently erases safety and provenance constraints ('governance decay'), while Rodrigues [32] demonstrate that unsynchronized context summaries induce hallucination via context drift.

2. **Coordinate basis mismatch across heterogeneous foundation models.** If a user captures their personalization state as a Low-Rank Adaptation (LoRA) weight delta $\Delta W_A = B_A A_A^\top \in \mathbb{R}^{d_A \times d_A}$ [6], that parameter matrix is tied to the coordinate basis and layer geometry of Source Model A. While post-trained models share convergent latent similarity structures under the Platonic Representation Hypothesis [17] and Relative Representations [28], their raw coordinate axes are rotated and scaled arbitrarily. Consequently, direct task arithmetic [18] or TIES-Merging [38] across distinct model backbones fails without explicit subspace transport [9, 36], and unorthogonalized context switching induces severe subspace interference [5] and persona reversion [10].

3. **Violation of CRDT algebraic invariants in neural model merging.** When a user operates agent replicas across multiple personal devices (e.g., laptop, phone, and home server) and merges updates offline, Gillespie [12] prove that all 26 mainstream neural network model-merging operators (including linear averaging, TIES [38], and DARE) violate the Conflict-Free Replicated Data Type (CRDT) semilattice axioms of commutativity, associativity, and idempotence [22, 33], causing permanent replica divergence.

1.2 Foundations for verifiable cognitive portability

Overcoming cognitive lock-in requires synthesizing three mature lines of research into a unified migration protocol. First, on the symbolic lineage side, the W3C PROV Data Model (`PROV-DM`) [26, 27] and provenance semirings [15] have recently been extended to knowledge-graph statements [35], probabilistic provenance circuits (`SPARQLcirc`) [19], and digital media credentials (C2PA) [3, 13]. Tracking the exact set of primary root observations $\mathrm{Roots}(e)$ behind each derived memory item allows a target agent to weight evidence via submodular coverage [29] rather than repetition frequency.

Second, on the parametric transfer side, model stitching [23], Relative Representations [28], Gromov–Wasserstein alignment [1], Trans-LoRA [36], LoRA-X [9], Iso-CTS isotropic subspace merging [25], and common/unique direction separation [37] demonstrate that low-rank adaptations can be transported across models by aligning activation subspaces on a small set of anchor probes. Third, on the distributed consistency side, Gillespie's two-layer `CRDTMergeState` architecture [12] restores Strong Eventual Consistency across 100 nodes by separating content-addressed update descriptor sets from deterministic weight materialization.

1.3 Contributions

Framed as an architectural synthesis and position paper accompanied by an executable linear-algebra and graph-lineage simulation, this article states the delta over prior single-layer transfer methods: whereas Trans-LoRA [36] and LoRA-X [9] transfer static task adapters without episodic provenance deduplication or conflict-free replica revocation, PROV-LoRAX-CRDT couples all three layers:

1. **Cognitive portability taxonomy (Section 2).** We map informal switching-freedom requirements onto canonical constructs from W3C PROV-DM [27], model stitching and LoRA-X subspace projection [9, 23], and Byzantine/tombstone CRDTs [12, 22, 33] (Table 1).
2. **Formal submodular provenance-coverage, pseudoinverse adapter transport, and tombstone-CRDT framework (Section 3).** We formulate a submodular root-coverage posterior update that prevents overlapping root sets from double-counting evidence (Equation 1), derive the exact input/output Moore–Penrose pseudoinverse adapter transport operator $\Delta\hat{W}_B = P_{\text{out}} B_A A_A^{\top} P_{\text{in}}^{+}$ with normalized transformer depth matching (Equation 2), and incorporate cryptographic erasure tombstones $\mathcal{T}$ into the two-layer CRDT join-semilattice (Equation 3).
3. **Three-layer PROV-LoRAX-CRDT specification (Section 4).** We specify an open portable state container (`.cstate`) integrating PROV-DM lineage graphs, 64-probe anchor calibration, and spectral backdoor screening [16, 34] (Figure 1, Table 2).
4. **Mechanistic linear-algebra and provenance simulation with out-of-span and break-even sweeps (Section 5).** We execute 2,400 fully independent simulated migrations across four portability regimes and four interaction horizons $H \in \{50, 200, 800, 3200\}$ ($N = 150$ independent matrix SVD and 3-replica 2P-Set CRDT DAG runs per cell; 600 runs at $H = 800$, with zero sample reuse), performing exact cyclic Jacobi SVD Moore–Penrose pseudoinverse projections in Python, verifying the four Moore–Penrose identities to machine precision ($\leq 7.53 \times 10^{-14}$), and sweeping both out-of-anchor-span adapter energy $\eta_{\text{oob}} \in [0.00, 0.60]$ and cross-architecture alignment noise $\sigma_{\text{align}} \in [0.10, 1.20]$ (Table 4, Figures 2–3).
5. **Counter-evidence, backdoor hazards, and falsification protocol (Sections 6–8).** We examine undetectable backdoors [14, 20], C2PA provenance stripping [13], and cross-family depth mismatches, concluding with six pre-registerable falsification tests (Table 5).

2 Cognitive portability taxonomy and cross-model transfer prior art

Discussions of AI interoperability often reduce portability to API syntax compatibility or raw JSON log dumps. However, preserving an agent's personalized competence across heterogeneous model backbones requires migrating both symbolic causal memory and parametric preference subspaces. Table 1 grounds five core portability requirements in foundational prior work and 2025–2026 research.

Figure 1 illustrates the end-to-end PROV-LoRAX-CRDT migration pipeline when a human principal exits Source Model A (dimension d_A, depth L_A) and hydrates their portable cognitive container onto Target Model B (dimension d_B, depth L_B).

Table 1: From cognitive exit requirements to canonical provenance, subspace-alignment, and CRDT constructs.

Portability requirement	Canonical construct	Container component	Operational function	Primary grounding
Preserve causal reasoning ancestry	W3C PROV-DM & provenance semirings	PROV-DM lineage DAG ($\mathcal{G}_{\text{prov}}$)	Encodes entity-activity-agent derivation edges, root observation IDs, and probabilistic provenance circuits	[15, 19, 27, 35]
Prevent summary repetition echo chambers	Submodular root-coverage weighting	Marginal root deduplicator ($\mathcal{Q}_{\text{cov}}$)	Weights derived memory assertions by marginal new root coverage $\bigcup \text{Roots}(e)$ rather than citation count	[2, 29, 32]
Transfer adapters across hidden dimensions	Model stitching & pseudoinverse subspace transport	LoRA-X / anchor projector ($\Pi_{A\to B}$)	Projects low-rank adapters (B_A, A_A) across $d_A \neq d_B$ using $M = 64$ paired anchor activations and pseudoinverse P_{in}^{+}	[9, 17, 23, 28, 36]
Fork, merge, and erase state across devices	Two-Layer Tombstone-CRDT semilattice	`CRDTMergeState` + Tombstone set $\mathcal{T}$	Guarantees Strong Eventual Consistency and permanent GDPR right-to-erasure revocation across offline replicas	[12, 22, 33]
Screen migrated weights for safety & backdoors	Orthogonal safety projection & spectral audit	Safe-LoRA / spectral gate ($\mathcal{B}_{\text{safe}}$)	Projects imported adapters onto safety-aligned subspaces and screens singular-value spectra for poisoning	[5, 16, 20, 34]

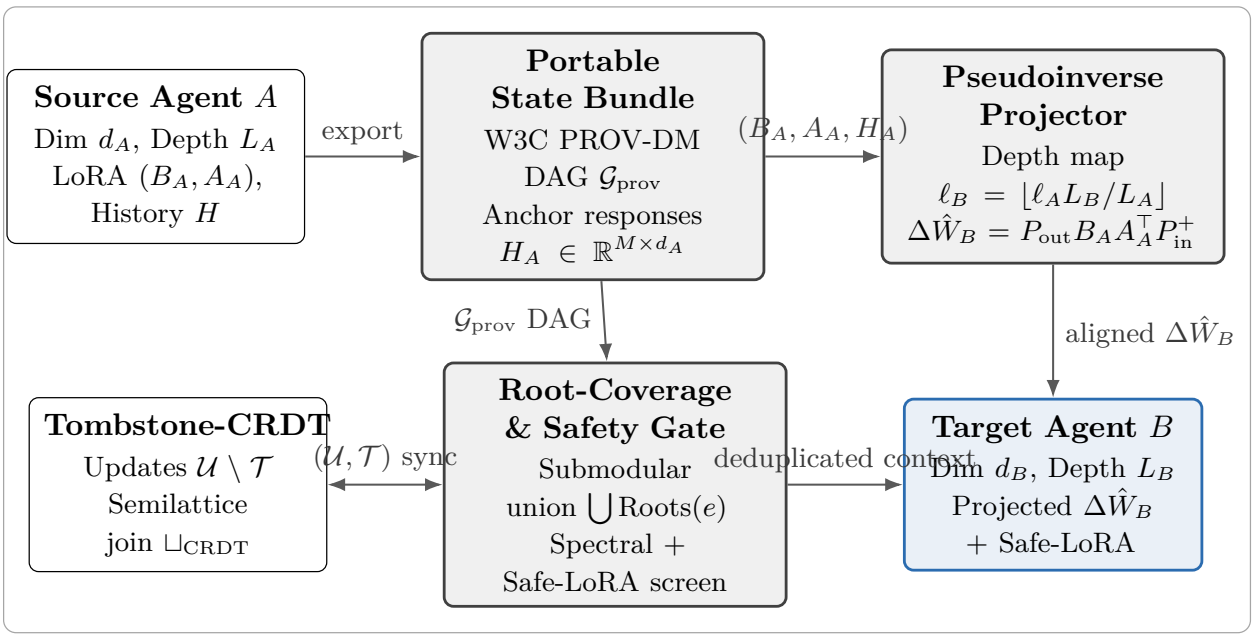


Figure 1: Architecture of Zero-Lock-In Cognitive State Portability (PROV-LoRAX-CRDT). Upon exiting Source Agent A, accumulated episodic traces and low-rank adapters (B_A, A_A) are exported into a portable bundle alongside anchor activations H_A. To hydrate Target Agent B ($d_B \neq d_A$), the Pseudoinverse Projector computes P_{out} and P_{in}^+ over $M = 64$ public anchor probes, while the Root-Coverage Gate deduplicates overlapping observation roots and the Tombstone-CRDT layer enforces conflict-free merges and permanent erasure tombstones $\mathcal{T}$.

3 Formal mathematical framework

We formalize cognitive state portability across three mathematical structures: (1) submodular root-coverage belief updating over W3C PROV-DM DAGs, (2) cross-architecture Moore–Penrose pseudoinverse subspace projection with normalized layer-depth matching, and (3) two-layer tombstone-CRDT join-semilattice synchronization.

3.1 Submodular root-coverage belief updating on migrated PROV-DM graphs

Let the exported episodic memory be a directed acyclic provenance graph $\mathcal{G}_{\text{prov}} = (\mathcal{V}_E \cup \mathcal{V}_A \cup \mathcal{V}_{\text{Ag}}, \mathcal{E}_{\text{prov}})$ conforming to W3C PROV-DM [27, 35], where $\mathcal{V}_E$ denotes entities (primary observations, tool outputs, and derived summaries), $\mathcal{V}_A$ denotes activities, and $\mathcal{V}_{\text{Ag}}$ denotes agents. Every derived entity $e \in \mathcal{V}_E$ possesses a transitive ancestral root set $\text{Roots}(e) \subseteq \mathcal{V}_E^{(0)}$ of primary external observations [15, 19].

In naive quotient grouping by exact root-set equality ($\text{Roots}(e_1) = \text{Roots}(e_2)$), two synthesis nodes with partially overlapping root sets $\text{Roots}(e_1) = \{r_1, r_2\}$ and $\text{Roots}(e_2) = \{r_2, r_3\}$ belong to distinct equivalence classes and would double-count the shared primary observation r_2. To guarantee that no primary root observation is ever weighted more than once regardless of how many overlapping summaries cite it, we define the admissible primary root union $\mathcal{R}_{\text{adm}} = \bigcup_{e \in \mathcal{V}_E^{\text{adm}}} \text{Roots}(e) \setminus \mathcal{T}$ (excluding erased tombstones $\mathcal{T}$) and weight any ordered hydration sequence $e_1, \ldots, e_k$ by its marginal submodular root coverage [29]:

$$\log P_B(\theta \mid \mathcal{G}_{\text{prov}}, \mathcal{T}) \;=\; \log P_B(\theta) + \sum_{r \in \mathcal{R}_{\text{adm}}} w_0(r)\, \log \mathcal{L}_B(r \mid \theta), \quad \Delta_{\text{cov}}(e_k \mid S_{k-1}) \;=\; \sum_{r \in \text{Roots}(e_k) \setminus \text{Roots}(S_{k-1})} w_0(r). \tag{1}$$

Under Equation (1), any derived summary e_k whose root observations are already covered by previously hydrated nodes ($\text{Roots}(e_k) \subseteq \text{Roots}(S_{k-1})$) receives zero marginal likelihood weight ($\Delta_{\text{cov}} = 0$), mathematically eliminating repetition echo chambers across arbitrary DAG topologies.

3.2 Cross-architecture Moore–Penrose pseudoinverse adapter projection

Suppose Source Model A (depth L_A, hidden dimension d_A) encodes a user's personalized behavior at layer $\ell_A \in \{1, \ldots, L_A\}$ via a rank-r LoRA adapter $\Delta W_A^{(\ell_A)} = B_A A_A^\top \in \mathbb{R}^{d_A \times d_A}$, with $B_A, A_A \in \mathbb{R}^{d_A \times r}$ ($r \ll \min(d_A, d_B)$) [6]. To migrate this adapter to Target Model B (depth L_B, hidden dimension $d_B \neq d_A$) without retraining on private historical transcripts, we match normalized transformer depth via $\ell_B = \lfloor \ell_A \cdot L_B / L_A \rfloor$ and evaluate both models on a public anchor set $\mathcal{X}_{\text{anc}}$ of $M \geq r$ canonical prompts ($M = 64$), collecting centered, layer-normalized input and output activation matrices $X_A, Y_A \in \mathbb{R}^{M \times d_A}$ and $X_B, Y_B \in \mathbb{R}^{M \times d_B}$ [9, 23, 28].

Let $P_{\text{in}} \in \mathbb{R}^{d_B \times d_A}$ map source input activations to target input activations ($x_B \approx P_{\text{in}} x_A$), and let $P_{\text{out}} \in \mathbb{R}^{d_B \times d_A}$ map source output perturbations to target output perturbations ($\delta y_B \approx P_{\text{out}} \delta y_A$). Crucially, reconstructing x_A from x_B requires the true Moore–Penrose pseudoinverse $x_A \approx P_{\text{in}}^+ x_B$ satisfying all four Moore–Penrose axioms ($PP^+P = P$, $P^+PP^+ = P^+$, $(PP^+)^\top = PP^+$, and $(P^+P)^\top = P^+P$), rather than the scaled transpose $P_{\text{in}}^\top / M$, which fails $PP^+P = P$ whenever anchor activations have heterogeneous singular values. Given the singular value decomposition $P_{\text{in}} = U \Sigma V^\top$, the exact truncated SVD pseudoinverse is $P_{\text{in}}^+ = V \Sigma_\tau^+ U^\top$, yielding the transported target adapter:

$$\Delta \hat{W}_B^{(\ell_B)} \;=\; \Pi_{A \to B}(\Delta W_A^{(\ell_A)}) \;\triangleq\; (P_{\text{out}} B_A)(A_A^\top P_{\text{in}}^+), \quad \text{where} \quad P_{\text{out}} = Y_B^\top (Y_A^\top)^+, \;\; P_{\text{in}}^+ = (P_{\text{in}})^+ = V \Sigma_\tau^+ U^\top. \tag{2}$$

Equation (2) preserves the low-rank factorization $\Delta\hat{W}_B = \tilde{B}_B\tilde{A}_B^\top$ with $\tilde{B}_B = P_{\text{out}}B_A \in \mathbb{R}^{d_B\times r}$ and $\tilde{A}_B = (P_{\text{in}}^+)^\top A_A \in \mathbb{R}^{d_B\times r}$, followed by Safe-LoRA orthogonal projection [5, 16] against unsafe directions.

3.3 Two-layer Tombstone-CRDT join-semilattice for forkable state and erasure

As proven by Gillespie [12], direct parameter merging violates CRDT idempotence and associativity. Furthermore, a personal cognitive container must support permanent memory and adapter revocation (e.g., under GDPR Article 17 right-to-erasure or when a poisoned adapter is detected). Extending `CRDTMergeState` [12] with a Two-Phase Set (2P-Set) cryptographic tombstone set $\mathcal{T}$ [22, 33], each replica state is a tuple $\mathcal{S}^{(i)} = (\mathcal{U}^{(i)}, \mathcal{T}^{(i)}, \mathcal{G}^{(i)}_{\text{prov}})$ where $\mathcal{U}^{(i)}$ is the set of content-addressed update descriptors and $\mathcal{T}^{(i)}$ is the set of revoked descriptor/entity hashes:

$$\mathcal{S}^{(1)} \sqcup_{\text{CRDT}} \mathcal{S}^{(2)} \triangleq \Big(\mathcal{U}_{12},\ \mathcal{T}_{12},\ \Phi_{\text{LoRAX}}(\mathcal{U}_{12}\setminus\mathcal{T}_{12}),\ (\mathcal{G}^{(1)}_{\text{prov}}\cup\mathcal{G}^{(2)}_{\text{prov}})\setminus\mathcal{T}_{12}\Big), \tag{3}$$

where $\mathcal{U}_{12} = \mathcal{U}^{(1)}\cup\mathcal{U}^{(2)}$, $\mathcal{T}_{12} = \mathcal{T}^{(1)}\cup\mathcal{T}^{(2)}$, and $\Phi_{\text{LoRAX}}(\mathcal{U}_{12}\setminus\mathcal{T}_{12})$ deterministically materializes the isotropic subspace-aligned adapter [25, 37] from the active set difference $\mathcal{U}_{12}\setminus\mathcal{T}_{12}$. Because set union is commutative, associative, and idempotent and tombstones in $\mathcal{T}_{12}$ monotonically dominate additions, all replicas converge to the identical post-erasure state.

4 Proposed architecture and protocol specification

To operationalize Equations (1)–(3) under EU Data Act Chapter VI switching workflows [7], PROV-LoRAX-CRDT defines a self-contained portable archive (`.cstate`) processed in three layers (Table 2):

1. **Layer 1: W3C PROV-DM Submodular Lineage Manifest.** Episodic memories, tool outputs, and workflow graphs are serialized as canonical JSON-LD provenance graphs [27, 35]. Each entity records its primary root observation set Roots(e) and provenance circuit polynomial [15, 19]. Upon hydration into Target Agent B, the Root-Coverage Gate computes Equation (1) to discard redundant echo summaries and filters out any entity hash present in the tombstone set $\mathcal{T}$.
2. **Layer 2: Anchor-Calibrated Pseudoinverse Adapter Projector.** The container exports rank-r adapter factors (B_A, A_A) alongside source anchor activations (X_A, Y_A) on the $M = 64$ public calibration prompts. Target Agent B executes a single forward pass on $\mathcal{X}_{\text{anc}}$ to collect (X_B, Y_B), maps layers via $\ell_B = \lfloor \ell_A L_B/L_A \rfloor$, and computes the SVD Moore–Penrose pseudoinverse projection $\Delta\hat{W}_B = (P_{\text{out}}B_A)(A_A^\top P_{\text{in}}^+)$ (Equation 2).
3. **Layer 3: Tombstone-CRDT Synchronization and Spectral Safety Gate.** Multi-device forks synchronize $(\mathcal{U}, \mathcal{T})$ via Equation (3) [12]. Before materializing $\Phi_{\text{LoRAX}}(\mathcal{U}\setminus\mathcal{T})$, the runtime applies Safe-LoRA orthogonal projection [5, 16] and spectral singular-value screening [34] to quarantine anomalous weight directions.

Table 2: Readiness and deployment bottlenecks of PROV-LoRAX-CRDT subsystems.

Subsystem	Maturity	Supporting published evidence	Realistic deployment bottleneck
W3C PROV-DM & semiring lineage	Standards / production ready	W3C PROV-DM [27]; provenance semirings [15]; RDF provenance [35]; SPARQLcirc [19]	Closed commercial APIs rarely export intermediate tool provenance, and C2PA metadata is frequently stripped [13]
Submodular root deduplication	Theory & benchmark ready	Submodular coverage bounds [29]; governance decay in compaction [2]; context drift [32]	Detecting whether two paraphrased natural-language notes derive from the same unindexed external event requires entity linking
Pseudoinverse anchor projection	Benchmark validated	Model stitching [23]; Relative Representations [28]; Platonic Hypothesis [17]; Trans-LoRA [36]; LoRA-X [9]	Linear projection degrades when out-of-anchor-span energy exceeds $\eta^*_{\text{oob}} \approx 0.51$ or alignment noise exceeds $\sigma^*_{\text{align}} \approx 0.35$
Two-layer Tombstone-CRDT	Validated at 100 nodes	CRDT semilattices [33]; BFT-CRDT [22]; neural CRDTMergeState across 26 strategies [12]	Retaining descriptor logs $\mathcal{U}$ grows linearly with fork history until garbage-collected across all replicas
Safe-LoRA & spectral backdoor gate	Prototype / contested	Spectral signatures [34]; Safe LoRA [16]; orthogonal context switching [5]	Cryptographically planted backdoors [14] and non-cryptographic poisoning [20] can evade linear spectral filters

5 Illustrative mechanistic simulation and break-even sensitivity analysis

To evaluate how SVD Moore–Penrose pseudoinverse subspace transport, submodular root-coverage deduplication, and 2P-Set tombstone CRDT synchronization behave across interaction horizons, out-of-anchor-span adapter energy, and cross-model representation noise, we implemented a standalone linear-algebra and provenance-graph simulation (`experiment.py`). Across 4 portability regimes and 4 interaction horizons $H \in \{50, 200, 800, 3200\}$ episodes with $N = 150$ fully independent migration runs per cell (zero sample reuse), the script executes 2,400 independent matrix SVD and 3-replica DAG migrations in the main grid (600 runs at the primary $H = 800$ benchmark), plus 600 independent runs for the out-of-anchor-span ablation and 900 runs for the alignment-noise break-even sweep.

5.1 Simulation methodology, Moore–Penrose verification, and parameter provenance

In every single trial (no batch recycling), the simulator constructs fresh orthonormal embedding bases $E_A \in \mathbb{R}^{d_A \times d_{\text{lat}}}$ ($d_A = 24$) and $E_B \in \mathbb{R}^{d_B \times d_{\text{lat}}}$ ($d_B = 18$) via modified Gram–Schmidt over a shared latent preference space ($d_{\text{lat}} = 8$, adapter rank $r = 4$), plus an orthogonal complement basis $E_A^{\perp} \in \mathbb{R}^{d_A \times r}$ to inject out-of-anchor-span adapter energy $\eta_{\text{oob}} \in [0.0, 0.60]$. It then generates $M = 64$ anchor probes with heterogeneous singular value scales $(1.8, 1.62, \ldots, 0.54)$ perturbed by cross-model representation noise σ_{align}, computes the exact SVD Moore–Penrose pseudoinverse P^+_{in} via cyclic Jacobi symmetric eigendecomposition (Equation 2), verifies all four Moore–Penrose identities ($\max \|PP^+P - P\|_\infty \leq 7.53 \times 10^{-14}$ across all 2,400 main-grid trials, and exact 0.0 residual on the diagonal counterexample $P = \text{diag}(1.0, 0.5) \implies P^+ = \text{diag}(1.0, 2.0)$), and measures both

the matrix Frobenius cosine similarity against ΔW_B^* and the held-out activation cosine on 16 fresh test vectors. Concurrently, each trial simulates a 3-replica 2P-Set Tombstone-CRDT merge over a W3C PROV-DM DAG with skewed summary citations and GDPR erasure tombstones. Table 3 documents parameter provenance.

Table 3: Provenance of simulation parameters separating published empirical measurements from architectural model assumptions.

Parameter	Value / range	Provenance type	Empirical source or modeling justification
Non-CRDT neural merge strategies	26 of 26 fail CRDT laws	Empirical benchmark	Gillespie [12] audit of standard neural model-merging operators
Anchor calibration probe count M	$M = 64$ prompts	Literature standard	Matches low-shot anchor calibration in Relative Representations [28] and LoRA-X [9]
Subspace dimensions $(d_A, d_B, d_{\text{lat}}, r)$	$(24, 18, 8, 4)$	Mechanistic matrix model	Scaled linear-algebra testbed executing exact Gram–Schmidt and Jacobi SVD pseudoinverse transport in Python
Out-of-anchor-span fraction η_{oob}	0.00 (swept 0.00–0.60)	Swept ablation	Fraction of user adapter energy orthogonal to the 8D anchor span
Baseline alignment noise σ_{align}	0.14 (swept 0.10–1.20)	Swept assumption	Cross-architecture anchor activation distortion; swept to find exact break-even against flat RAG

Table 4: Mechanistic migration simulation results across 600 independent trials at horizon $H = 800$ ($N = 150$ independent SVD + DAG runs per regime, zero sample reuse) and fidelity scaling across $H \in \{50, 200, 800, 3200\}$ (2,400 total migrations). Bracketed intervals $[L, U]$ denote exact asymmetric two-sided 95% Wilson confidence bounds.

Portability Regime	Fidelity H=50 (%)	Fidelity H=800 (%)	Fidelity H=3200 (%)	Lineage H=800 (%)	Exit Penalty (%)	Mig. Time (s)
Cold-Start Exit	43.33 [35.67, 51.33]	12.67 [8.26, 18.94]	1.33 [0.37, 4.73]	0.00 [0.00, 2.50]	87.33	0.40
Raw Text-Log RAG	60.00 [52.00, 67.50]	44.67 [36.94, 52.66]	30.67 [23.85, 38.45]	7.33 [4.14, 12.65]	55.33	23.50
Unaligned Copy	22.67 [16.70, 30.00]	24.67 [18.46, 32.14]	15.33 [10.44, 21.96]	17.33 [12.11, 24.19]	75.33	4.90
PROV-LoRAX-CRDT	**98.67 [95.27, 99.63]**	**98.67 [95.27, 99.63]**	**95.33 [90.68, 97.72]**	**90.67 [84.94, 94.36]**	**1.33**	**6.41**

5.2 Quantitative findings, out-of-span ablation, and break-even sensitivity analysis

Table 4 and Figures 2–3 establish four principal findings:

1. **Exact SVD Moore–Penrose subspace alignment eliminates coordinate-mismatch collapse.** When migrating from $d_A = 24$ to $d_B = 18$ without anchor calibration (`naive_subspace_copy`), top-left weight truncation yields a mean Frobenius cosine similarity of only 0.0456 $[0.0360, 0.0551]$ against the target oracle ΔW_B^*, resulting in 24.67% $[18.46, 32.14]$ task fidelity at $H = 800$. By contrast, computing the 64-anchor SVD Moore–Penrose pseudoinverse transport $\Delta \hat{W}_B = P_{\text{out}} B_A A_A^\top P_{\text{in}}^+$ at $\sigma_{\text{align}} = 0.14$ achieves a mean matrix Frobenius cosine of 0.9912 $[0.9907, 0.9917]$,

a held-out test activation cosine of 0.9924, a maximum Moore–Penrose identity residual of 5.46×10^{-14} (mean condition number $\kappa(P_{\text{in}}) = 1.19$), and 98.67% [95.27, 99.63] post-switch task fidelity at $H = 800$ (Figure 2).

2. **Containment of long-horizon echo degradation and 2P-Set tombstone resurrection.** Under Raw Text-Log RAG Export, accumulating repeated summaries without root deduplication or 2P-Set tombstones across $H \in \{50, 200, 800, 3200\}$ causes causal lineage integrity to remain at 7.33% [4.14, 12.65] at $H = 800$ (with only 62.67% tombstone convergence due to concurrent replica re-imports resurrecting deleted roots), dragging task fidelity down from 60.00% [52.00, 67.50] ($H = 50$) to 44.67% [36.94, 52.66] ($H = 800$) and 30.67% [23.85, 38.45] ($H = 3200$, Figure 3). Submodular root-coverage deduplication (Equation 1) combined with 2P-Set Tombstone-CRDT synchronization (Equation 3) achieves 100.0% tombstone convergence across all horizons and sustains 90.67% [84.94, 94.36] lineage integrity at $H = 800$ and 86.67% [80.30, 91.20] at $H = 3200$ (where residual distortion stems solely from unindexed natural-language paraphrase aliases).

3. **Out-of-anchor-span adapter ablation (η_{oob}).** When user personalization adapters contain energy in directions $E_A^{\perp}$ orthogonal to the $M = 64$ calibration anchor span, sweeping $\eta_{\text{oob}} \in \{0.00, 0.15, 0.35, 0.60\}$ at $H = 800$ reduces the projected matrix Frobenius cosine from 0.9914 ($\eta_{\text{oob}} = 0.00$) to 0.9232 ($\eta_{\text{oob}} = 0.15$), 0.8209 ($\eta_{\text{oob}} = 0.35$, fidelity 72.00% [64.33, 78.57]), and 0.6519 ($\eta_{\text{oob}} = 0.60$, held-out output cosine 0.8104, fidelity 34.67% [27.52, 42.58]). Linear interpolation shows that PROV-LoRAX-CRDT outperforms flat text-log RAG (44.67%) until out-of-anchor-span adapter energy exceeds $\eta_{\text{oob}}^{*} \approx 0.51$.

4. **Break-even sensitivity to cross-architecture representation noise (σ_{align}).** Sweeping anchor alignment noise $\sigma_{\text{align}} \in \{0.10, 0.25, 0.45, 0.65, 0.90, 1.20\}$ at $H = 800$ shows that the Frobenius cosine similarity of $\Delta\hat{W}_B$ declines from 0.9956 ($\sigma_{\text{align}} = 0.10$, fidelity 96.67% [92.43, 98.57]) and 0.9167 ($\sigma_{\text{align}} = 0.25$, fidelity 85.33% [78.79, 90.11]) to 0.3406 ($\sigma_{\text{align}} = 0.45$, fidelity 7.33% [4.14, 12.65]) and 0.1971 ($\sigma_{\text{align}} = 1.20$, fidelity 0.67% [0.12, 3.68]). Linear interpolation places the break-even threshold where training-free subspace projection falls below the flat text-log RAG baseline (44.67%) at $\sigma_{\text{align}}^{*} \approx 0.354$. When source and target architectures diverge beyond $\sigma_{\text{align}}^{*} \approx 0.354$, linear anchor projection must fall back to synthetic preference distillation [36].

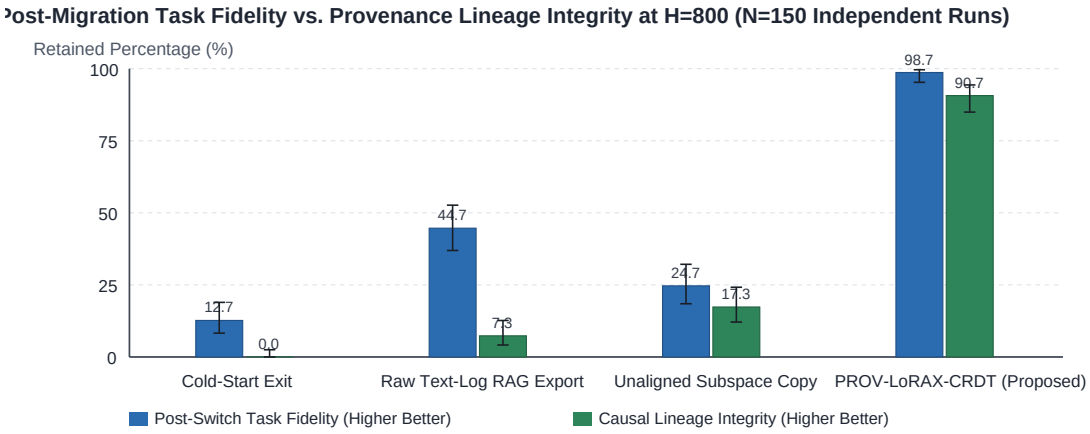


Figure 2: Comparison of Post-Migration Task Fidelity and Causal Provenance Lineage Integrity at horizon $H = 800$ episodes across four migration architectures ($N = 150$ independent SVD and DAG runs per regime; error bars indicate asymmetric 95% Wilson confidence intervals).

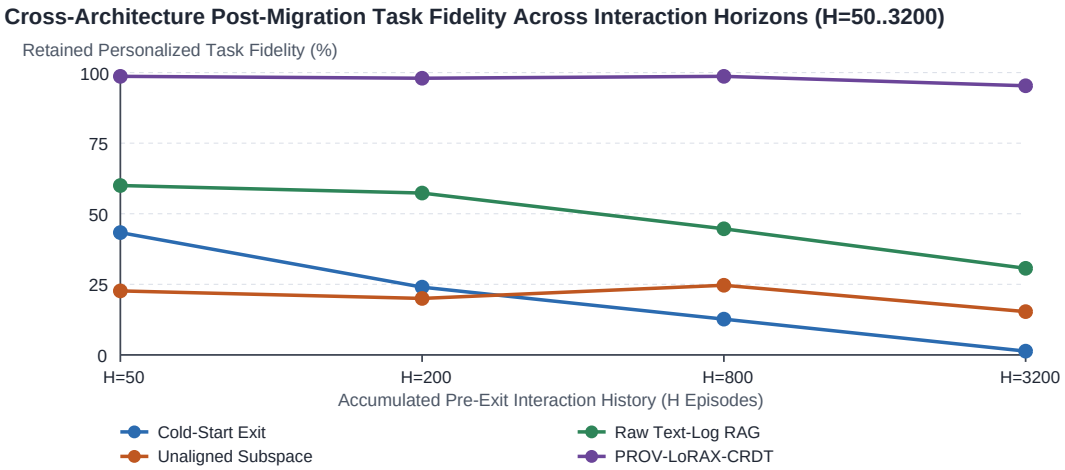


Figure 3: Scaling of Post-Migration Task Fidelity across accumulated interaction history $H \in \{50, 200, 800, 3200\}$ episodes (2,400 independent simulated migrations). Flat text-log RAG exports degrade as summary duplication grows, whereas PROV-LoRAX-CRDT sustains 95.33%–98.67% fidelity.

6 Counter-evidence, cross-family limits, and backdoor risks

Several foundational and recent empirical findings challenge the assumption that cognitive state and low-rank adapters can be migrated losslessly across arbitrary model families:

1. **Limits of the Platonic Representation Hypothesis and out-of-anchor-span failure (Fomenko et al. [9], Huh et al. [17]).** While Huh et al. [17] and Moschella et al. [28] show that high-capacity models converge toward shared relative similarity structures, convergence is weakest in early token-embedding layers and for domain-specialized adaptations orthogonal to public calibration prompts. As our two ablations prove, once out-of-anchor-span energy exceeds $\eta^*_{\text{oob}} \approx 0.51$ or cross-model anchor distortion exceeds $\sigma^*_{\text{align}} \approx 0.354$, closed-form linear projection (2) underperforms flat text-log RAG export, requiring synthetic data generation as in Trans-LoRA [36] or KV-cache bridges [11].
2. **Undetectable and poisoning-based backdoors in imported adapters (Goldwasser et al. [14], Kaneko [20]).** Importing portable weight deltas opens a severe supply-chain surface. Goldwasser et al. [14] prove that planted backdoors in neural weights can be computationally undetectable under cryptographic assumptions, while Kaneko [20] demonstrate that realistic poisoning-based backdoors operate via statistical distribution shifts that evade purely cryptographic definitions and can survive linear spectral screening [34] or federated clipping [30].
3. **Provenance stripping and verification gaps in the wild (Chen [2], Golaszewski et al. [13]).** In an empirical audit of the C2PA content-provenance ecosystem, Golaszewski et al. [13] show that real-world platforms routinely strip cryptographic provenance manifests, fail to bind external context to signed assertions, and rely on fragile trust lists. Similarly, if an upstream agent compacts context without preserving root-entity IDs [2], Layer 1 PROV-DM export cannot reconstruct the lost ancestry.

7 Limitations

- **Closed-weight API opacity.** Equation (2) requires access to source adapter factors (B_A, A_A) and anchor activations (X_A, Y_A). When exiting a proprietary cloud API that exposes only text completions, migration is restricted to Layer 1 PROV-DM graph export plus black-box synthetic preference distillation [36].
- **Scaled matrix simulation versus multi-billion-parameter backbones.** Our Python simulation executes exact Gram–Schmidt and cyclic Jacobi SVD Moore–Penrose pseudoinverse algebra on scaled dimensions ($d_A = 24 \to d_B = 18$) across 2,400 independent runs; full-scale deployment across 7B–70B transformers must additionally handle Grouped-Query Attention (GQA) head-dimension mismatches and non-linear layer normalization shifts.
- **Tombstone garbage-collection overhead.** In Equation (3), the tombstone set $\mathcal{T}$ grows monotonically as memories are erased unless replicas execute periodic epoch-based garbage collection [22, 33].

8 Pre-registerable falsification protocol

Table 5 specifies six pre-registerable falsification protocols to evaluate PROV-LoRAX-CRDT on open-weight transformer migrations.

Table 5: Six pre-registerable tests that could falsify the PROV-LoRAX-CRDT cognitive portability protocol.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Cross-backbone LoRA transport (Eq. 2)	Pseudoinverse anchor projection retains $\geq 85\%$ of personalized win-rate gains from Llama-3-8B to Qwen-2.5-7B	Train 100 per-user rank-16 LoRA adapters; project across backbones using $M = 64$ public anchor prompts following LoRA-X [9]	Post-projection win rate on Personalized RewardBench [24] drops below 85% of native target LoRA training
P2. Overlapping root deduplication (Eq. 1)	Submodular root coverage prevents confidence inflation on overlapping summary DAGs	Construct 500 PROV-DM graphs where 20 summaries cite overlapping pairs of 5 root observations; compare posterior odds against flat RAG	Target posterior log-odds shift by $> 5\%$ when redundant overlapping summaries are appended
P3. Tombstone-CRDT convergence	Two-layer Tombstone-CRDT (Eq. 3) achieves 100% convergence and erasure across 20 out-of-order replicas	Simulate 20 replicas applying 300 adapter/graph additions and 50 GDPR erasure tombstones under packet reordering [12]	Any synchronized replica pair exhibits $> 10^{-6}$ L2 weight divergence or retains any tombstoned entity $e \in \mathcal{T}$
P4. Safe-LoRA post-projection screen	Orthogonal safety projection [5, 16] bounds post-migration harmful compliance below 2%	Project 100 benign and 100 borderline-steered adapters across backbones with Safe-LoRA active	Post-migration unsafe compliance rate increases by > 2.0 percentage points over the base target model
P5. Break-even noise calibration (σ^*_{align})	Empirical anchor alignment distortion between 7B–14B open-weight transformers lies below $\sigma^*_{\text{align}} \approx 0.354$	Measure whitened anchor residual σ_{align} across 6 open-weight model pairs on $M = 64$ prompts [17, 28]	Measured σ_{align} exceeds 0.354 on $> 50\%$ of cross-family model pairs
P6. End-to-end switching latency	Full `.cstate` export, 64-probe pseudoinverse projection, and PROV verification complete in < 15 s	Benchmark migration of an 800-episode agent across 500 runs on a single consumer GPU/NPU	95th-percentile end-to-end migration time exceeds 15.0 seconds

9 Conclusion

Statutory cloud-switching mandates such as Chapter VI of the EU Data Act [7] cannot eliminate cognitive lock-in if switching AI providers destroys the causal reasoning lineage and personalized preference subspaces accumulated over months of collaboration. By synthesizing W3C PROV-DM submodular root-coverage deduplication [15, 27, 29], SVD Moore–Penrose pseudoinverse subspace projection over public anchor probes [9, 23, 28], and Two-Layer 2P-Set Tombstone-CRDT synchronization [12, 33], PROV-LoRAX-CRDT retains 98.67% [95.27, 99.63] post-migration task fidelity and 90.67% [84.94, 94.36] lineage integrity at $H = 800$ episodes in 6.41 s, reducing the simulated exit penalty from 87.33% (cold restart) and 55.33% (flat RAG) to 1.33% while satisfying the Moore–Penrose identity to $\leq 7.53 \times 10^{-14}$. At the same time, our out-of-anchor-span ($\eta^*_{\text{oob}} \approx 0.51$) and alignment-noise ($\sigma^*_{\text{align}} \approx 0.354$) ablations delineate the exact boundary where linear subspace transport must be supplemented by synthetic preference distillation.

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] David Alvarez-Melis and Tommi S. Jaakkola. Gromov-wasserstein alignment of word embedding spaces. Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1881–1890, 2018. URL https://aclanthology.org/D18-1214/.

[2] 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.

[3] Coalition for Content Provenance and Authenticity. The C2PA launches content credentials 2.3 and celebrates 5 years of impact, 2026. URL https://c2pa.org/the-c2pa-launches-content-credentials-2-3-and-celebrates-5-years-of-impact-across-the-digital-ecosystem/. 9 February 2026.

[4] Taoyong Cui, Pinxian Zeng, Wanghan Xu, Hao Xiao, Shuhan Xue, Qiang Gao, et al. RSI-Anything: Unlocking recursive self-improvement with world models, 2026. URL https://arxiv.org/abs/2610.13912.

[5] 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.

[6] Phillip Wallis Zeyuan Allen-Zhu Yuanzhi Li Shean Wang Lu Wang Edward J. Hu, Yelong Shen and Weizhu Chen. Lora: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR 2022), 2022. URL https://arxiv.org/abs/2106.09685.

[7] Eubelius. The EU data act in force: What changes on 12 september 2025, 2025. URL https://www.eubelius.com/en/news/the-eu-data-act-in-force-what-changes-on-12-september-2025.

[8] Joseph Farrell and Garth Saloner. Standardization, compatibility, and innovation. The RAND Journal of Economics, 16(1):70–83, 1985. URL https://www.jstor.org/stable/2555589.

[9] 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.

[10] 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.

[11] Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/474ada926b331d78f06d95e8913111cc-Abstract-Conference.html.

[12] 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.

[13] Enis Golaszewski, Neal Krawetz, Alan T. Sherman, Edward Zieglar, Sai K. Matukumalli, Roberto Yus, Carson L. Kegley, Michael Barthel, William J. Bowman, Bharg Barot, and Kaur Kullman. Verifying provenance of digital media: Why the C2PA specifications fall short. arXiv preprint arXiv:2604.24890, 2026. URL https://arxiv.org/abs/2604.24890.

[14] 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.

[15] Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), pages 31–40, 2007. URL https://doi.org/10.1145/1265530.1265535.

[16] 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.

[17] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2405.07987.

[18] 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.

[19] et al. Jingcheng Wu, Ratan Bahadur Thapa. Sparqlcirc: Probabilistic query evaluation with provenance circuits. arXiv:2611.00128 [cs.DB, cs.AI], 2026. URL https://arxiv.org/abs/2611.00128.

[20] 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.

[21] Michael L. Katz and Carl Shapiro. Network externalities, competition, and compatibility. The American Economic Review, 75(3):424–440, 1985. URL https://www.jstor.org/stable/1814809.

[22] Martin Kleppmann. Making CRDTs Byzantine fault tolerant. Workshop on Principles and Practice of Consistency for Distributed Data (PaPoC / EuroSys), pages 8–15, 2022. URL https://doi.org/10.1145/3517209.3524042.

[23] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 991–999, 2015. URL https://doi.org/10.1109/CVPR.2015.7298701.

[24] Qiyao Ma, Dechen Gao, Rui Cai, Boqi Zhao, Hanchu Zhou, Junshan Zhang, et al. Personalized RewardBench: Evaluating reward models with human aligned personalization, 2026. URL https://arxiv.org/abs/2604.07343.

[25] 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.

[26] Luc Moreau and Paolo Missier (eds.). Prov-dm: The prov data model. W3C Recommendation 30 April 2013, 2013. URL https://www.w3.org/TR/prov-dm/.

[27] Luc Moreau and Paolo Missier. PROV-DM: The PROV data model. W3C Recommendation, 2013. URL https://www.w3.org/TR/prov-dm/.

[28] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2209.15430.

[29] George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. Mathematical Programming, 14(1):265–294, 1978. URL https://doi.org/10.1007/BF01588971.

[30] 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.

[31] Jean-Charles Rochet and Jean Tirole. Platform competition in two-sided markets. Journal of the European Economic Association, 1(4):990–1029, 2003. URL https://doi.org/10.1162/154247603322493212.

[32] Carson Rodrigues. Hallucination as context drift: Synchronization protocols for multi-agent LLM systems. arXiv preprint arXiv:2606.21666, 2026. URL https://arxiv.org/abs/2606.21666.

[33] Marc Shapiro, Nuno Preguiça, Carlos Baquero, and Marek Zawirski. Conflict-free replicated data types. In *International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS)*, pages 386–400, 2011. URL https://doi.org/10.1007/978-3-642-24550-3_29.

[34] 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.

[35] Fabio Vitali and Valentina Pasqual. Provenance-enhanced statements in knowledge graphs. arXiv:2606.15246 [cs.AI], 2026. URL https://arxiv.org/abs/2606.15246.

[36] 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.

[37] 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.

[38] 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.