REVBOUND: REVOCATION CUTS FOR TRANSFORMED AGENT MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

An agent reads an authorized memory fragment, compresses it into a plan, and caches that plan for a peer. Access is then revoked. The source disappears from retrieval, yet its derivative can still enter the next model call. RevBound treats this lifecycle as a path-coverage problem over execution provenance: a *revocation cut* must intersect every denied-source path after its last mutable transformation. The Object-Alignment Principle identifies when revocation evaluations protect the same output under the same relational source closure, policy version, and decision boundary. RevBound realizes such a cut at model-call admission with typed user–agent and agent–resource lineage, a linearizable policy fence, transitive parent closure, and an `AllowedBundle`-only invocation path. In 1,160 matched fixed-parent attacks spanning re-summarization, paraphrase, merging, and relay, retrieval-time filtering exposes 600 forbidden derivatives (51.7%); two independently implemented final-boundary serializers expose 56 (4.8%) and 36 (3.1%), while RevBound exposes none (one-sided 95% upper bound $\leq 0.26\%$). RevBound also records zero violations over 156,000 churned bundle reads and 1.8M fragment checks at 7.6 ms p95. On SWE-Coop it resolves $28.4\pm1.1\%$ of issues, compared with $23.6\pm1.0\%$ for shared vector memory, and a second model/tool stack reproduces the boundary-class gap, with trace-audited leaks falling from 55/300 to 1/300. Revocation for transforming agent memory must therefore be enforced on the model-admitted object, where complete lineage and non-bypassable routing turn source denial into descendant exclusion.

1 INTRODUCTION

A code-review agent reads an authorized repository fragment at t_0, summarizes it into a plan, and relays the plan to a collaborator. At t_r, repository access is revoked. A retrieval filter now does exactly what its contract requires: it removes the original fragment. The cached plan nevertheless remains eligible for prompt assembly at t_e, so the next model call consumes a descendant of the revoked source. No local access-control decision is incorrect; the system protected an intermediate object while a transformed object crossed the model boundary.

This read–transform–revoke–consume sequence recurs in long-horizon agents because observations, reflections, tool traces, plans, and messages persist across model calls and agents (Park et al., 2023; Schick et al., 2023; Yao et al., 2023; Shinn et al., 2023; Wang et al., 2024; Packer et al., 2023; Li, 2026; Li et al., 2023; Rezazadeh et al., 2025; Yu et al., 2026). It changes the revocation question from "is the current retrieval set allowed?" to "does any denied source still have a dependency path into the object admitted to the model?"

We answer with a path property. For policy version v and model-admitted bundle B, a *revocation cut* intersects every dependency path from a source denied under v after that path's last mutable transformation. The cut checks the path's relational source closure and remains current until model-call admission. A constructive non-equivalence result shows how any earlier decision can be bypassed by cached, history-dependent state. A sufficiency result connects the cut to the implemented contract: authenticated principals, lineage-sound tracked transformations, a linearizable policy fence, and a single guarded model-invocation route.

RevBound realizes this cut by preserving exact user–agent and originating agent–resource relations through every tracked transformation. Immediately before admitting a model call, it recomputes their

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

transitive union, validates a fenced policy lease, and passes only a typed `AllowedBundle` to the inference transport. A certificate binds every component to the policy version, closure digest, and deterministic decision. Figure 1 traces the same fragment from authorized read to the final cut.

The central experiment replays 1,160 identical fragments, rewrites, policy updates, prompts, and seeds across four enforcement designs. Retrieval-time filtering records 600 final-bundle leaks (51.7%). Two separately implemented serializers that move the decision after transformation, using fragment ancestry and rewritten-span labels, record 56 (4.8%) and 36 (3.1%); RevBound records 0 (upper bound $\leq 0.26\%$). The ordering holds for re-summarization, paraphrase, partial-copy merging, and cross-agent relay. Component ablations then identify closure propagation and final gating as separate requirements, while 156,000 churned reads test the policy fence at up to 120 revocations/min.

The *Object-Alignment Principle* makes the experimental consequence explicit. It describes an evaluation by $\sigma(\mathcal{E}) = (O, \Pi, v, \partial)$: protected output, relational provenance closure, effective policy version, and decision boundary relative to the last mutable transformation. Systems provide close evidence about the same revocation property only when these coordinates align, even when their internal data structures differ. This criterion separates final-output revocation from write-time authority, retrieval filtering, tool-call information-flow enforcement, and parameter-level unlearning, which protect different objects or intervals.

The remaining experiments test whether the cut survives pressure beyond the fixed-parent intervention. Held-out adaptive transformations expose attribution misses; fixed-retrieval RepoHoldout controls memory contents and ranking; SWE-Coop measures end-to-end collaborative utility; and a Llama-3.3/τ-bench (Yao et al., 2025) stack tests transfer. In the latter, trace-audited leaks fall from 55/300 for earlier-object filtering to 1/300 for RevBound, with task success increasing in both domains.

Our contributions are:

1. **Revocation cuts and object alignment.** A path-based definition, non-equivalence construction, and four-coordinate comparison rule identify the object and evidence required for model-input revocation.
2. **A lineage-sound model-call boundary.** RevBound combines exact relational lineage, tracked transformation closure, admission-time policy fencing, and `AllowedBundle`-only invocation.
3. **Mechanism-resolving evidence.** Matched attacks and two independent final-boundary serializers establish a boundary-class separation; component ablations expose the additional roles of closure and mandatory routing.
4. **Operational evidence across regimes.** Churn, held-out adaptation, fixed retrieval, collaborative coding, and a second model/tool stack quantify covered-source exclusion, cost, utility compatibility, and attribution pressure.

2 Why Earlier Security Objects Miss Transformed State

Agent-memory systems determine what state is stored, shared, and retrieved (Packer et al., 2023; Li, 2026; Rezazadeh et al., 2025; Yu et al., 2026). Collaborative Memory is the closest access-control setting: it attaches immutable provenance and rechecks dynamic user–agent–resource permissions when producing filtered memory views (Rezazadeh et al., 2025). RevBound targets the later object produced after those views are summarized, merged, templated, or relayed, and retains the originating agent–resource pairing needed to re-evaluate each derivative.

Recent memory-security mechanisms make provenance operational but answer neighboring questions. TMA-NM binds non-malleable authority to write-time origin, PPMF prevents source authority from increasing before high-risk tool calls, and MemLineage gates sensitive actions using a provenance chain (Louck, 2026; Xu et al., 2026; Ouyang & Hou, 2026). Their principal policy is trust or authority under poisoning; ours is requester-specific access revocation under a changing authorization graph, checked on model-visible context. Agentic unlearning removes targeted information from parameters and memory (Wang et al., 2026), whereas RevBound neither modifies model weights nor claims erasure: it excludes policy-ineligible dependencies from a particular model-call bundle.

CaMeL separates trusted control flow from untrusted data and enforces capabilities at tool invocation, while FIDES propagates confidentiality and integrity labels through planner and tool execution (Debenedetti et al., 2026; Costa et al., 2025). These systems motivate non-bypassable mediation and relational labels; RevBound contributes the dynamic retrospective case in which content was legitimately consumed, transformed, and cached before its source authorization changed. In the broader security literature, FLAM joins authorization and information flow under mutable trust, and Dynamic Release formalizes policy upgrades and revocation (Arden et al., 2015; Li & Zhang, 2022). Revocation cuts specialize those concerns to provenance-carrying agent transformations and identify the concrete output whose exclusion is measured.

Complete mediation, dynamic policy semantics, and provenance propagation supply three necessary ingredients, but they do not by themselves identify which use must be mediated after an agent materializes new state. Revocation cuts connect these ingredients on an execution graph: dynamic authorization labels the denied sources, provenance identifies their descendant paths, and the cut places mediation after the last mutable transformation at model-call admission. Object alignment turns that composition into an evaluation rule, so a baseline is close only when it measures the same consumed object under equivalent closure and policy timing.

This distinction yields four experimental classes. Write-time admission decides before storage; provenance logging records history; retrieval filtering decides current candidates; model-call admission decides the bytes exposed after prompt transformations. Only the last class shares RevBound's policy-decision boundary. Within that class, closure representation and routing completeness still matter, so we call SB-1 and SB-2 *boundary-matched comparators*, not complete cuts. Their residual failures are evidence about those implementation dimensions rather than counterexamples to the cut definition.

3 Revocation Cuts and Object Alignment

3.1 Actors and permissions

We use U for users, A for agent identities, and R for tools/resources. Permissions evolve over time:

- $G_{UA}(t) \subseteq U \times A$: which agents a user may invoke at time t.
- $G_{AR}(t) \subseteq A \times R$: which resources an agent may access at time t.

Policies change due to role updates, token expiry, project transitions, or incident response. Each committed policy version v defines $G_{UA}(v)$ and $G_{AR}(v)$. We assume tenant-scoped graphs and authenticated long-lived agent principals; provenance carries the tenant identifier, and cross-tenant closure is rejected before eligibility is evaluated.

3.2 Memory fragments

Each immutable fragment m contains content, a type (Fact, ToolTrace, Decision, Plan, Warning, Resolution), typed provenance, trust metadata, and parent links. For requester u, its transitive authorization closure is

$$\Pi^*(m; u) = \big(E^*_{UA}(m; u), E^*_{AR}(m)\big),$$

where $E^*_{UA}(m; u) = \{(u, a') : a' \text{ contributed to } m\}$ and $E^*_{AR}(m)$ preserves every exact originating pair (a', r) through which a contributing agent accessed a resource. Keeping (a', r) as a relation, rather than independent agent and resource sets, prevents a resource obtained by a' from being incorrectly checked against the requesting agent a. Derived fragments materialize the union of their parents' typed edges; missing or inconsistent parents are rejected or confined to a raw-only tier with no prompt route.

Recorded lineage versus content dependence. Let $D^*(m; u) = (D^*_{UA}(m; u), D^*_{AR}(m))$ denote the typed authorization edges of sources on actual content-dependence paths to m. Recorded lineage is *sound* when

$$D^*_{UA}(m; u) \subseteq E^*_{UA}(m; u) \quad \text{and} \quad D^*_{AR}(m) \subseteq E^*_{AR}(m).$$

Over-approximation may deny additional content but preserves exclusion; omitted dependencies can escape any lineage-based decision. RevBound establishes soundness for the covered transformation

API by deriving parent handles from every model-visible input to a rewrite, merge, template, or relay call inside the trusted wrapper, rather than accepting a writer-selected parent list.

3.3 THREAT MODEL

For shared agent memory with dynamic permissions, we target *content-level* retroactive leakage. The adversary may be a malicious user in the same tenant, a compromised agent that chooses adversarial outputs, or an untrusted resource whose output is summarized and relayed. It may retain earlier text, invoke allowed rewrite operations, and adapt after coarse allow/deny outcomes. Authenticated principals, the tracked transformation wrapper, linearizable policy fences, and mandatory `AllowedBundle` invocation form the trusted enforcement base. A compromised writer controls content but cannot delete wrapper-attached input lineage. Opaque or uninstrumented derivations enter a raw-only tier, while the integrity composition in Table 4 evaluates tampering and replay by a compromised store.

3.4 EXECUTION-PROVENANCE GRAPH AND REVOCATION CUTS

Let $H = (V, E)$ record one agent execution: source and memory nodes feed rewrite, merge, template, and relay nodes, which feed token nodes in the bundle B admitted to a model call. An edge denotes actual content dependence; $D^*(z; u)$ follows those edges, while $\Pi^*(z; u)$ is the recorded closure available to enforcement. For version v, a decision node is v-*valid* for a path when it checks a sound closure under an admission-valid fence for v after the path's last adversary-controlled transformation.

A set of v-valid decision nodes C_v is a *revocation cut* for B when every path from a source denied under v to any token in B intersects C_v. Cut completeness is therefore a path-coverage property, not a property of an isolated retrieval call. The guarded model-invocation boundary supplies a natural cut because every covered dependency crosses it after prompt transformation and immediately before model-call transport.

3.5 THE OBJECT-ALIGNMENT PRINCIPLE

For evaluation $\mathcal{E}$, define the signature $\sigma(\mathcal{E}) = (O, \Pi, v, \partial)$, where O is the object admitted to the model call, Π its relational historical-closure rule, v the policy version fenced at admission, and ∂ the policy-decision boundary relative to the last mutable transformation. We write

$$\mathcal{E}_1 \simeq_{\mathrm{OA}} \mathcal{E}_2 \quad \Longleftrightarrow \quad O_1 \equiv_{\mathrm{vis}} O_2 \ \wedge\ \Pi_1 \equiv_{\mathrm{src}} \Pi_2 \ \wedge\ v_1 \equiv_{\mathrm{pol}} v_2 \ \wedge\ \partial_1 \equiv_{\mathrm{cut}} \partial_2.$$

The four relations require the same model-visible measurement unit, the same typed ancestor-edge semantics, equivalent policy histories, and cut-equivalent last decisions with no unchecked mutable transformation afterward. Distinct implementations may satisfy the relation. SB-1 and SB-2 match ∂, isolating the policy-decision boundary coordinate while leaving closure representation and route completeness to the component analysis.

Proposition 1 (Post-decision boundary non-equivalence)**.** *Let x depend on source s, let an authorization edge carried by s be denied at version v, and let a history-dependent transformation T emit $y = T(x)$ after the last decision on x. If the decision set fails to cut the path $s \rightsquigarrow x \rightsquigarrow y \rightsquigarrow B$, compliance of the currently retrieved object does not imply compliance of final bundle B. Equivalently, filtering before T and closure-aware filtering after T are not interchangeable.*

Proof. Retrieve x while authorized, materialize y, deny an authorization edge carried by s, and filter the current candidate set. The filter excludes x, but T's cached output remains on an uncut path to B. A final closure-aware decision rejects y because its typed closure contains the denied edge. The two placements produce different outputs under the same policy version. Appendix A gives the path argument. □

Proposition 2 (Sound final-cut sufficiency)**.** *Under A1–A4, if every component m of $B(u, a, v)$ crosses RevBound's model-call admission decision and passes the eligibility predicate on a sound $\Pi^*(m; u)$, then*

$$D^*_{UA}(m; u) \subseteq E^*_{UA}(m; u) \subseteq G_{UA}(v), \qquad D^*_{AR}(m) \subseteq E^*_{AR}(m) \subseteq G_{AR}(v)$$

for every admitted component.

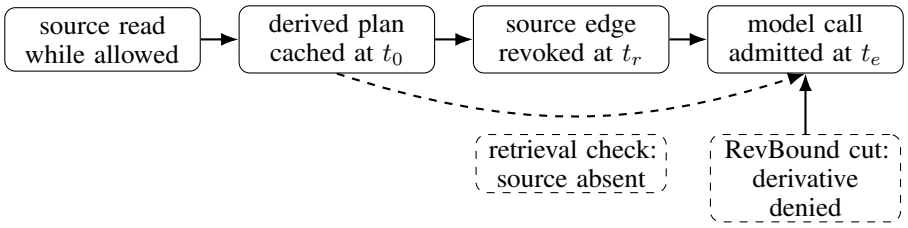


Figure 1: One source across the full revocation lifecycle. The retrieval check correctly observes that the source is absent after t_r, while the cached derivative retains an unchecked dashed path. RevBound cuts that path at model-call admission by evaluating the derivative's typed closure.

Proof. Algorithm 1 tests every recorded typed edge under one admission-valid policy version. Soundness places every actual dependency edge inside that tested closure; a denied pair removes m, and guarded invocation prevents unchecked components from reaching model transport. Thus the admission decisions intersect every covered source path to B. Appendix A states the graph argument. □

4 REVBOUND: A RELATIONAL MODEL-CALL GATE

RevBound makes model-call admission the last policy-sensitive operation before inference transport. Immutable typed edges determine which authorization relations reach each component; an admission-time fence fixes the effective policy version; `AllowedBundle` is the only model-consumable representation; and a decision certificate records the version, closure digests, and reason codes. These mechanisms bind the authorization decision to the bytes admitted to the model call.

4.1 ELIGIBILITY PREDICATE

For a request by user u through agent a, let $\Pi^*(m; u)$ be the typed closure defined above. Under admission version v, m is eligible exactly when

$$(u, a) \in G_{UA}(v) \quad \wedge \quad E^*_{UA}(m; u) \subseteq G_{UA}(v) \quad \wedge \quad E^*_{AR}(m) \subseteq G_{AR}(v).$$

The first term authenticates the live requesting edge. The second checks the requester's current access to every contributing agent. The third checks each resource against the agent that actually accessed it; for example, a revoked (a', r) cannot be hidden by an allowed (a, r). RevBound applies this predicate to every assembled component after rewrite and merge. Ingestion rejects missing parents, while serialization independently rejects malformed or inconsistent relational closure with a deterministic reason code.

4.2 POLICY LINEARIZATION, FENCING, AND FAST CHECKS

Policy versions are totally ordered by the authoritative engine. An update linearizes when that engine commits version v_r; its acknowledgement carries a tenant fence that invalidates lower-version admission leases. The guarded invocation path obtains an authenticated lease for v after the last mutable transformation and revalidates it immediately before writing the `AllowedBundle` to inference transport. If a revocation linearizes between those steps, revalidation restarts the decision at $v' \geq v_r$. Successful revalidation and transport admission form one model-call linearization point; a revocation committed afterward governs the next call, and a re-grant is a distinct higher-version update.

Replicas and bitmap caches may accelerate membership tests, but a row is accepted only when its tenant and version equal the fenced snapshot. A replica that cannot prove v meets the current fence, an expired lease, or a partition at revalidation fails closed. This is the mechanism exercised by the lag and outage rows of Table 6; the injected stale-bitmap bug deliberately bypasses the equality check and therefore exposes the bounded violation window reported there. Compressed bitmaps preserve batch speed without changing these version rules.

Algorithm 1 RevBound cut construction at model-call admission

1: Input: requester (u, a), assembled components $\{m_i\}$, guarded model-call handle
2: Obtain authenticated admission lease for policy version v; fail closed if no current fence is available
3: Fetch bitmap rows whose tenant and version equal v
4: **for** each fragment m_i **do**
5: Materialize $\Pi^*(m_i; u) = (E^*_{UA}, E^*_{AR})$; deny missing or inconsistent parents
6: Deny if $(u, a) \notin G_{UA}(v)$ or $E^*_{UA} \not\subseteq G_{UA}(v)$
7: Deny if $E^*_{AR} \not\subseteq G_{AR}(v)$
8: Otherwise allow m_i and emit allow reason code
9: **end for**
10: Revalidate the admission lease; restart on a higher fence
11: Admit typed `AllowedBundle` through the model-call handle; append its version, closure digests, and decisions

Algorithm 1 runs after rewrite and merge operations. Recomputed relational closure captures the typed authorization edges of every admitted component; lease revalidation prevents a policy change from racing model-call admission; and the certificate binds each allowed component to the decision that authorized its bytes.

Non-bypassable transformation and invocation. Rewrite, merge, template, and relay wrappers attach the handles of every model-visible input and materialize their relational union; agents cannot supply a smaller parent set. The client exposes model access only through a guarded `invoke(AllowedBundle)` path that rejects missing lease metadata, invalid closure digests, or malformed provenance and performs the final revalidation. Opaque outputs remain in a separate raw-only forensic interface with no route into prompt assembly or model invocation.

4.3 Conflict surfacing

Contradictions never overwrite fragments. RevBound instead creates a `Resolution` fragment to (i) connect the conflicting fragments; (ii) record which arbitration rule was used; and (iii) retain pointers to verifiers. This makes the process auditable and prevents silent drift in context.

Resolution ownership. When a conflict signal is raised, trusted workflow code creates Resolution fragments as *derived* artifacts that must cite every piece of parent evidence via `derived_from`. As a result, provenance closure applies to them after revocation, and the arbitration trail remains auditable. Deployments may also limit authorship of `Resolution` fragments and impose rate limits on them.

5 Security Argument and Executable Assumptions

Assumptions exercised by the implementation. A1 authenticates requester and agent identities within one tenant. A2 admits every model call through `invoke(AllowedBundle)` and revalidates its policy lease at transport admission. A3 makes covered transformations lineage-sound by attaching all model-visible input handles inside trusted wrappers and preserving exact user–agent and agent–resource pairs under union. A4 totally orders updates at the authoritative engine, rejects unfenced replicas and partitions, and represents re-grants as higher versions. These assumptions define the histories covered by Proposition 2.

Executable invariants. Every admitted component satisfies both typed edge relations at one admission-valid version; every actual dependency introduced through the covered API is included in its recorded closure; and no unchecked representation can enter inference transport. The fixed-parent attacks directly test these invariants. SB-1 shares the final policy-decision boundary but reconstructs fragment ancestry without typed-route enforcement; SB-2 propagates span labels but does not require a parent-complete fragment closure or the `AllowedBundle` type. Their residual events therefore locate closure and routing incompleteness within the final-boundary family. Churn, fault injection,

and degraded-provenance studies separately pressure policy fencing, cache consistency, and lineage soundness.

6 EXPERIMENTAL PROTOCOL

Questions. The experiments ask, in order: (Q1) how many revoked derivatives survive when the same transformations are checked before versus after rewrite; (Q2) how final gating and derived closure contribute, and where combined closure/routing incompleteness remains; (Q3) whether the resulting cut remains stable under churn and held-out transformations; and (Q4) whether it preserves useful shared memory across tasks and model stacks.

Units, aggregation, and matched budgets. Every pooled security count comprises five prespecified blocks whose disjoint denominators sum to the displayed n; raw x/n pools those blocks. RepoHoldout partitions 180 issues into five 36-issue blocks, and each τ-bench domain partitions 150 interactions into five 30-interaction blocks. SWE-Coop instead reruns the complete 300-issue set under each of five seeds (1,500 issue-runs) and reports mean$\pm$standard deviation across full-set seed rates. Continuous episode quantities use median [p25,p75], bounded rewrite rounds use mean [p25,p75], and request latencies use p50/p95/p99. Wilson intervals accompany nonzero pooled counts and exact one-sided Clopper–Pearson bounds accompany zeros. Comparisons pair every item within a block and freeze prompts, tools, retrieved candidates, rewrite outputs, policy histories, budgets, and schedules. Point differences between methods are therefore matched outcome contrasts; task-rate comparisons use the reported full-set seed dispersion.

6.1 MODELS, HARDWARE, WORKLOADS, AND COMPARATORS

The primary stack uses Qwen2.5-32B-Instruct (temperature 0.2, top-p 0.9), bge-large-en-v1.5 embeddings, and bge-reranker-large on Ubuntu 22.04 with 4$\times$A100 80 GB GPUs, CUDA 12.4, and PyTorch 2.3.1. ACSS-Permissions v2.0 drives 0, 20, 60, and 120 revokes/min. SWE-Coop evaluates 300 SWE-bench Lite issues (Jimenez et al., 2024) under each of five seeds with three cooperating agents per issue; RepoHoldout contains 180 issues from disjoint repositories. The second stack uses Llama-3.3-70B-Instruct (temperature 0.2, top-p 0.9) with the same retrieval models on 4$\times$H100 80 GB GPUs, CUDA 12.4, PyTorch 2.5.1, and vLLM 0.8.5. It evaluates fixed held-out sets of 150 τ-bench Retail and 150 Airline interactions.

Appendix Table 8 maps the four interventions component by component. SB-1 traverses fragment parents at serialization but lacks RevBound's relational schema and guarded output type. The independently implemented SB-2 propagates source labels through rewritten spans; span attribution can omit a parent that fragment closure retains. Agreement between these implementations provides evidence for a final-boundary class contrast, while their residual events and the component ablations identify lineage soundness and guarded routing as additional requirements.

For adaptive evaluation, attack-generation prompts and transformation operators are disjoint from protocol development; the attacker observes coarse allow/deny categories and adapts for at most three rewrite or relay rounds. A final-bundle *leak* is a model-visible proposition or instruction whose content depends on a source denied at admission; topical overlap without dependence is insufficient. The endpoint audit receives the revoked source, policy event, admitted bundle, and complete transformation trace after method identifiers are removed. Its prespecified binary rule marks exact copying directly and marks a semantic paraphrase only when the trace supports preservation of the source-specific proposition or directive; it uses no score threshold. *Extra deny* means that the monitored method blocks a bundle accepted by the fenced clean reference. Fixed attacks have mechanical labels from explicit parentage and directly test A1–A4; adaptive and cross-stack counts are trace-audited stress outcomes for dependencies beyond the recorded parent set. Appendix D details construction and labeling.

7 EVIDENCE FOR OBJECT-ALIGNED REVOCATION

The fixed-parent intervention produces the largest separation; subsequent stress tests locate its dependence on closure, fencing, and attribution.

Table 1: Matched test of the object checked after a source is revoked. Panel A reports final-bundle ASR with 95% Wilson intervals, or an exact one-sided bound for zero events; Panel B reports median denial time among denied episodes, extra denials, and p95 latency. SB-1 and SB-2 share the final locus but use fragment reconstruction and span labels, respectively. Pooling yields 600/1,160 earlier-object events, 56/1,160 for SB-1, 36/1,160 for SB-2, and 0/1,160 for RevBound.

Panel A: final-bundle attack success

Attack style	Earlier object	SB-1	SB-2 token lineage	RevBound
Re-summarization	219/320; 68.4 [63.2,73.3]%	23/320; 7.2 [4.8,10.7]%	15/320; 4.7 [2.9,7.6]%	0/320; UB $\leq 0.93\%$
Paraphrase	152/280; 54.3 [48.4,60.0]%	16/280; 5.7 [3.5,9.1]%	10/280; 3.6 [1.9,6.5]%	0/280; UB $\leq 1.06\%$
Partial copy + merge	128/300; 42.7 [37.2,48.3]%	11/300; 3.7 [2.1,6.5]%	7/300; 2.3 [1.1,4.7]%	0/300; UB $\leq 0.99\%$
Cross-agent relay	101/260; 38.8 [33.1,44.9]%	6/260; 2.3 [1.1,4.9]%	4/260; 1.5 [0.6,3.9]%	0/260; UB $\leq 1.14\%$
Pooled	600/1,160; 51.7 [48.9,54.6]%	56/1,160; 4.8 [3.7,6.2]%	36/1,160; 3.1 [2.2,4.3]%	0/1,160; UB $\leq 0.26\%$

Panel B: deny seconds [quartiles]; extra deny; p95 ms

Attack style	SB-1	SB-2 token lineage	RevBound
Re-summarization	0.10[0.05,0.19]; 11/320; 9.2	0.11[0.06,0.20]; 9/320; 8.7	0.04[0.02,0.08]; 8/320; 8.1
Paraphrase	0.12[0.06,0.21]; 13/280; 8.9	0.13[0.07,0.22]; 11/280; 8.5	0.06[0.03,0.12]; 10/280; 8.0
Partial copy + merge	0.15[0.08,0.27]; 15/300; 9.0	0.16[0.09,0.28]; 14/300; 8.8	0.08[0.04,0.16]; 12/300; 8.4
Cross-agent relay	0.09[0.04,0.17]; 9/260; 8.7	0.10[0.05,0.18]; 8/260; 8.3	0.05[0.03,0.10]; 7/260; 7.9

7.1 Derived-fragment laundering stress test

We replay the complete lifecycle from Figure 1: a source is revoked after its first authorized use, then reintroduced through re-summarization, paraphrase, partial-copy merging, or cross-agent relay. ASR is the fraction of episodes whose model-call bundle contains a forbidden derivative. Table 1 is the denominator-complete comparison across four implementations; Table 3 complements it with per-style dispersion and time-to-denial.

On the matched episodes, SB-1 and SB-2 yield paired risk differences of 46.9 and 48.6 percentage points from the earlier-object control, with the same ordering in all four attack styles. This implementation-diverse agreement establishes the boundary-class separation. RevBound removes the 56 and 36 residual events; the component intervention in Table 5 explains why by separately restoring 166/156,000 violations without derived closure and 1,693/156,000 without the final gate. Together, the comparator map and ablations identify final placement, sound closure, and guarded routing as cumulative cut requirements.

Held-out adaptation (Appendix Table 9) leaves the earlier-object control at 48.4–61.2% ASR, while both boundary-matched serializers remain below 7.2%. RevBound records 6/500 trace-audited attribution misses across the two families at 8.4–8.6 ms p95. New operators and attacker feedback preserve the large separation while locating the remaining pressure in semantic attribution rather than explicit-parent enforcement.

7.2 Clean-run exclusion under churn

Table 2 tests whether policy fencing and relational closure remain stable as versions turn over. The 156,000 denominator comprises 39,000 concurrent multi-user bundle reads at each churn level;

Table 2: ACSS-Permissions clean runs. RevBound records 0 observed violations over 156,000 bundle reads and 1.8M fragment checks; the bundle-level Clopper–Pearson 95% upper bound is $\leq 1.9 \times 10^{-5}$. Version turnover raises tail latency while the admission fence preserves exclusion.

Churn (revokes/min)	Bundle reads	RevBound violations	Latency p50/p95/p99 (ms)
0	39,000	0 / 39,000	1.3 / 2.1 / 3.0
20	39,000	0 / 39,000	1.5 / 3.2 / 4.8
60	39,000	0 / 39,000	1.9 / 5.4 / 7.6
120	39,000	0 / 39,000	2.4 / 7.6 / 10.8

expanding those bundles induces 1.8M fragment decisions. Injected cache and replica faults are excluded here and reported separately in Table 6.

Integrity composition. Appendix Table 4 composes the revocation cut with `MemLedger`, our local client-verifiable ledger comparator, to measure policy eligibility, record tampering, and stale-proof replay under compromised storage.

7.3 End-to-end utility, fixed retrieval, and cross-stack transfer

Three controls connect exclusion to useful work. RevBound resolves 28.4±1.1% of SWE-Coop issues versus 23.6±1.0% for shared vector memory, with fewer duplicate calls (Appendix Table 12). Fixed-memory RepoHoldout keeps resolution close at 51, 54, and 56 of 180 for the earlier-object control, SB-1, and RevBound, while leaks fall from 11 to 2 to 0 (Table 13). On the Llama-3.3/τ-bench stack, earlier-object versus RevBound leaks are 25/150 versus 0/150 on Retail and 30/150 versus 1/150 on Airline, with task success increasing from 104 to 109 and 89 to 94 (Table 15). The cut therefore preserves effective collaboration while reproducing the protected-output separation across retrieval controls, tasks, and model stacks.

8 Implications of Object-Aligned Revocation

The 51.7% versus 3.1–4.8% separation shows that retrieval compliance and model-input compliance diverge once memory is transformed. Reporting the protected bundle, typed closure, policy version, and decision boundary prevents write-time trust, retrieval filtering, action gating, and unlearning from being treated as interchangeable revocation evidence. RevBound's 0/1,160 fixed-parent events and 7.6–8.8 ms p95 further show that the relational cut can occupy the model-call path at modest measured cost; the ablations assign complementary roles to final gating, derived closure, and guarded routing.

The formal result covers calls admitted through tracked transformations whose authenticated lineage over-approximates actual dependence and whose final lease validation follows policy-update linearization. It excludes denied dependencies from subsequent calls, while already admitted calls and facts independently derived from authorized sources retain their policy status. Degraded parentage, adaptive attacks, and cross-stack tracing measure the transition beyond exact closure: 6/500 and 1/300 trace-audited events quantify attribution pressure alongside the fixed-parent guarantee.

9 Conclusion

Revocation is a property of the object admitted to the model, not merely the fragments returned by retrieval. Across 1,160 matched attacks, retrieval-time filtering admits 600 revoked derivatives; two final-boundary serializers admit 56 and 36, and RevBound none. Revocation cuts expose the uncovered paths, while object alignment identifies which comparisons test the same claim. With sound relational lineage, admission fencing, and guarded routing, RevBound records zero covered-lineage violations under churn at 7.6 ms p95 and retains effective shared-memory task performance. The resulting principle is actionable for agent systems: track transformations end to end and enforce revocation where their outputs become model input.

AI USE STATEMENT

Language-model tools assisted in manuscript preparation, including prose and presentation editing and reference and submission-format checks.

REPRODUCIBILITY STATEMENT

Section 6 specifies models, decoding, hardware, workloads, comparators, units, seed-block aggregation, metrics, and uncertainty conventions. Algorithm 1 defines the enforcement path; Appendix D and Tables 5–15 provide workload construction, component interventions, fault behavior, and transfer results.

REFERENCES

Owen Arden, Jed Liu, and Andrew C. Myers. Flow-limited authorization. In *2015 IEEE 28th Computer Security Foundations Symposium*, pp. 569–583, 2015. doi: 10.1109/CSF.2015.42. URL https://doi.org/10.1109/CSF.2015.42.

Manuel Costa, Boris Köpf, Aashish Kolluri, Andrew Paverd, Mark Russinovich, Ahmed Salem, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Securing AI agents with information-flow control. arXiv preprint arXiv:2505.23643, 2025. URL https://arxiv.org/abs/2505.23643.

Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2026. doi: 10.48550/arXiv.2503.18813. URL https://arxiv.org/abs/2503.18813.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/edac78c3e300629acfe6cbe9ca88fb84-Abstract-Conference.html.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html.

Haichang Li. Memory as a service (MaaS): Purpose-bound memory mediation for cooperative agents. In *KDD 2026 Workshop on Agentic Software Engineering*, 2026. doi: 10.48550/arXiv.2506.22815. URL https://arxiv.org/abs/2506.22815.

Peixuan Li and Danfeng Zhang. Towards a general-purpose dynamic information flow policy. In *2022 IEEE 35th Computer Security Foundations Symposium*, pp. 260–275, 2022. doi: 10.1109/CSF54842.2022.9919639. URL https://doi.org/10.1109/CSF54842.2022.9919639.

Yedidel Louck. Securing LLM-agent long-term memory against poisoning: Non-malleable, origin-bound authority with machine-checked guarantees. arXiv preprint arXiv:2606.24322, 2026. URL https://arxiv.org/abs/2606.24322.

Ciyan Ouyang and Rui Hou. MemLineage: Lineage-guided enforcement for LLM agent memory. arXiv preprint arXiv:2605.14421, 2026. URL https://arxiv.org/abs/2605.14421.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. arXiv preprint arXiv:2310.08560, 2023. URL https://arxiv.org/abs/2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Alireza Rezazadeh, Zichao Li, Ange Lou, Yuying Zhao, Wei Wei, and Yujia Bao. Collaborative memory: Multi-user memory sharing in LLM agents with dynamic access control. arXiv preprint arXiv:2505.18279, 2025. URL https://arxiv.org/abs/2505.18279.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68539–68551, 2023. doi: 10.52202/075280-2997. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R. Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Bin Wang, Fan Wang, Pingping Wang, Jinyu Cong, Yang Yu, Yilong Yin, Zhongyi Han, and Benzheng Wei. Agentic unlearning: When LLM agent meets machine unlearning. arXiv preprint arXiv:2602.17692, 2026. URL https://arxiv.org/abs/2602.17692.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=ehfRiF0R3a.

Jinghan Xu, Yiyong Xiao, Wanru Shao, Hankai Liu, and Xinjin Li. Memory provenance laundering in LLM agents: A non-amplification firewall for persistent memory. arXiv preprint arXiv:2607.29167, 2026. URL https://arxiv.org/abs/2607.29167.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R. Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2406.12045. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b126cc38b8638e07bef37e7b2bb72bf-Abstract-Conference.html.

Zhongming Yu, Naicheng Yu, Hejia Zhang, Wentao Ni, Mingrui Yin, Jiaying Yang, Yujie Zhao, and Jishen Zhao. Multi-agent memory from a computer architecture perspective: Visions and challenges ahead. arXiv preprint arXiv:2603.10062, 2026. URL https://arxiv.org/abs/2603.10062.

A Revocation-Cut Proofs

A.1 Post-decision boundary non-equivalence

Let s be a source, x an authorized fragment whose relational closure contains an edge contributed by s, and T a permitted history-dependent transformation. Choose $t_0 < t_r < t_e$: x is retrieved at t_0, that edge is revoked at t_r, and $T(x)$ is admitted to a model call at t_e. A retrieval decision at t_e can exclude x from the current candidate set. Because $T(x)$ was materialized from historical state and its path to model-call admission does not cross that decision, the guarded object can still contain $T(x)$. The earlier protected object satisfies its predicate while the model-call bundle contains a derivative carrying a denied edge. A final decision evaluates $\Pi^*(T(x); u)$ under an admission-valid version and removes the path. Summaries, merges, templates, and cross-agent relays differ only in the implementation of T.

A.2 Sound final-cut sufficiency

Consider any admitted component $m \in B(u, a, v)$. A2 requires its path to pass through guarded model-call admission and makes v valid at that linearization point. A3 derives $\Pi^*(m; u) = (E^*_{UA}(m; u), E^*_{AR}(m))$ by taking relational unions over every wrapper-observed input and establishes $D^*_{UA} \subseteq E^*_{UA}$ and $D^*_{AR} \subseteq E^*_{AR}$. Algorithm 1 admits m only if the live requester edge belongs to $G_{UA}(v)$, every $(u, a') \in E^*_{UA}(m; u)$ belongs to $G_{UA}(v)$, and every exact originating pair $(a', r) \in E^*_{AR}(m)$ belongs to $G_{AR}(v)$. A1 authenticates those principals and A4 ensures that a concurrent update either precedes admission or forces a restart at a higher version. Therefore every actual covered source-to-m path intersects one decision that checks a sound over-approximation of its typed authorization edges under the same admission-valid v. A denied pair keeps m out of the model-call `AllowedBundle`; applying the argument to every component proves Proposition 2.

B Additional System Diagnostics

Table 3: Laundering dispersion and response time by attack style. Seed-mean ASR characterizes the earlier-object control; RevBound deny time runs from the illicit derived write to the final-cut decision. Table 1 reports the aligned x/n comparison.

Attack style	Earlier-object ASR (%)	RevBound deny time (s)
Re-summarization	68.4 ± 3.2	0.04 [0.02, 0.08]
Re-encoding (paraphrase)	54.3 ± 3.8	0.06 [0.03, 0.12]
Partial copy + merge	42.7 ± 4.1	0.08 [0.04, 0.16]
Cross-agent relay	38.8 ± 3.6	0.05 [0.03, 0.10]

`MemLedger` is our local client-verifiable append-only ledger comparator. It authenticates fragment and provenance records and rejects stale proofs before RevBound evaluates policy eligibility; it does not implement RevBound's authorization predicate.

Table 4: Composition under compromised storage (1.0M trials, five seeds, 120 revokes/min, bundle size 32). RevBound and MemLedger protect complementary properties. Their fused 12.6 ms path shares snapshot fetch and decision logging, explaining why its p95 is below the 15.4 ms sum of isolated p95 values. Replay acceptance is not defined without a verifier.

Configuration	Policy violation	Tamper accepted	Verified replay accepted	p95 ms
Cost anchor only (no checks)	7.8%	12.4%	9.6%	0.0
MemLedger only	7.8	0.0	0.0	7.8
RevBound only	0.0	12.4	– (no verifier)	7.6
RevBound + MemLedger	0.0	0.0	0.0	12.6

Table 5: Per-churn ablation results. Count cells include raw $x/39{,}000$, the pooled rate per 10k reads, and its 95% Wilson interval on the same scale; Panel B reports latency at every churn setting.

Panel A: exclusion and utility

Churn	Without derived closure	Without final gate	Conflict overwrite resolve
0	18/39,000; 4.6 [2.9,7.3]	216/39,000; 55.4 [48.5,63.3]	26.4±1.1%; [25.1,27.7]
20	29/39,000; 7.4 [5.2,10.7]	337/39,000; 86.4 [77.7,96.2]	26.3±1.0%; [25.1,27.5]
60	47/39,000; 12.1 [9.1,16.0]	492/39,000; 126.2 [115.5,137.9]	26.0±1.3%; [24.4,27.6]
120	72/39,000; 18.5 [14.7,23.3]	648/39,000; 166.2 [153.8,179.5]	25.7±1.4%; [24.0,27.4]

Panel B: p95 boundary latency (ms)

Churn	RevBound	Without derived closure	Conflict overwrite
0	2.1	1.8	2.0
20	3.2	2.8	3.1
60	5.4	4.7	5.2
120	7.6	6.8	7.4

Table 6: Injected failure modes. The stale-bitmap denominator is 25,000 pooled reads across five seed blocks for each severity; 2/25,000 and 8/25,000 map to 0.008% and 0.032%. Outage, unfenced lag, and partition activate fail-closed fencing, while bypassing the bitmap-version check creates a short measurable leak window.

Failure	Final-bundle result	Recovery / utility
Policy service outage	0 violations; all denied	immediate fallback deny
Unfenced replica (10 s lag)	0 violations	+0.4% deny
Bypassed bitmap-version check	2–8 / 25,000 per severity; 0.008–0.032%	repaired in 0.3 [0.1,0.6] s
Partition during lease validation	0 violations; stricter deny	+1.2% deny

C Ablations and Fault Diagnostics

C.1 Ablation study

Table 5 expands the earlier 18–72 and 216–648 extrema into 39,000-read churn strata. Removing derived closure produces 166/156,000 pooled violations; removing the final gate produces 1,693/156,000. Conflict overwrite preserves exclusion but lowers aggregate resolve rate from 28.4±1.1% to 26.1±1.2%.

C.2 Implementation fault tests

Version-consistency checks repair stale bitmaps in 0.3 [0.1,0.6] s. The exact denominator-to-rate mapping separates the deliberately bypassed check from the fenced 156,000 clean reads in Table 2.

D Workload Construction and Endpoint Labels

Seed blocks and denominators. Count-based studies use five prespecified blocks generated from independent random seeds. The displayed denominator is the sum across blocks: 232 episodes per block for the 1,160 fixed-attack study, 50 per block for each 250-episode adaptive family, 7,800 reads per block at each ACSS churn rate, 36 issues per block for the 180-issue RepoHoldout study, and 30

interactions per block for each 150-interaction τ-bench domain. SWE-Coop instead evaluates the complete 300-issue set under each of five seeds, for 1,500 issue-runs, and reports full-set seed means and standard deviations. Block-level rates produce reported dispersion where shown; pooled counts produce Wilson intervals or the exact zero-event bound. This construction keeps the item, attack, policy history, and random seed paired across methods.

Fixed-parent attacks. Each episode begins with a source fragment that is authorized and consumed once. A policy update then revokes one user–agent or agent–resource edge carried by that source. Re-summarization, paraphrase, partial-copy merging, and cross-agent relay each produce a derivative with an explicit `derived_from` path. The final-bundle label is therefore mechanical: an admitted component is a leak exactly when its transitive typed closure contains the revoked edge at the admission version. The four styles contribute 320, 280, 300, and 260 episodes, with the same derivative, revocation point, and model-call request replayed across methods.

Adaptive families and endpoint rubric. The obfuscated-plan family preserves a revoked directive through held-out plan rewrites; the multi-hop family relays its content through successive agent messages. Each family contains 250 episodes and permits at most three transformations after observing coarse allow/deny feedback. Development prompts and operators are excluded from these sets. For these studies and the cross-stack study, the endpoint audit receives the canonical source, policy event, transformation trace, and final bundle after method identifiers are removed. The binary label is positive when the trace and output preserve a source-specific proposition or directive; exact copying is sufficient, paraphrase counts when the same claim or instruction survives, and shared topic alone is insufficient. The reported adaptive counts use this fixed rubric without a score threshold.

Churn and task workloads. ACSS-Permissions interleaves multi-user bundle reads with 0, 20, 60, or 120 committed revocations per minute; every admitted component is checked against the fence active at its model-call linearization point. SWE-Coop assigns three agents to each SWE-bench Lite issue under one shared tool budget, while RepoHoldout uses 180 issues from repositories disjoint from the primary set and freezes retrieval candidates across methods. The τ-bench overlay applies the same policy-update and endpoint-label rules to fixed Retail and Airline interactions. Success, tool-call, and duplicate-call metrics remain native to each task; revocation leakage is evaluated separately at the model-call bundle.

E Additional Object-Aligned Evaluation

E.1 Provenance-aware hygiene tier

The poisoning study composes RevBound with a two-tier hygiene interface. All writes enter a forensic raw tier; the planning path consumes an actionable tier whose promotion hooks record verifier identity and version. Low-reputation sources and high-impact fragment types remain raw-only until verification succeeds. Timeouts and missing evidence deny promotion, while final-bundle eligibility remains independently enforced under the admission-fenced policy version.

Table 7: RevBound mechanisms and their path-level effects at model-call admission.

Property	Statement (informal)	Enforced by
Admission-consistent revocation	revoked edges deny under one effective policy state	version fence + lease revalidation
Derived-path closure	derived artifacts inherit every cited source	schema checks + provenance union
Decision certification	each admitted component has a replayable reason	deterministic cut + evidence record
Conflict-preserving derivation	resolutions retain every conflicting parent	resolution fragments + typed links

Table 8: Intervention map for the four matched enforcement designs. Every row receives the same fragments, transformations, policy history, prompts, and seeds.

Method	Decision object	Dependency representation	Policy timing	Model route
Earlier object	retrieved fragments	fragment metadata before rewrite	lookup before transformation	ordinary assembly
SB-1	post-transform bundle	parent-graph reconstruction	lookup at serialization	ordinary serializer
SB-2	post-transform bundle	source labels on rewritten spans	lookup at serialization	ordinary serializer
RevBound	model-call bundle	sound typed relational closure	fenced revalidation at admission	guarded bundle

Table 9: Adaptive final-bundle attacks under open-world lineage-attribution pressure. Each held-out family contains 250 obfuscated-plan or 250 multi-hop-relay episodes across five blocks. The endpoint audit counts trace-supported semantic dependence.

Family	Method	ASR x/n; 95% CI	Mean rounds [IQR]	Extra deny	p95 ms
Obfuscated plan	Earlier object	153/250; 61.2 [55.0,67.0]%	1.6[1,2]	3/250	5.8
	SB-1	18/250; 7.2 [4.6,11.1]%	2.4[2,3]	15/250	9.8
	SB-2	12/250; 4.8 [2.8,8.2]%	2.5[2,3]	18/250	9.1
	RevBound	4/250; 1.6 [0.6,4.0]%	2.7[2,3]	12/250	8.6
Multi-hop relay	Earlier object	121/250; 48.4 [42.3,54.6]%	1.8[1,2]	2/250	5.6
	SB-1	12/250; 4.8 [2.8,8.2]%	2.5[2,3]	11/250	9.5
	SB-2	8/250; 3.2 [1.6,6.2]%	2.6[2,3]	14/250	8.8
	RevBound	2/250; 0.8 [0.2,2.9]%	2.8[2,3]	9/250	8.4

These four rows bind policy state, path closure, decision evidence, and conflict handling to the same model-admitted object. Integrity composition, response shaping, and hygiene extend the cut with separately measured properties.

E.2 Result-indexed validity matrix

Table 10 compactly separates the measured regimes so that fixed-parent exclusion, open-world attribution, utility, and storage integrity are not conflated.

E.3 Degraded provenance stress test

We reuse the laundering harness and hold prompts, retrieval candidates, attack scripts, policy histories, and five seed blocks fixed while weakening provenance quality. Each row aggregates 200 post-revocation episodes (4 attack styles $\times$ 10 episodes/style/block $\times$ 5 blocks). A *leak* means forbidden content still reaches the final bundle; *extra deny/raw-only* means the monitored path conservatively blocks or quarantines content that would otherwise have remained actionable.

Omitted parents and mixed transparent/opaque derivations route 83/200 and 74/200 episodes to raw-only handling. Partial instrumentation creates 7/200 leaks, empirically locating A2–A3 coverage as the path segment required for a complete cut.

E.4 Fixed-retrieval safety–utility frontier

The end-to-end SWE-Coop configuration first tests whether the guarded memory path remains useful under a complete collaborative workload:

The fixed-retrieval comparison holds memory contents, ranked candidates, prompts, conflict handling, and tool budgets constant across enforcement designs:

Table 10: Validity conditions for each result family. Panel A identifies the protected output and intervention; Panel B records held-out pressure, measured cost, and the failure surface exercised.

Panel A: alignment contract

Setting	Protected object / revocation cut	Attacker and trust assumptions	Security property	Strongest comparator
Clean-run churn	Final bundle / admission-fenced gate	Revoked typed edges; A1–A4 trusted	Fenced exclusion under churn	RevBound across four churn rates
Fixed laundering	Post-rewrite bundle / model-call gate	Explicit-parent attacker; matched prompts, candidates, policy versions	Placement and completeness effects	Earlier object, SB-1, SB-2, RevBound
Adaptive laundering	Three-round bundle / model-call gate	Coarse feedback; trace endpoint audit; held-out operators	Open-world attribution stress	SB-1 and token-lineage SB-2
Fixed retrieval	Code-repair bundle / model-call gate	Disjoint repositories; identical memory and retrieval	Safety–utility behavior	Earlier object, SB-1, three closure policies
Cross-stack	Retail/Airline bundle / model-call gate	Dynamic overlay; trace endpoint audit; matched budgets	Cross-stack protected-output effect	Two boundary-matched comparators on Llama-3.3
Degraded provenance	Final bundle with weakened lineage	Omitted, ambiguous, opaque, or partial parentage	Fail-closed erosion behavior	Raw-only and conservative-union policies
Compromised storage	Final bundle plus record integrity	Tamper/replay store; honest client verifier	Revocation plus integrity	MemLedger, RevBound, composed path

Panel B: held-out pressure, cost, and evidence

Setting	Adaptive / held-out condition	Benign utility or overhead	Known failure mode	Evidence
Clean-run churn	120 revokes/min	7.6 ms p95	Stale bitmap inconsistency	Table 2
Fixed laundering	Four rewrite styles	Per-style deny, extra-deny, p95	Incomplete parentage	Table 1; Appendix Table 3
Adaptive laundering	Disjoint attacks, three rounds	8.4–9.8 ms p95	Unrecorded semantic dependence	Table 9
Fixed retrieval	180 disjoint-repository issues; fan-in 1–64	Resolve, calls, duplicate ratio, extra deny	High fan-in availability cost	Tables 13, 14
Cross-stack	150 Retail + 150 Airline interactions	Task success, calls, duplicate ratio, latency	Domain-specific attribution miss	Table 15
Degraded provenance	Four lineage degradations	Raw-only / extra-deny and p95	Partial instrumentation leaks	Table 11
Compromised storage	1.0M tamper/replay trials	12.6 ms composed p95	Store omission	Table 4

Reading the matrix. The zero-event fixed-parent result is separate from trace-audited semantic dependence and from integrity or poisoning outcomes. Only the fixed-parent and clean-churn rows directly test Proposition 2.

Table 11: Degraded-provenance stress test. Omitted or opaque parents trigger raw-only handling, ambiguous lineage primarily increases denial, and partial instrumentation produces the largest leak count. The pattern maps a controlled transition from exact closure to conservative availability loss and then measurable leakage.

Degradation case	Leak count (x/n)	Extra deny or raw-only (x/n)	p95 latency (ms)
Omitted parents	0 / 200	83 / 200 raw-only	8.3
Ambiguous parentage	1 / 200	61 / 200 extra deny	8.9
Mixed transparent/opaque derivations	0 / 200	74 / 200 raw-only	8.7
Partial instrumentation	7 / 200	18 / 200 extra deny	6.8

Table 12: SWE-Coop results on all 300 SWE-bench Lite issues under each of five seeds (1,500 issue-runs). RevBound attains 28.4±1.1% resolution and a 0.15 median duplicate-call ratio.

Method	Resolve rate (%)	Tool calls/issue	Duplicate ratio
Chat-only	19.4 ± 0.8	10.8 [9.8,12.2]	0.34 [0.27,0.40]
Shared vector	23.6 ± 1.0	10.0 [9.1,11.2]	0.24 [0.19,0.30]
RevBound	28.4 ± 1.1	9.0 [8.2,10.2]	0.15 [0.11,0.20]

Starting from this controlled setting, we vary provenance fan-in and closure policy. Exact closure resolves explicit ancestors; conservative union merges every candidate parent scope; raw-only fallback withholds high-uncertainty derivatives.

The frontier separates retrieval quality from boundary policy. At fan-in 64, exact closure resolves 53/180 issues with 11 extra denials; conservative union resolves 47/180 with 22, and raw-only fallback resolves 43/180 with 32. All three retain zero observed leaks, making the availability cost of lineage uncertainty directly measurable.

E.5 Cross-stack Retail and Airline evaluation

The transfer study uses Llama-3.3-70B-Instruct and a dynamic-revocation overlay on fixed τ-bench Retail and Airline splits. Five seeds reuse identical simulated-user prompts, task goals, tool schemas, memory candidates, and revocation schedules across methods. As in the adaptive study, endpoint verification counts semantic source dependence beyond the recorded parent set, testing attribution robustness outside explicit-parent closure.

Across both domains, earlier-object filtering leaks in 55/300 interactions, SB-1 in 8/300, SB-2 in 5/300, and RevBound in 1/300. Retail task success increases from 104/150 to 109/150 between earlier-object filtering and RevBound; Airline increases from 89/150 to 94/150. Semantic endpoint verification therefore reproduces the boundary-class safety gap under a different model, hardware platform, interaction protocol, and non-coding workload.

F Complementary Security Evaluation

F.1 Side-channel evaluation

We model an attacker who can issue repeated requests and observe response timing and denial patterns at the model-facing memory gateway. Concretely, the attacker observes end-to-end wall clock latency and response-shape metadata such as the number of returned fragments and denials. We evaluate a balanced binary task where the hidden bit is whether a target permission edge is active in the admission-fenced version, so chance performance is 50%. Leakage is operationalized as the accuracy of a simple classifier trained on those response features, reported as mean±standard deviation across five seeded classifier fits on the same balanced protocol. Fixed-shape responses return a constant-size envelope with masked entries, and constant-latency mode adds padding. Overhead is measured relative to standard variable-shape operation.

Table 13: RepoHoldout results over 180 disjoint-repository issues in five 36-issue blocks. Fixed memory and retrieval preserve close resolution counts while final-bundle leaks fall from 11 to 2 to 0.

Method	Resolved	Tool calls	Duplicate ratio	Leak / extra deny
Earlier object	51/180; 28.3%	10.5[9.4,11.6]	0.23[0.18,0.29]	11/180 leak
SB-1	54/180; 30.0%	10.2[9.1,11.4]	0.20[0.16,0.25]	2/180 leak; 7/180 deny
RevBound	56/180; 31.1%	9.9[8.9,11.0]	0.18[0.14,0.23]	0/180 leak; 6/180 deny

Table 14: Safety–utility frontier on SWE-Coop-RepoHoldout (180 issues, five seeds). Panel A reports issue resolution, final-bundle leakage, and extra denial. Panel B reports median duplicate ratio, median tool calls, and p95 boundary latency. Exact closure preserves the strongest utility at moderate fan-in; conservative union and raw-only fallback trade availability for robust zero observed leakage as ancestry broadens.

Panel A: issue outcomes

Closure policy	Fan-in	Resolved	Leak	Extra deny
Exact	1	56/180	0/180	4/180
Exact	4	57/180	0/180	5/180
Exact	16	55/180	0/180	7/180
Exact	64	53/180	0/180	11/180
Conservative union	1	55/180	0/180	5/180
Conservative union	4	56/180	0/180	8/180
Conservative union	16	51/180	0/180	13/180
Conservative union	64	47/180	0/180	22/180
Raw-only fallback	1	53/180	0/180	9/180
Raw-only fallback	4	52/180	0/180	14/180
Raw-only fallback	16	48/180	0/180	21/180
Raw-only fallback	64	43/180	0/180	32/180

Panel B: operational cost

Closure policy	Fan-in	Duplicate ratio	Tool calls	p95 ms
Exact	1	0.18[0.14,0.23]	9.8[8.8,10.9]	7.9
Exact	4	0.17[0.13,0.22]	9.7[8.7,10.8]	8.3
Exact	16	0.19[0.15,0.24]	10.0[8.9,11.2]	9.1
Exact	64	0.22[0.17,0.28]	10.6[9.4,11.9]	11.8
Conservative union	1	0.19[0.15,0.24]	9.9[8.9,11.0]	7.6
Conservative union	4	0.18[0.14,0.23]	9.8[8.8,11.0]	8.0
Conservative union	16	0.23[0.18,0.29]	10.7[9.5,12.0]	8.8
Conservative union	64	0.27[0.21,0.34]	11.6[10.3,13.0]	10.9
Raw-only fallback	1	0.21[0.16,0.27]	10.3[9.2,11.5]	6.8
Raw-only fallback	4	0.24[0.19,0.30]	10.8[9.6,12.1]	7.1
Raw-only fallback	16	0.29[0.23,0.36]	11.7[10.4,13.2]	7.7
Raw-only fallback	64	0.35[0.28,0.42]	12.8[11.3,14.5]	8.5

The 72.4% standard-mode probe accuracy exposes response-shape information. Fixed-shape and constant-latency modes approach chance with distinct overhead points, giving deployments a measurable tuning surface for auxiliary leakage.

The measured frontier supports selecting a response-shaping point jointly with probe-rate limits, coarsened denial categories, and tenant-specific monitoring.

F.2 Poisoning robustness

PoisonBench-Memory injects adversarial fragments via untrusted resources and measures whether poisoned content is later reused in an episode's execution. ASR counts episodes where the injected objective influences an agent decision or action, such as when a poisoned instruction or template

Table 15: Cross-stack results on 150 held-out interactions per domain, partitioned into five 30-interaction seed blocks. Success reports pooled count and rate; calls and duplicate ratios are medians [p25,p75]. Latency reports p95 plus standard deviation across blocks. Both boundary-matched implementations sharply reduce final-bundle leakage, and RevBound preserves the strongest task completion with modest boundary cost.

Retail Method	Task success	Leak	Extra deny	Tool calls	Duplicate ratio	p95 ms
Chat-only	92/150; 61.3%	–	–	8.7[7.2,10.4]	0.18[0.13,0.25]	–
Shared vector	101/150; 67.3%	31/150	0/150	8.1[6.8,9.7]	0.14[0.10,0.20]	2.7±0.2
Earlier object	104/150; 69.3%	25/150	4/150	7.9[6.7,9.5]	0.13[0.09,0.19]	5.9±0.3
SB-1	106/150; 70.7%	3/150	8/150	8.0[6.8,9.6]	0.12[0.08,0.18]	9.4±0.5
SB-2	107/150; 71.3%	2/150	9/150	7.8[6.6,9.3]	0.11[0.08,0.17]	8.9±0.4
RevBound	109/150; 72.7%	0/150	7/150	7.7[6.5,9.2]	0.10[0.07,0.16]	8.5±0.4
Airline Method	Task success	Leak	Extra deny	Tool calls	Duplicate ratio	p95 ms
Chat-only	79/150; 52.7%	–	–	10.3[8.6,12.1]	0.24[0.18,0.31]	–
Shared vector	87/150; 58.0%	37/150	0/150	9.8[8.2,11.7]	0.19[0.14,0.26]	2.9±0.2
Earlier object	89/150; 59.3%	30/150	5/150	9.6[8.0,11.5]	0.18[0.13,0.25]	6.2±0.4
SB-1	91/150; 60.7%	5/150	10/150	9.5[7.9,11.3]	0.16[0.11,0.22]	9.7±0.6
SB-2	93/150; 62.0%	3/150	12/150	9.3[7.8,11.1]	0.15[0.10,0.21]	9.2±0.5
RevBound	94/150; 62.7%	1/150	9/150	9.2[7.7,11.0]	0.14[0.10,0.20]	8.8±0.5

Table 16: Side-channel mitigation trade-off. Fixed response shape reduces policy-state classifier accuracy by 15.6 points for 4.2% latency overhead; constant-latency padding reaches 53.2% accuracy for 8.6% overhead.

Mode	Classifier accuracy (%)	Latency overhead (%)	Notes
Standard (variable-shape)	72.4 ± 2.8	baseline	deny patterns leaky
Fixed-shape responses	56.8 ± 3.4	$+4.2 \pm 0.8$	substantially reduced, still above chance
Constant-latency mode	53.2 ± 3.1	$+8.6 \pm 1.4$	strongest mitigation; 3.2 points above chance

is retrieved and followed. Detection steps count retrieval rounds from the first injected write until the poisoned fragment, or one of its descendants, is quarantined or demoted from the actionable tier. The experiment measures the incremental protection obtained by composing provenance-aware promotion and quarantine with final-bundle eligibility.

Table 17: PoisonBench-Memory results over five seeded workload runs. The composed hygiene path reduces attack success by 24.2, 19.8, and 15.2 points across the three injection styles. Non-zero RevBound-stack ASR quantifies residual semantic attack risk, and detection latency is reported in retrieval rounds.

Injection type	Naive ASR	RevBound-stack ASR	Detection rounds
Instruction injection	36.4 ± 3.0	12.2 ± 1.7	5 [4, 7]
Data poisoning	29.2 ± 3.4	9.4 ± 1.3	6 [5, 9]
Consistency traps	22.8 ± 2.8	7.6 ± 1.1	7 [6, 10]

Promotion and quarantine materially reduce every attack type, while the 7.6–12.2% residual range separates semantic verifier quality from RevBound's revocation property. This distinction gives operators a direct criterion for strengthening content verification without changing the model-call admission contract.

F.3 Operational contract details

Assumption-to-failure mapping. A1 binds graph edges to authenticated principals. A2 makes `AllowedBundle` the unique prompt-consumption type and validates its lease at admission. A3 supplies complete relational parentage for anti-laundering closure. A4 orders policy updates, fences stale replicas, and fails closed when freshness cannot be established. Table 11 exercises A2–A3 erosion, Table 6 injects A4 consistency faults, and Table 4 adds a client-verifiable integrity path for tamper and replay.

Operational invariants. Assuming authenticated identities, an admission-valid policy version for $G_{UA}(v)$ and $G_{AR}(v)$, and fail-closed validation of relational provenance, no admitted fragment carries a typed edge denied at that version. Because derived fragments union exact parent relations and model wrappers accept only `AllowedBundle`, revoking any represented source edge denies every cited descendant after the update linearizes. Propositions 1–2 connect this implementation invariant to path coverage.

G Evaluation Accounting and Deployment Guidance

RevBound treats audit records, denial categories, and verifier outputs as sensitive derived data. Fixed-shape envelopes, coarse reason codes, rate limiting, and access-controlled evidence records complement the final-bundle cut. The central tables report denominator-complete bundle, episode, issue, and interaction outcomes; the appendix connects each aggregate to uncertainty, operational cost, and a specific cut assumption.

Benchmark-specific compute accounting. ACSS-Permissions, laundering, degraded provenance, composition, and fault injection run the gateway, policy engine, and logging stack on Ubuntu 22.04 and record per-block wall-clock time, policy-version traces, and decision logs; together these campaigns consume 61 A100 GPU-hours and 184 CPU-hours. SWE-Coop, RepoHoldout, frontier, and PoisonBench-Memory use Qwen2.5-32B-Instruct on 4×A100 80 GB and consume 551 GPU-hours, for 612 A100 GPU-hours overall. The Llama-3.3 τ-bench campaign runs on 4×H100 80 GB and consumes 428 H100 GPU-hours.

The evaluation stack uses Qwen2.5-32B-Instruct (Apache-2.0), Llama-3.3-70B-Instruct (Llama 3.3 Community License), BAAI bge-large-en-v1.5 and bge-reranker-large (MIT), SWE-bench Lite materials (MIT), and τ-bench materials (MIT). ACSS-Permissions, PoisonBench-Memory, Conflict-Inject, adaptive attack splits, and policy traces follow the protocols reported in the paper. The protected-object matrix and denominator ledger make the decision boundary, unit, uncertainty convention, and hardware explicit for every central revocation result.