AUDIENCE-BOUND PERSISTENT MEMORY: AUTHORIZATION ACROSS THE MEMORY LIFECYCLE

Anonymous authors
Paper under double-blind review

ABSTRACT

A personal language agent that acts for its owner across private and shared conversations can learn a fact from one audience and later place it in the context it assembles for another. We study *authorization before context* across the whole memory lifecycle. Each memory item carries the audience present when it was recorded; derived items are partitioned by audience, receive the intersection of their sources' audiences, or are suppressed; an audience widens only by an explicit, object-specific grant; and an item enters a model attempt only when every current viewer belongs to one of its authorized audiences, with unresolved viewers failing closed to public-only. Under explicit identity, provenance and complete-mediation assumptions, this admission is sound and policy-complete on the exact assembled context, enforced by exclusion rather than by model behavior. We realize it in two independently persisted reference architectures, a flat store and a relationship graph, and, descriptively, in a native agent-memory runtime. In a prospectively frozen confirmation over 10,000 multi-party histories, no forbidden item entered any architecture's context, whereas unscoped retrieval exposed forbidden items in 82% of its contexts. Entitled recall matched policy-equivalent baselines exactly and exceeded unscoped retrieval by 0.30 Recall@5, with a Holm-confirmed advantage that grows with distractors. No architecture produced a wrong-principal substitution, but unscoped substitutions were too rare to establish the prespecified joint decision.

1 INTRODUCTION

A personal language agent acts as a delegated proxy for its owner. On the owner's behalf it takes part in private conversations with the owner and in one-to-one and group conversations with other participants, who may be people or other agents, across transports such as email, team chat and text messages. What it learns in all of these conversations accumulates in one long-term memory, although each fact was shared with a particular *audience*: the participants present when it was said. For each model call, the agent assembles a context from the current conversation's recent messages and from long-term memory, which holds both what was said and what is derived from it, such as summaries and knowledge-graph facts. The resulting threat is compositional: a fact shared with one audience could be stored, transformed, retrieved on a later turn and placed in the context assembled for a different audience. Once it is there, later defenses are too late. The model has already read it; redacting the response cannot undo that exposure, and scoring the answer cannot detect it, because a model can answer acceptably even when its context (instructions, conversation history and retrieved memory) contains a forbidden fact.

Three common designs fall short. Relevance-only retrieval ignores who is present on the current turn. Output filtering acts after the model has read the context. Per-user or per-session silos avoid exposure by giving up legitimate recall: a fact shared in a group that includes Bob is exactly what Bob's later one-to-one conversation is entitled to recall (Figure 1). We instead treat the problem as one of authorization in the memory system and the agent harness around the model, not as a behavior the model should learn. We keep one shared memory but label each item with its audience, read from the transport when the item is written (always including the owner), carry the label through every transformation, and check every path from memory into a model's context before the model is called: an item is admitted only when every current viewer already belongs to one of its authorized audiences.

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.

This is a relationship check in the style of Pang et al. (2019); related predicates appear in access control, in authorization-first retrieval over static documents (Namboothiri, 2026), and at a personal agent's memory-to-context transition (§2). What is missing is its *lifecycle*: how labels are bound at write time, computed for derived memory, widened only by explicit decisions, kept current, and enforced for every physical model call across different stores. Because a fully mediated retriever never returns an inadmissible item, the central empirical question is the cost in recall: whether evidence the audience is entitled to see is still found, and how retrieval behaves when highly relevant facts about *other* participants compete for the same slots.

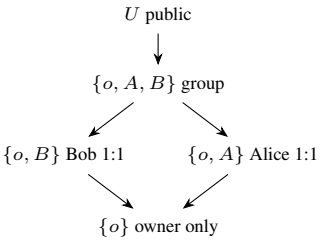
Contributions.

1. **Lifecycle authorization model** (§3): origin-audience binding at write time; four closed derivation dispositions; exact, non-unioned declassification grants that never carry over to derived items; and admission against one authorization snapshot per physical model attempt.
2. **Conditional guarantee** (§3.5): under the identity, provenance and complete-mediation assumptions of §3.4, no model call's context contains an inadmissible item (soundness), and an exhaustive query returns every admissible one (policy completeness). In both reference architectures an excluded item also never occupies a candidate slot, so it cannot displace an admissible one.
3. **Three realizations** (§4): two independently persisted reference architectures that share only the policy package, and a realization inside an existing agent-memory runtime, validated descriptively.
4. **A prospectively frozen evaluation** (§5): a 10,000-history paired confirmation with a pre-specified non-inferiority margin, closed testing, Holm-controlled superiority, conformance testing and complete outcome accounting. No forbidden item entered any of 200,000 exact contexts; all eight non-inferiority checks pass with no recall cost against policy-equivalent baselines, and recall superiority and the dose slope are confirmed; substitutions (0 vs. 7) were too rare to confirm H3a/b, so the joint decision is not established (§6).

2 Related Work

Authorization before context. Authorization-First Retrieval (Namboothiri, 2026) argues that, in multi-agent retrieval over documents, authorization must precede any learned component, since retrieve-then-filter pipelines expose unauthorized content structurally. Filtered vector search shows that attribute filters break the recall assumptions of approximate indexes (Chronis et al., 2025). Liu (2026) applied an audience-membership test at a personal agent's memory-to-context transition, proved one-way confinement and fail-closed behavior for the rule, and reported no forbidden inclusions on a 79-scenario synthetic suite against unscoped projections that include them by construction. We add what they leave open: derived labels, exact grants, lifecycle currentness, per-attempt admission, independent realizations and a prospectively powered recall study.

Governed and multi-user memory. Collaborative Memory (Rezazadeh et al., 2025) shares memory among users and agents through configured, time-varying access graphs and provenance-tagged



Item (audience)	Group turn	Bob 1:1	Unresolved
f_1 recorded in group $\{o, A, B\}$	✓	✓	–
f_2 recorded with Alice $\{o, A\}$	–	–	–
f_3 public (U)	✓	✓	✓

Figure 1: Admission by audience membership. Left: the audience lattice; an item flows downward, to containers whose viewers are a subset of its audience, never upward. Right: which items enter each turn's context. Bob's one-to-one recalls the group fact f_1 but never Alice's fact f_2; a turn with unresolved viewers fails closed to public items.

fragments; governed shared-memory services use scopes and roles (Margalit et al., 2026; Bhardwaj et al., 2026); and multi-user agents and deployment-time memorization expose the privacy failure surface (Yang et al., 2026b; Chen et al., 2026b). Our audiences are derived from the transport identities of the participants present rather than configured, and the endpoint is the exact context of each model attempt.

Privacy and memory benchmarks. Contextual-integrity and privacy benchmarks for memory and multi-user agents mostly score model output or task utility (Mireshghallah et al., 2026; Gupta et al., 2026; Ren et al., 2026), and multi-party memory benchmarks score utility (Yang et al., 2026a; Hu et al., 2026). We score the assembled context itself (GateMem's construct limitation is in App. D).

Authority, integrity and information flow. Other work asks what memory authorizes an agent to *do* or whether memory can be trusted (Zhan et al., 2026; Cerruti et al., 2026; Xu et al., 2026; Louck, 2026; Wang et al., 2026; Chen et al., 2026c; Saidi, 2026); we share its write-time origin binding but address read-audience confidentiality only. Our labels and grants follow lattice and decentralized information-flow control and declassification (Bell & La Padula, 1976; Denning, 1976; Myers & Liskov, 1997; Sabelfeld & Sands, 2009), and our derived-data rules follow provenance-based policy (Bates et al., 2013; den Hartog & Zannone, 2016); agent information-flow systems (Costa et al., 2025; Debenedetti et al., 2025; Palumbo et al., 2026) instead govern control and effect flows under prompt injection. We implement only the recipient slice of contextual integrity (Nissenbaum, 2009). Appendix D gives a detailed comparison.

3 Lifecycle Authorization Model

This section extends a single audience-membership admission rule to writes, derivations, grants and model attempts.

3.1 Participants, viewers and audiences

Let U be the set of *participants*, people or agents, with a distinguished owner $o \in U$ and an anonymous participant $\star \in U$ that stands for any viewer who cannot be resolved. A *container* C is one conversation turn; its *viewers* $V(C) \subseteq U$ are the participants present on that turn. The agent takes part on the owner's behalf, so $o \in V(C)$ always. When the viewer evidence (the transport identities of those present, §3.4) is missing, stale, low-confidence or conflicting, $V(C) = U$ (fail closed). Groups and channels are containers, not participants, and an audience is identified by its member set, not by the channel on which it is observed.

Each memory item m carries an *audience* $\mathrm{Aud}(m) \subseteq U$, the participants who may see it; the owner belongs to every audience. An audience has one of three kinds: *public*, $\mathrm{Aud}(m) = U$, which any participant may see; *owner-private*, $\{o\}$; and *shared*, the resolved viewers A of the container where m was recorded, with $o \in A \subsetneq U$. Only a public audience contains $\star$. We use *participant* throughout; the frozen protocol's names *principal postings* and *wrong-principal substitution* use *principal* in the same sense.

The lifecycle adds three objects. *Write-time binding* sets $\mathrm{Aud}(m)$ of a recorded item to the resolved viewers of its origin container (or U for a public write) and fails closed: a write whose origin viewers cannot be determined is refused rather than bound to a guessed audience. A *derived* item receives an audience from its sources (§3.3). A *grant* adds one exact, object-specific audience of resolved participants; $\mathrm{Grants}(m, t)$ holds the grants in force at time t, and the authorized audiences $\mathrm{AuthSets}(m, t) = \{\mathrm{Aud}(m)\} \cup \mathrm{Grants}(m, t)$ are evaluated separately, never unioned. Otherwise an audience can only be narrowed (App. E.2).

3.2 Admission

Let $\mathrm{Current}(m, t)$ hold when m and every transitive source are provenance-complete at time t and none is superseded, frozen, revoked or quarantined. Container C at time t admits

$$\mathrm{Adm}(C, t) = \{\, m : \ \mathrm{Current}(m, t) \ \wedge \ \exists A \in \mathrm{AuthSets}(m, t) : \ V(C) \subseteq A \,\}. \tag{1}$$

Table 1: Eq. (1) by viewer evidence (every admitted item must also be current).

Viewer evidence for C	Admitted memory
Owner only, $V(C) = \{o\}$	Every item; reading widens no audience
Resolved viewers	Items with some $A \in \mathrm{AuthSets}(m, t)$, $V(C) \subseteq A$
Missing, stale, low-confidence or conflicting	$V(C) = U$: public items only

Table 2: Derivation dispositions for an item derived from sources S.

Disposition	Resulting audience	When
Partition	One item per source audience	Sources may stay separate
Intersection	$\bigcap_{s \in S} \mathrm{Aud}(s)$ (always $\ni o$)	Indivisible artifact
Declassification	Adds one grant; $\mathrm{Aud}(m)$ unchanged	Authenticated, object-specific decision
Suppress / quarantine	None: not created, or quarantined and never current	Incomplete provenance, or a source reachable only via a grant

Without grants and lifecycle state this is exactly the rule of Liu (2026). Table 1 instantiates it; admission is by exclusion, so an inadmissible item never reaches the model.

3.3 Derivation and grants

Every persisted transformation takes one of four closed dispositions (Table 2). Derivation reads each source's Aud, never its grants, so a grant cannot become authority for a new item that might outlive it.

3.4 Threat model, attempts and trusted computing base

For a container C, the protected assets are the items, derived artifacts and fields that are not admissible for C under Eq. (1). Table 3 maps the failures we consider to the mechanism that addresses each. The endpoint is the exact context of each *physical* model attempt, checked against one authorization snapshot pinned at its linearization point and sealed before the model is called; later mutations govern later attempts, and retries or fallbacks are new attempts. The *oracle* is the code that evaluates Eq. (1); Figure 2(a) shows where it acts.

Transports and identity. A *transport* is the messaging service that carries a conversation, such as an email service, a team-chat workspace or a text-messaging network. Each names senders and recipients by its own authenticated identifier: an email address, a workspace account identifier, a phone number. An *identity resolver* maps each one to a single participant, and ambiguous or conflicting resolution fails closed. The trusted computing base is therefore these transport identities, the resolver, provenance capture, the oracle, stored audiences and grants, and every read path's call to the oracle; the model, rankers, answers, output filters and stored text are untrusted.

3.5 Guarantee

The four properties of the single-rule setting (Liu, 2026) carry over with $\mathrm{AuthSets}$ in place of Aud, and derivation adds a fifth (Table 4; proofs in App. A).

Let $\widehat{\mathrm{Adm}}(C, t)$ be what the *implementation* admits and $Q(C)$ the items satisfying the query's non-authorization filters.

Proposition 1 (Conditional soundness and policy completeness)**.** *If the trusted computing base is correct and every memory-to-context path invokes the oracle on one snapshot, then (i)* $\widehat{\mathrm{Adm}}(C, t) \subseteq \mathrm{Adm}(C, t)$*, and (ii) a backend that declares exhaustive retrieval returns every item of* $\mathrm{Adm}(C, t) \cap Q(C)$.

Part (i) says the implementation admits nothing outside Eq. (1); part (ii) says it rejects nothing inside it. Neither is a retrieval-quality promise, since ranking and a finite k can still lose entitled evidence, so recall is measured.

Table 3: Threats and the mechanism addressing each. Forged transport identities and wrong identity resolution are trusted-base failures, not defended.

Threat or failure	Addressed by
A viewer elicits a fact recorded for another audience	Admission by exclusion (P1)
Relevant facts about another participant compete for slots	Authorization before ranking; slot noninterference (§3.5, H3)
Missing, stale, low-confidence or conflicting viewer evidence	Fail closed: public-only reads (P3), refused writes
Poisoned memory tries to widen its own audience	Write-time binding (P4)
Summaries, merges and derivations mix audiences	Derivation dispositions (P5)
Direct lookup, graph or aggregate paths bypass retrieval	Complete mediation (Prop. 1)
Source or grant changes during an attempt	One snapshot per attempt

Table 4: Properties of Eq. (1). P1–P4 extend the single-rule properties; P5 is new.

Property	Statement
P1 one-way confinement	$m \in \mathrm{Adm}(C,t) \Rightarrow V(C) \subseteq A$ for some $A \in \mathrm{AuthSets}(m,t)$
P2 anti-monotonicity	$V(C_1) \subseteq V(C_2) \Rightarrow \mathrm{Adm}(C_2,t) \subseteq \mathrm{Adm}(C_1,t)$
P3 fail-closed soundness	$V(C) = U \Rightarrow$ only current public items are admitted
P4 poisoning containment	An injected item's audience is fixed at write time and cannot widen itself
P5 derivation non-expansion	An intersection item's audience lies within every source's; no grant is inherited

Context exclusion and slot noninterference. The proposition guarantees *context exclusion*: the model never sees an inadmissible item. Both reference architectures add *slot noninterference*: every candidate is authorized before ranking by per-item scores fixed before authorization, so an inadmissible item never occupies a fetch or context slot and cannot displace an admissible one (confirmed in all 10,000 histories, §6.3). Full *selection noninterference* would also remove inadmissible text from shared ranking statistics, which requires per-audience index statistics; it is not part of the guarantee.

Scope. Grant revocation, deletion and member removal are simple under Eq. (1) and take effect from the next attempt's snapshot (App. E.2). Out of scope are transport authentication, same-audience contextual-integrity norms, physical erasure and unlearning, revocation within a sealed attempt, side channels, action authorization and general prompt-injection defense.

4 Architecture and Implementation

4.1 Common contract and reference architectures

Figure 2(b) shows the realizations. All share one code-owned policy package: canonical identifiers and audience digests, the oracle for Eq. (1) and Table 1 with closed reason codes, the derivation and grant rules of §3.3, and the exact-context artifact schema. They share neither storage nor candidate queries, so differential conformance testing between them is not tautological, and each constrains candidates *before* ranking. B-Flat is a durable normalized store with separate item, audience, source, derivation and grant records; it authorizes the exhaustive candidate set and then ranks deterministically. B-Graph is a durable relationship store whose typed participant, audience, item, provenance, derivation and grant relations constrain candidate enumeration before the same ranking and context assembly. Both persist to SQLite through independent schemas and query paths and are the two confirmatory architecture classes.

Four evaluation-only arms share the same queries, memory state, model, prompts and deterministic lexical index (built once; fetch budget 20, $k = 5$): unscoped relevance retrieval, a fixed session silo, a bounded post-filter (rank the top 20 without authorization, then drop inadmissible ones before assembly; not output filtering), and principal postings. The last stores, for each participant p, the postings $N_p = \{(m, A) : p \in A \in \mathrm{AuthSets}(m, t)\}$ and admits m when some pair (m, A) appears in the postings of every current viewer; it is policy-equivalent to the oracle, so it is a strong baseline.

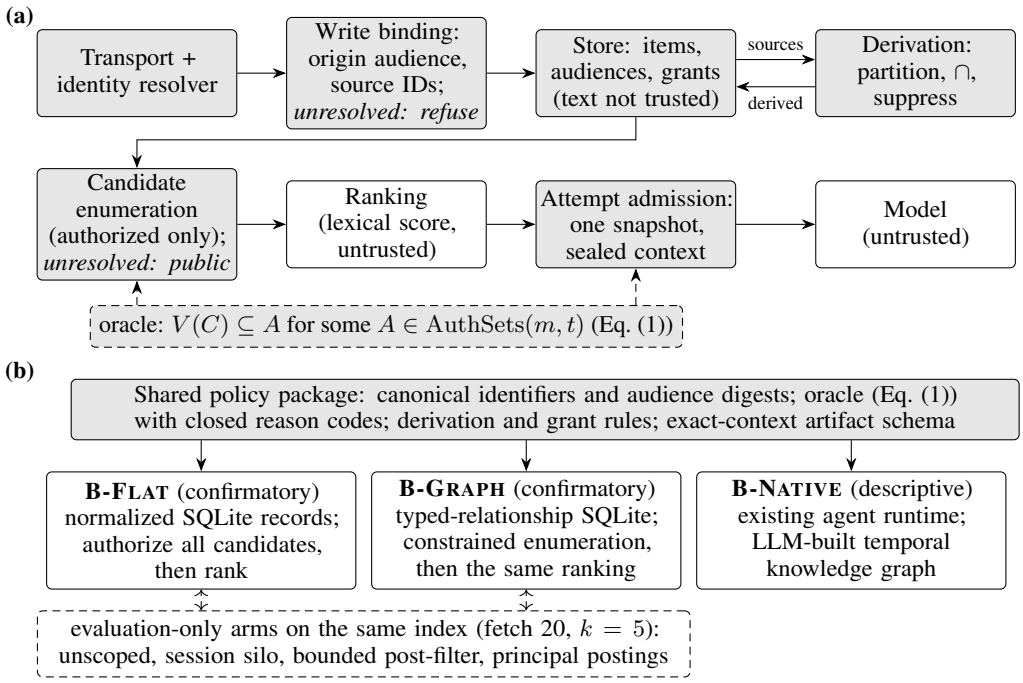


Figure 2: (a) Write–transform–read lifecycle. Shaded components belong to the trusted computing base, and italics mark the fail-closed rules; the same oracle constrains candidate enumeration before ranking and re-checks every physical model attempt; ranking, stored text and the model are not trusted to enforce confidentiality. (b) Realizations: two independently persisted reference architectures share only the policy package and are compared with evaluation-only arms; the native realization is reported descriptively.

4.2 NATIVE REALIZATION

B-NATIVE realizes the same contract inside an existing personal-agent memory runtime whose long-term store includes a temporal knowledge graph built by an LLM extraction pipeline. We extended the existing component responsible for each write, derivation, retrieval, context-carrier and provider boundary rather than adding a parallel authorization service; native-specific mechanisms are in App. E.3. It is validated descriptively, outside the confirmatory family (§6.4), for persistence, restart, concurrency and mediation in the research configuration.

5 EVALUATION DESIGN

Protocol, sample plan and decision rules were frozen before any confirmatory outcome was inspected (amendments in App. B.3); Table 5 and Figure 3 summarize them.

Table 5: Hypotheses, tests and outcomes. H1 holds on every exact context; H2 and the Holm family (decided under the worst-case supplement) are complete.

	Claim	Test	Status
H1	No forbidden item in any context	Exposure per attempt; suites	**0 of 200,000** (§6.1)
H2	Entitled recall is non-inferior	8 paired bounds > -0.05	**All 8 pass** (§6.2)
H3a/b	Fewer wrong-principal substitutions	Answer-level rate; Holm	Null not rejected (Tab. 7)
H3c	Advantage grows with distractors	Slope on $\log_2(d{+}1)$; Holm	Null rejected (Tab. 7)
H4a/c/d	Derivation never widens audiences	Transformation, mutation	Cases pass (§6.4)
H4b	Intersection keeps availability	Intersection vs. suppression	8/8 vs. 0/8 (App. E.7)

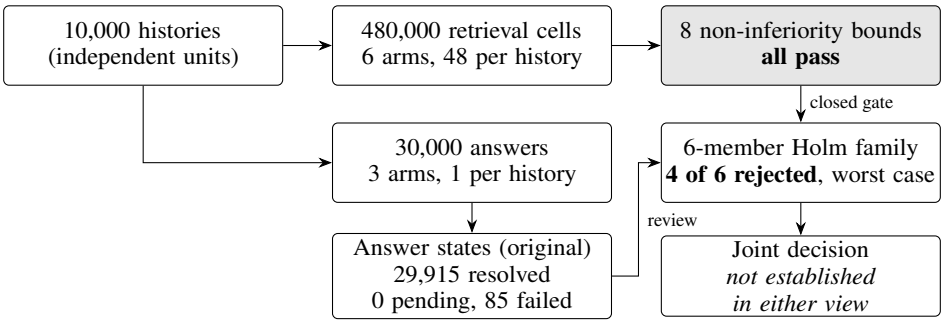


Figure 3: Confirmatory pipeline and status. Retrieval gates are complete and pass; every contract-valid answer is judged, and the worst-case supplement decides the Holm family (§6.3). Answer states include all saved judgments; the amended recovery view differs only in answer states (App. C).

5.1 Design

Confirmation frame. The frame is 10,000 independently generated multi-party histories (draws 40–10,039, 1,250 in each of eight themes), from a controlled event law whose ledger fixes participants, containers, audiences, viewers, gold evidence and answer targets before language rendering. Histories cover cross-channel recall, same-predicate distractors about *other* participants at doses $d \in \{0, 1, 2, 4, 8\}$ on the temporal-update question (paired advantage $\delta(d)$; the six-question endpoint uses $d = 8$), and abstention when only unauthorized evidence exists. Draws 0–39, including the 32-history pilot, are excluded. Inference concerns this generator, not natural prevalence.

Arms, models and other evidence. Each history yields 48 paired retrieval cells over the six arms of §4.1 and one answer each from unscoped retrieval, B-Flat and B-Graph. All confirmation model calls use one pinned model (`gpt-5.6-luna`, medium reasoning, no fallback, observed identity checked; limits in App. B), and one shared lexical index per history (FTS5 BM25) ranks candidates without embeddings. An external frame of independently authored corpora with their original gold labels (500 LongMemEval (Wu et al., 2025), 1,981 LoCoMo (Maharana et al., 2024) and 68 EverMemBench (Hu et al., 2026) questions) shows recall parity across all six arms (App. E.4).

5.2 Inference

For history h and an architecture–comparator pair, $D_h \in [-1, 1]$ is the mean over six equally weighted questions of the paired authorized Recall@5 difference; the estimand is $\mu = \mathbb{E}[D_h]$, with no imputation. Because histories are independent and bounded, the frozen test is the one-sided empirical-Bernstein bound of Maurer & Pontil (2009), $\bar{D} - \sqrt{2V_n \ln(2/\alpha)/n} - 7r \ln(2/\alpha)/(3(n-1))$, with $n = 10{,}000$, sample variance V_n and statistic range r ($r = 2$ for D_h). H2 passes only if all eight bounds at $\alpha = 0.025$ exceed -0.05, i.e., on average at most one extra complete miss of gold evidence per twenty questions. The gate then opens a six-member Holm family at $\alpha = 0.05$ (recall superiority, reduced substitution and the H3c slope for each architecture), with p-values from the frozen tests (Table 8). The plan gave a minimum simultaneous Monte Carlo lower bound of 0.935 on joint power over 40 scenarios (target 0.90).

Answer states and views. Each answer is resolved, pending semantic review, a terminal failure or an unresolved recovery (App. C). We report the original protocol and an amended view in which terminal citation-contract failures had one recorded recovery attempt, and select neither as primary.

Answer judging. Under an approved amendment, AI judges apply the frozen rubric to non-exact answers. Each judging route is a recorded provenance stratum, no judge saw keys or prior labels, and every judgment is schema- and citation-validated before it is saved (App. E.5).

Table 6: Entitled-recall non-inferiority (10,000 histories, both views): comparator Recall@5 (six questions, $d = 8$), paired difference and one-sided 97.5% bound (margin -0.05).

Comparator	R@5	Mean Δ	97.5% bound
Unscoped	0.674	$+0.298$	$+0.2900$
Session silo	0.740	$+0.233$	$+0.2283$
Bounded post-filter	0.972	0.000	-0.0020
Principal postings	0.972	0.000	-0.0020
B-FLAT, B-GRAPH	0.972	all 8 pass	

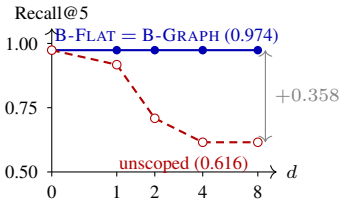


Figure 4: Recall@5 of the temporal-update question at distractor dose d (descriptive; 95% intervals within ± 0.008).

6 RESULTS

6.1 NO FORBIDDEN ITEM REACHES THE MODEL (H1)

Exclusion holds by construction (Proposition 1), and the campaign confirms it: none of the 200,000 exact contexts assembled by B-FLAT and B-GRAPH, nor the 180,000 of the silo, post-filter and postings arms, contained a forbidden item, whereas unscoped retrieval placed 220,385 forbidden items into 82,266 of its 100,000 contexts (82%), per the scorer's stored exact-context counts. Conformance testing agrees: 94 differential cases passed, 71 mutants were killed and 24 transformation cases passed on an earlier revision, and all 30 selected checks passed on the evaluated one.

6.2 ENTITLED RECALL IS NON-INFERIOR (H2)

Authorization before ranking costs no recall. All eight prespecified non-inferiority checks pass at the -0.05 Recall@5 margin (Table 6), identically for both architectures and answer views. Against the policy-equivalent bounded post-filter and principal postings, the paired difference is exactly zero in every history (eight distractors never exhausted the post-filter's 20-item window); against unscoped and session-silo retrieval, the architectures recover more entitled evidence ($+0.298$ and $+0.233$); the ordering holds at every k from 1 to 20 (Figure 6).

6.3 WRONG-PRINCIPAL COMPETITION (H3) AND THE JOINT DECISION

The Holm tests confirm recall superiority over unscoped retrieval ($+0.298$) and an advantage that grows with distractor dose (H3c slope $+0.130$ on $\log_2(d+1)$) in both views; both are retrieval-only (Table 7; Figure 5 in App. C). Figure 4 shows the mechanism: unscoped Recall@5 falls from 0.974 to 0.616 as distractors about other participants displace entitled evidence, whereas both architectures hold 0.974 because no distractor reaches their candidates: in all 10,000 histories they fetched the same 20 candidates and context at every dose, and unscoped retrieval did so in none. Neither architecture produced a wrong-principal substitution: B-FLAT 0 of 9,965 and B-GRAPH 0 of 9,966 judged answers, against 7 of 9,984 for unscoped retrieval (amended view: 0, 0 and 7 of 9,992–9,997). At this base rate the frozen test cannot confirm the reduction (App. C), so H3a/b is an underpowered null, not evidence against the mechanism; an exploratory sign test (not prespecified) finds all 7 discordant pairs favoring each architecture (one-sided $p = 0.0078$); the joint decision is *not established in either view*.

6.4 DERIVATION AND NATIVE VALIDATION (H4)

Transformation and mutation tests exercise H4a, H4c and H4d case by case; in the matched H4b study, derived items stayed available under intersection in 8 of 8 cases and in none under blanket suppression (App. E.7). In the native frame, AI review marked 195 of 210 answers correct in 42 scenarios (four unfavorable cases in App. E.6), and a scoped native qualification passed all 39 phases.

Outcome accounting. All 10,000 histories, 480,000 retrieval cells and 30,000 planned answers are retained. Under the original protocol, 85 answers (35/34/16 for B-FLAT/B-GRAPH/unscoped) failed the citation contract; the amended view recovered 67, leaving 18 in ten histories (App. C).

Table 7: Six-member Holm family ($\alpha = 0.05$), worst-case missingness supplement. Substitution: identification range of the unscoped-minus-target difference, upper p (†).

	Original view			Amended view		
Holm member	Estimate	p	Rej.	Estimate	p	Rej.
Recall, B-FLAT	$+0.298$	$< 10^{-6}$	yes	$+0.298$	$< 10^{-6}$	yes
Recall, B-GRAPH	$+0.298$	$< 10^{-6}$	yes	$+0.298$	$< 10^{-6}$	yes
Substitution, B-FLAT	$[-0.0028, +0.0023]$	$1^\dagger$	no	$[+0.0000, +0.0010]$	$1^\dagger$	no
Substitution, B-GRAPH	$[-0.0027, +0.0023]$	$1^\dagger$	no	$[-0.0001, +0.0010]$	$1^\dagger$	no
Dose slope, B-FLAT	$+0.130$	$< 10^{-6}$	yes	$+0.130$	$< 10^{-6}$	yes
Dose slope, B-GRAPH	$+0.130$	$< 10^{-6}$	yes	$+0.130$	$< 10^{-6}$	yes
Joint decision	not established			not established		

7 LIMITATIONS

The guarantee is only as strong as its trusted base: a misresolved transport identity yields a wrong decision that the oracle executes faithfully, so identity resolution is where a deployment must invest; the identity study (App. E.7) shows that every fail-closed policy removes exposure at a measured cost in entitled recall. The property is read-audience confidentiality at the context boundary; full selection noninterference additionally requires per-audience index statistics (§3.5). Confirmatory inference covers one controlled generator, one answer model and one lexical ranker, and the external corpora establish recall parity without cross-audience ground truth. Semantic outcomes are AI judgments without the approved consistency audit, and external answers mix three generation models, so neither is independent or human validation. Native evidence is descriptive, and its historical conformance receipts predate the evaluated revision.

8 CONCLUSION

Authorization before context extends across the memory lifecycle: no forbidden item entered any exact context, recall matched policy-equivalent baselines at every k and increasingly exceeded unscoped retrieval under distractors, and no architecture substituted a wrong principal, although unscoped substitutions were too rare to establish the joint decision. Confidentiality in agent memory is therefore a property of the memory architecture rather than of model behavior: binding audiences at write time and admitting before ranking costs no recall and removes the slot competition that degrades unscoped memory, with a guarantee that holds on the exact context of every model attempt.

Future work. The same boundary discipline extends from confidentiality to integrity and action. Origin-bound authority would admit content not authored by the owner or an explicitly delegated proxy only as attributed, quarantined data, never as instructions, trusted memory or effect authority, containing prompt-injection damage by construction rather than by detection (Louck, 2026; Debenedetti et al., 2025; Costa et al., 2025). An effect gateway would commit external actions only through deterministic, source-closed, single-use authorizations bound to exact arguments, so a hijacked model can at most propose an effect. Task runtimes for open-ended research, browsing and code execution would be treated as compromised, their outputs reaching memory or effects only through these mediated boundaries. Within the present design, per-audience index statistics would add full selection noninterference, and substitution studies at higher base rates would resolve H3a/b.

AI USE STATEMENT

Generative AI was used substantially and in roles the ICLR 2027 policy requires us to disclose. AI coding assistants were used to implement the reference architectures, the native-runtime changes, the evaluation harness, the analysis code and tests, under the authors' specifications and review; they also drafted and reviewed specifications, experimental and statistical designs, and performed technical reviews of implementation and analysis packets. AI judgments are part of the evidence: pre-confirmation answer review combined approved AI-assisted judgments with ten judgments by a study author; the native answers were labeled by the executing AI session; and confirmation answers were judged by AI judging sessions on several recorded routes (the approved AI consistency audit was

not run before submission) (Section 5.2 and App. E.5), whose limits we report. AI coding sessions also operated, checked and admitted the final judging windows under study approval. AI assistants also helped formulate the formal model and its mathematical claims, drafted the proofs, interpreted results, created figures, performed literature search, citation verification and reference formatting, and drafted and revised this manuscript. The synthetic histories were generated by AI-assisted generator code whose language is rendered by deterministic templates without model calls. The evaluated answer models, including the three external generation models, are subjects of study, not author aids. The authors directed the research, claims and limitations and take responsibility for the final content, including text, claims and artifacts produced with generative AI.

Ethics statement

All confirmatory and native experiments use synthetic histories and run in fresh isolated research state; no private conversation, real participant identity or production trace is used or released. Independently authored corpora are used under their recorded licenses with disclosed adaptations: LongMemEval (MIT), EverMemBench (Apache-2.0) and GateMem (CC-BY-4.0); LoCoMo (CC-BY-NC-4.0) and GroupMemBench (no asserted license) are not redistributed. The mechanism restricts what memory reaches a model and does not create new capabilities for misuse; a misconfigured identity resolver could, however, grant memory to the wrong audience, which is why the guarantee is stated conditionally and unresolved viewers fail closed.

Reproducibility statement

Section 5 and Appendix B specify the frozen protocol, sample plan, model and embedding pins, decision rules and answer-state taxonomy. An anonymized research artifact accompanies this submission as supplementary material (CC BY-NC 4.0 for project-authored material; third-party notices preserved). It contains the policy oracle, the independent B-FLAT and B-GRAPH stores and the evaluation-only baselines, the controlled generator, the conformance suites, 365 tests and the analysis code, with a claim-to-evidence map naming each table's and figure's command, inputs and expected output. One command, `reproduce_all.sh`, runs every check from a clean copy without model calls: the generator regenerates the controlled scenarios byte-identically; the conformance replay reproduces 94 of 94 differential cases, 71 killed mutants and 24 transformation cases; the supporting replay reproduces all four tables; and a headline replay regenerates every headline number from per-history reductions with the frozen estimator code (H1 counts, Tables 6 and 7 in both views, Figures 4 and 6, substitution counts, the sign test and slot noninterference), checking each value exactly against the submission run. The native source slice and the large per-history outputs and judgment records are released with the final version. Model-backed commands for fresh stochastic replication are documented separately and require the reader's own model access.

References

Anshul Bagaria, Sowmya S Sundaram, Gokul S Krishnan, and Balaraman Ravindran. Judging LLM-as-a-judge: Concerning rubric artifacts in LLM-based automated text generation evaluation, 2026. URL https://arxiv.org/abs/2609.02942.

Adam Bates, Ben Mood, Masoud Valafar, and Kevin Butler. Towards secure provenance-based access control in cloud environments. In *Proceedings of the Third ACM Conference on Data and Application Security and Privacy (CODASPY '13)*, pp. 277–284. ACM, 2013. doi: 10.1145/2435349.2435389. URL https://sts.cs.illinois.edu/papers/paper/TowardsSecureProvena20130219.html.

D. Elliott Bell and Leonard J. La Padula. Secure computer system: Unified exposition and Multics interpretation. Technical report, MITRE Corporation, Bedford, MA, March 1976. URL https://doi.org/10.21236/ADA023588. DTIC accession ADA023588.

Varun Pratap Bhardwaj, Garima Singh, and Arun Pratap Bhardwaj. SuperLocalMemory 4.0: The governed memory operating system for AI agents, 2026. URL https://arxiv.org/abs/2608.08253. Version 2.

Tommaso Cerruti, Mika Okamoto, and Ansel Kaplan Erol. Agent memory is a surface for endogenous authorization laundering, 2026. URL https://arxiv.org/abs/2609.01836.

Jianlin Chen, Wenhui Chen, Ziyao Lin, and Chi Man Vong. A judge should know what changed: Construct validity for LLM-as-a-judge evaluation, 2026a. URL https://arxiv.org/abs/2608.24419.

Lei (Rachel) Chen, Guilin Zhang, Kai Zhao, Dalmo Cirne, Andy Olsen, Xu Chu, Zeke Miller, Alet Blanken, Amine Anoun, and Jerry Ting. Deployment-time memorization in foundation-model agents, 2026b. URL https://arxiv.org/abs/2606.10062. ICML 2026 Workshop on Memory in Foundation Models (MemFM), per arXiv comments.

Xuanze Chen, Xukang Xie, Wentao Fu, Jiajun Zhou, Shanqing Yu, and Qi Xuan. MemSecBench: Tracking agent memory poisoning from persistence to consequence and repair, 2026c. URL https://arxiv.org/abs/2607.27080.

Yannis Chronis, Helena Caminal, Yannis Papakonstantinou, Fatma Özcan, and Anastasia Ailamaki. Filtered vector search: State-of-the-art and research opportunities. *Proceedings of the VLDB Endowment*, 18(12):5488–5492, 2025. doi: 10.14778/3750601.3750700. URL https://www.vldb.org/pvldb/vol18/p5488-caminal.pdf.

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, 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, 2025. URL https://arxiv.org/abs/2503.18813.

Jerry den Hartog and Nicola Zannone. A policy framework for data fusion and derived data control. In *Proceedings of the 2016 ACM International Workshop on Attribute Based Access Control (ABAC '16)*, pp. 47–57. ACM, 2016. doi: 10.1145/2875491.2875492. URL https://doi.org/10.1145/2875491.2875492.

Dorothy E. Denning. A lattice model of secure information flow. *Communications of the ACM*, 19 (5):236–243, May 1976. doi: 10.1145/360051.360056. URL https://doi.org/10.1145/360051.360056.

Shubham Gupta, Nazanin Mohammadi Sepahvand, Abhinav Kumar, Cem Subakan, Spandana Gella, Pierre-André Noël, Perouz Taslakian, Eugene Bagdasarian, and Valentina Zantedeschi. PiSAs: Benchmarking contextual integrity in multi-user agentic systems, 2026. URL https://arxiv.org/abs/2607.05318.

Chuanrui Hu, Tong Li, Xingze Gao, Hongda Chen, Yi Bai, Dannong Xu, Tianwei Lin, Xiaohong Li, Yunyun Han, Jian Pei, and Yafeng Deng. Evaluating long-horizon memory for multi-party collaborative dialogues, 2026. URL https://arxiv.org/abs/2602.01313.

Sibo Liu. Authorization before context: A model-neutral audience boundary against cross-audience memory leakage in agentic systems, 2026. URL https://arxiv.org/abs/2608.17148. Non-archival workshop paper, AdvML-Frontiers x CoTMA at COLM 2026.

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

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Yanki Margalit, Nurit Cohen-Inger, Erni Avram, Ran Taig, and Oded Margalit. Governed shared memory for multi-agent LLM systems, 2026. URL https://arxiv.org/abs/2606.24535.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Conference on Learning Theory (COLT)*, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Niloofar Mireshghallah, Neal Mangaokar, Narine Kokhlikyan, Arman Zharmagambetov, Manzil Zaheer, Saeed Mahloujifar, and Kamalika Chaudhuri. CIMemories: A compositional benchmark for contextual integrity in LLMs. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9a2bcfaf383638e166162a25b6dff125-Abstract-Conference.html.

Andrew C. Myers and Barbara Liskov. A decentralized model for information flow control. In *Proceedings of the Sixteenth ACM Symposium on Operating Systems Principles (SOSP)*, pp. 129–142. ACM, October 1997. doi: 10.1145/268998.266669. URL https://doi.org/10.1145/268998.266669.

Rohith Namboothiri. Authorization-first retrieval: Enforcing least privilege in multi-agent RAG systems. In *Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026)*, pp. 256–271. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.trustnlp-main.15. URL https://aclanthology.org/2026.trustnlp-main.15/.

Helen Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009. doi: 10.1515/9780804772891. URL https://www.sup.org/books/law/privacy-context.

Nils Palumbo, Sarthak Choudhary, Jihye Choi, Guy Amir, Prasad Chalasani, and Somesh Jha. Formal policy enforcement for real-world agentic systems, 2026. URL https://arxiv.org/abs/2602.16708.

Ruoming Pang, Ramon Caceres, Mike Burrows, Zhifeng Chen, Pratik Dave, Nathan Germer, Alexander Golynski, Kevin Graney, Nina Kang, Lea Kissner, Jeffrey L. Korn, Abhishek Parmar, Christina D. Richards, and Mengzhi Wang. Zanzibar: Google's consistent, global authorization system. In *2019 USENIX Annual Technical Conference (USENIX ATC 19)*, pp. 33–46. USENIX Association, 2019. URL https://www.usenix.org/conference/atc19/presentation/pang.

Yangda Peng, Yunjia Qi, Haotian Xia, Guanzhong He, Xintong Shi, Richeng Xuan, Songyuanyi Lu, Yixian Liu, Zhichao Hu, Yuhong Liu, and Hao Peng. Can LLM-as-a-judge reliably verify rubrics in agentic scenarios?, 2026. URL https://arxiv.org/abs/2606.29920. Version 2.

Zhe Ren, Yibo Yang, Yimeng Chen, Zijun Zhao, Benshuo Fu, Zhihao Shu, Bingjie Zhang, Yangyang Xu, Dandan Guo, and Shuicheng Yan. GateMem: Benchmarking memory governance in multi-principal shared-memory agents, 2026. URL https://arxiv.org/abs/2606.18829.

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, 2025. URL https://arxiv.org/abs/2505.18279.

Andrei Sabelfeld and David Sands. Declassification: Dimensions and principles. *Journal of Computer Security*, 17(5):517–548, 2009. doi: 10.3233/JCS-2009-0352. URL https://doi.org/10.3233/JCS-2009-0352.

Walid Saidi. MutMem: Cryptographically authorized mutation in persistent agent memory, 2026. URL https://arxiv.org/abs/2608.02843.

Yiqi Wang, Zihao Yan, Jiaqi Zhang, Zhangkai Wu, Mingkai Zheng, Zequn Sun, Yanming Zhu, and Taotao Cai. MAP-Graph: Provenance-aware shared memory for multi-agent workflows, 2026. URL https://arxiv.org/abs/2608.10509.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html.

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

Jingbo Yang, Kwei-Herng Lai, Xiaowen Wang, Shiyu Chang, Yaar Harari, and Evgeniy Gabrilovich. GroupMemBench: Benchmarking LLM agent memory in multi-party conversations, 2026a. URL https://arxiv.org/abs/2605.14498.

Shu Yang, Shenzhe Zhu, Hao Zhu, José Ramón Enríquez, Di Wang, Alex Pentland, Michiel A. Bakker, and Jiaxin Pei. Multi-user large language model agents, 2026b. URL https://arxiv.org/abs/2604.08567.

Qiuyang Zhan, Rui Zhang, Sheng Guo, Lepeng Zhao, and Zhuotao Liu. When memory becomes authority: Benchmarking authority collapse at the memory consolidation boundary, 2026. URL https://arxiv.org/abs/2608.01679.

A PROOFS

Notation follows §3. Fix an authoritative snapshot at time t.

P1. Immediate from Eq. (1): membership in $\mathrm{Adm}(C,t)$ requires some $A \in \mathrm{AuthSets}(m,t)$ with $V(C) \subseteq A$. Hence no item reaches a viewer outside every authorized audience. □

P2. If $V(C_1) \subseteq V(C_2)$ and $m \in \mathrm{Adm}(C_2,t)$, the witness A with $V(C_2) \subseteq A$ also satisfies $V(C_1) \subseteq A$, and currentness does not depend on C; so $m \in \mathrm{Adm}(C_1,t)$. Adding a current grant to $\mathrm{AuthSets}(m,t)$ can only add admitting containers. □

P3. With unresolved viewers, $V(C) = U \ni \star$. Owner-private and shared audiences exclude $\star$ (§3.1), and so does every grant, which names resolved participants only; so $V(C) \subseteq A$ holds only for $A = \mathrm{Aud}(m) = U$, and $\mathrm{Adm}(C,t) = \{m : \mathrm{Aud}(m) = U \wedge \mathrm{Current}(m,t)\}$. Implementations store the three audience kinds as typed labels and an unresolved viewer state as $V(C) = U$, so this is exactly the rule "unresolved viewers see public items only". □

P4. Write-time binding sets $\mathrm{Aud}(m') = A$ for an injected item m', where A is the set of resolved viewers of its origin container, independent of content; a write whose origin viewers cannot be determined is refused. A grant requires an authenticated, object-specific decision that content cannot mint, and derivation reads Aud only. By P1, m' reaches only containers with $V(C) \subseteq A$ unless such a decision is made. This is a confidentiality property, not an integrity guarantee. □

P5. For an intersection-derived item d with sources S, $\mathrm{Aud}(d) = \bigcap_{s \in S} \mathrm{Aud}(s) \subseteq \mathrm{Aud}(s)$ for every s; $o \in \mathrm{Aud}(d)$ because o belongs to every audience, and $\star \in \mathrm{Aud}(d)$ only if every source is public. Any $V(C) \subseteq \mathrm{Aud}(d)$ therefore satisfies $V(C) \subseteq \mathrm{Aud}(s)$ for every source. Because derivation reads Aud and never Grants, no source grant enters $\mathrm{Aud}(d)$. □

Proposition 1. The canonical basis selector $\sigma(m,C,t)$ enumerates the members of $\mathrm{AuthSets}(m,t)$ separately and returns $\mathrm{Aud}(m)$ if $V(C) \subseteq \mathrm{Aud}(m)$, otherwise the canonical minimum grant audience containing $V(C)$, and $\perp$ if none exists. The implementation admits m iff $\mathrm{Current}(m,t)$ and $\sigma(m,C,t) \neq \perp$. A caller-supplied binding that is invalid, stale or swapped is rejected even when another valid binding exists; that is a fail-closed error, not evidence of inadmissibility, and is tested separately; part (ii) below concerns requests without such an error.

(i) By complete mediation, m entered the attempt only through this check against the snapshot at t, so $\mathrm{Current}(m,t)$ holds and $\sigma(m,C,t) = A \neq \perp$ with $A \in \mathrm{AuthSets}(m,t)$ and $V(C) \subseteq A$; hence $m \in \mathrm{Adm}(C,t)$. *(ii)* If $m \in \mathrm{Adm}(C,t)$, some $A \in \mathrm{AuthSets}(m,t)$ contains $V(C)$; since σ enumerates every member of $\mathrm{AuthSets}(m,t)$ it returns an authorizing audience, so the implementation admits m, and an exhaustive backend returns every admitted item of $Q(C)$. □

B FROZEN PROTOCOL, MUTATIONS AND DEVIATIONS

B.1 FROZEN PROTOCOL SUMMARY

Table 8: Frozen confirmation protocol.

Item	Frozen value
Confirmation histories	10,000 (draws 40–10,039), eight themes $\times$ 1,250
Retrieval cells	48 per history; 480,000 total; one deterministic replicate
Answers	1 per history for unscoped, B-FLAT, B-GRAPH; 30,000 planned
Attempts	≤ 2 identical-input attempts per answer; ≤ 330 per 100-history chunk; $\leq 33{,}000$ total
Retrieval	shared deterministic lexical index per history (FTS5 BM25, reciprocal-rank fusion); fetch budget 20; $k = 5$; ties by item identifier
Answer model	`gpt-5.6-luna`, medium reasoning, no fallback, observed-identity check
Embedding	`text-embedding-3-small` pinned (1,536 dims, unit-ℓ_2, cosine); not used by confirmation ranking
Primary endpoint	paired authorized Recall@5 difference
Non-inferiority	margin -0.05; one-sided empirical-Bernstein bound, $\alpha = 0.025$; all 8 must pass
Superiority	6 contrasts after the gate; Holm familywise $\alpha = 0.05$
H3c	paired OLS slope of $\delta(d)$ on $\log_2(d+1)$, $d \in \{0, 1, 2, 4, 8\}$
Stopping	no interim outcome looks; no optional stopping
Missing outcomes	no imputation, substitution or silent exclusion
Pilot	32 held-out histories, 5 answer replicates, excluded from confirmation
Planned joint power	min. simultaneous MC lower bound 0.935 over 40 scenarios (target 0.90)
Arm order	seeded Latin rotation
Execution	subscription transport; 3 shared workers; no paid fallback
Extraction fitness	20 episodes vs. human annotation: entity recall 1.0, relation recall 28/29, precision 1.0

B.2 MUTATION OPERATORS

Common operators: reverse the subset relation; treat unknown viewers as owner-private; omit one current viewer; add one unauthorized audience; bypass authorization on direct lookup; merge derived audiences by union; admit incomplete-provenance summaries; let a source grant authorize a new persisted derivation or let a derived artifact inherit it; skip endpoint cascade on graph edges; remove an owner-policy branch; let query order select the reason code; reduce bounded over-fetch compensation; union two separately approved grants; reject a valid grant on an owner-private item or mutate the origin label when applying it; let a model generate or rewrite an authoritative identifier or source closure; authorize after ranking; change the sealed context after its snapshot. Native-only operators additionally target write-producer inventory, identity-only entity reuse, stale currentness or authorization decisions, snapshot placement relative to capacity waiting and retries, post-transport currentness reads, binding forgery, duplicate authorization-registry entries, realm and audience partitioning, projection collapse and bounded refill.

B.3 AMENDMENTS AND DEVIATIONS

1. **Pre-confirmation human review.** The originally required concealed-repeat or independent-adjudicator reliability check was replaced, by a prospective amendment recorded before confirmation, with study-approved AI-assisted judgments plus ten direct judgments by a study author without arm or system labels (descriptive agreement 10/10; repeat concealment not established). The amendment changed no hypothesis, population, sample, margin, multiplicity rule, power target or success rule.
2. **Source adjudication.** One LoCoMo question whose target is a named vehicle and whose retained answer abstained is recorded as an answer miss, not a wrong-principal substitution;

its retrieval recall is unscorable because the original question has no evidence annotations. This was a post-review source adjudication, not a blinded rating.

3. **Extraction fitness.** An earlier fitness run failed; a bounded 20-episode replacement passed. The failed run and its charges are retained and receive no fitness credit.
4. **Pilot.** The pilot was not rerun after later code repairs; it is reported historically at its original revisions and excluded from confirmation.
5. **Citation recovery.** After the original protocol completed, a separately recorded amendment allowed a recovery attempt for original terminal citation-contract failures. Both the original-protocol and the amended views are reported; neither is selected as primary.
6. **Analysis correction.** A comparison in the analysis adapter required completion publication order to equal physical attempt order and thereby rejected three valid answers of one history. The corrected analysis compares completed records by unique observation identity while preserving full-content equality; evidence-bound revalidation then resolved one of the three answers and left two awaiting review (both since judged). The other 9,999 histories and all 10,000 retrieval reductions are byte-identical before and after, no answer was regenerated, no admission mismatch remains, and an independent technical review accepted the correction.
7. **Judging route.** For semantic answer review, an approved study amendment restricted to this phase replaced the requirements that judging use the pinned model and that the answer model never be the sole judge with AI judging and no human assessors; unavailable model metadata is recorded as unknown. It does not amend any other call class. Judging began in one AI coding-assistant session; a later approved override moved it to parallel judge sessions (`gpt-6-sol`, low reasoning effort) on disjoint batches under one coordinating session (`gpt-6-astra`, extra-high effort). A judge context that reaches its recorded context limit receives no further work; each new `gpt-6-sol` context is qualified individually, and no judge context has seen keys or prior labels. A later study-approved switch moved the remaining queue to fresh-session `claude-opus-5-5` judges at low effort, which share one procedure-level qualification (18 controls) reused unchanged across windows; each route is recorded as a provenance stratum. When the validator rejected one internally inconsistent judgment, the queue stopped by design; an study-approved successor with the same judges, rubric and validators resumed it from the last durable snapshot, with its production checks done by the executing AI coding session under study approval (neither independent nor human), and the 17 subjects of the rejected batch were judged once more (the rejected response is retained). Parked subjects were then closed in one final pass: 170 were admitted from existing saved judgments (86 previously validated; 84 validated by the unchanged validator at admission) and 107 received one fresh judgment, leaving none pending; both successor windows were checked the same way. Rubric, schema and citation validation are unchanged, and earlier saved judgments are retained.
8. **Audit scope.** The study leads approved an AI consistency audit by the same judging route (App. E.5) in place of independent semantic validation (not run before submission; §5.2), and required that this limitation be reported.
9. **Recovery hold.** After 67 of the 85 original failures were recovered, further recovery calls were put on hold; the 18 remaining identities are retained.
10. **Missingness supplement.** After the recovery hold, a lightweight worst-case supplement over compatible completions of genuinely missing binary outcomes was conditionally authorized; it keeps the shared unscoped outcome, the eight gates and the full Holm family, is reported separately from the frozen unavailable result, and was executed once all labels were saved (§6.3).
11. **External answer models.** External answers were generated under three recorded model strata (App. E.4) instead of the single pinned model, and eight valid responses recorded outside the run journal leave a selection question open: for seven identities the earliest valid completion differs from the journal's recorded choice. Both selections are retained and neither is promoted; external results are reported per stratum.

C Outcome Accounting

Retrieval. Planned retrieval cells per arm: 100,000 each for unscoped, B-FLAT and B-GRAPH; 60,000 each for silo, bounded post-filter and principal postings; 480,000 in total. All 10,000 per-history retrieval reductions are complete and identical across the two answer views.

Answers. Table 9 gives per-arm states.

Table 9: Answer states per arm (10,000 planned answers each; generated from the accepted export, excluding the review overlay). *Pending* means contract-valid but not yet semantically reviewed; it is not a correctness judgment; every pending answer was judged afterwards (§6.3).

	Original protocol			Amended recovery		
Arm	Resolved	Pending	Terminal	Resolved	Pending	Unresolved
B-FLAT	4,812	5,153	35	4,825	5,168	7
B-GRAPH	4,807	5,159	34	4,815	5,177	8
Unscoped	5,103	4,881	16	5,113	4,884	3
Total	14,722	15,193	85	14,753	15,229	18

No answer is in the *disputed*, *exact-value contradiction* or *admission failed* state in either view. Under the original protocol, 60 histories contain at least one terminal failure; under the amended view, 10 histories contain at least one unresolved recovery. A history can contain answers in several states.

Physical attempts. 30,524 historical physical answer attempts and their charges are retained unchanged.

Missing answers. The original view has 85 missing answers in 60 histories (35/34/16 by B-FLAT/B-GRAPH/unscoped); the amended view has 18 in 10 histories (7/8/3). All 18 lie in the attribution-conflict slot at dose 8 and in the single frozen six-question stratum; by source theme they are care visit 9, client design 3, family travel 3, volunteer shift 2 and housing inspection 1. These counts support no missing-at-random assumption. Pending semantic reviews are existing responses, not missing answers. Further answer recovery is on hold.

Missingness supplement. The worst-case supplement (conditionally authorized after the recovery hold; App. B.3, item 10) varies, for each view, only the genuinely missing binary substitution outcomes over every compatible completion, keeps the one shared unscoped outcome per history, and reapplies the frozen bound, eight gates and six-member Holm family through integer sufficient statistics (at most eight completions per history; a bounded search for the original 85). Pending reviews, disputes and contradictions are not treated as missing. It is reported alongside, never instead of, the frozen unavailable result.

Why H3a/b cannot be rejected. With n complete histories the empirical-Bernstein bound subtracts at least the range term $7r\ln(2/\alpha)/(3(n-1))$; with $r = 2$, $\alpha = 0.05$ and $n = 9{,}954$ this is about 0.0017, larger than the complete-case mean substitution difference of about $7/9{,}954 \approx 0.0007$. The test therefore cannot reject even before the missing outcomes are bounded, and the worst-case envelope p is 1 in both views.

D Extended Related-Work Matrix

Table 10 summarizes what each line of work establishes and how this paper differs.

Authorization before retrieval. Authorization-First Retrieval (Namboothiri, 2026) argues, for multi-agent retrieval-augmented generation over documents, that authorization must precede any learned component and reports that retrieve-then-filter pipelines expose unauthorized content structurally. Filtered vector search is a known systems trade-off in which attribute filters break the recall assumptions of approximate indexes (Chronis et al., 2025). We adopt both lessons and do not claim

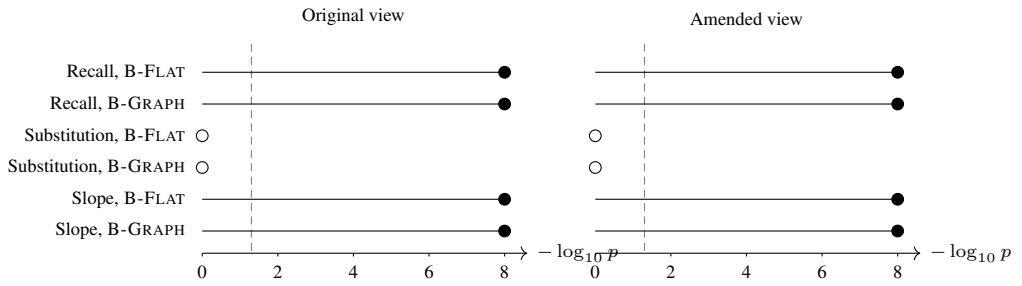


Figure 5: Holm family per view: $-\log_{10}$ of each member's p-value (upper envelope for substitution); filled markers are Holm rejections, and the dashed line marks $p = 0.05$.

them; our setting adds persistent writes whose audiences are derived from the conversation transport, transformations whose labels must be computed, overlapping audiences, declassification, and a recall study against a policy-equivalent principal-postings baseline.

Governed and multi-user agent memory. Collaborative Memory (Rezazadeh et al., 2025) shares memory among users and agents through time-varying access graphs, provenance-tagged fragments and read/write policies that project filtered views. Yang et al. (2026b) formalize multi-user agents as a multi-principal decision problem and observe privacy violations that grow over multi-turn interaction. Governed shared-memory systems include a multi-tenant fleet-memory service whose authors report, among other failure modes, a sub-tenant scope bypass on a get-by-identifier path (Margalit et al., 2026), and a local-first memory system with scope isolation, role-based access and verified erasure (Bhardwaj et al., 2026); Chen et al. (2026b) treat deployment-time memorization as a privacy–utility frontier and observe residue of deleted raw data in derived tiers. Relative to these, our audiences are derived from the transport identities of the participants present rather than configured, derivation and direct lookup are mediated by the same closed rules, and the endpoint is the exact assembled context of each physical model attempt.

Memory privacy benchmarks. CIMemories (Mireshghallah et al., 2026) measures whether models control information flow from persistent memory by task context and finds violations that accumulate across tasks; PiSAs (Gupta et al., 2026) measures cross-user spillage through outputs, inter-agent messages and memory; GateMem (Ren et al., 2026) scores utility, access control and forgetting jointly for multi-principal shared memory; GroupMemBench (Yang et al., 2026a) and EverMemBench (Hu et al., 2026) evaluate multi-party and long-horizon memory utility. Most of these score model output or task utility; we score the assembled context directly and use independently authored corpora under a disclosed audience overlay, reporting GateMem's missing audience labels as a construct limitation (§5.1; Table 12).

Authority, integrity and provenance laundering. A separate line of work concerns what memory authorizes an agent to *do*, or whether memory can be trusted. Consolidation can erase the source constraints of a retained claim (Zhan et al., 2026); memory writers can invent authority that executors act on (Cerruti et al., 2026); consolidation can rewrite external observations as user history (Xu et al., 2026); and poisoning defenses based on content or lineage can be unsound under laundering, motivating write-time, origin-bound, non-malleable authority (Louck, 2026). MAP-Graph (Wang et al., 2026) separates hard permission eligibility from graded path trust and risk-gated actions; MemSecBench (Chen et al., 2026c) measures the persistence and consequences of memory poisoning; MutMem (Saidi, 2026) makes memory mutations cryptographically authorized. We share the write-time origin-binding principle but address a different property (read-audience confidentiality from source to context) and claim no action-authority, integrity or poisoning-defense result.

Information flow and declassification. Our labels descend from lattice and decentralized information-flow control (Bell & La Padula, 1976; Denning, 1976; Myers & Liskov, 1997), our grants from the declassification literature (Sabelfeld & Sands, 2009), and our derived-data rules from provenance-based access control and derived-data policy (Bates et al., 2013; den Hartog & Zannone, 2016). Agent-level information-flow systems such as FIDES (Costa et al., 2025) and

CaMeL (Debenedetti et al., 2025), and the reference-monitor framing of FORGE (Palumbo et al., 2026), target control and effect flows under prompt injection. Contextual integrity (Nissenbaum, 2009) motivates the setting; we implement only its recipient/audience slice, not information-type or transmission-principle norms.

Table 10: Related work: what each line establishes and how this paper differs.

Work	What it establishes	Distinction
AFR (Namboothiri, 2026)	Authorization before learned retrieval for multi-agent RAG	Persistent writes, derived labels, overlapping audiences, grants, recall study
Collaborative Memory (Rezazadeh et al., 2025)	Access graphs, provenance-tagged fragments, filtered views	Participation-derived audiences; closed derivation; exact-context endpoint
GateMem (Ren et al., 2026)	Joint utility, access-control and forgetting benchmark	No checkpoint has authoritative audience labels; reported as a construct limitation
CIMemories (Mireshghallah et al., 2026)	Contextual-integrity violations accumulate across tasks	Deterministic exclusion before the model; recall measured
PiSAs (Gupta et al., 2026)	Cross-user spillage across surfaces	Structural memory-to-context exclusion
AuthMem-Bench (Zhan et al., 2026)	Authority collapse at consolidation	Read confidentiality, not authority to act
TMA-NM (Louck, 2026)	Origin-bound non-malleable authority against poisoning	Shares write-time origin binding; different property
EAL-Bench (Cerruti et al., 2026), PPMF (Xu et al., 2026)	Memory-created or laundered action authority	No action-policy claim
MAP-Graph (Wang et al., 2026)	Eligibility + path trust + action gating	Exact audiences, lifecycle, recall study
MemSecBench (Chen et al., 2026c), MutMem (Saidi, 2026)	Poisoning persistence; authorized mutation	Integrity is out of scope
Governed memory (Margalit et al., 2026; Bhardwaj et al., 2026)	Multi-scope governed memory services	Transport-derived audiences; derivation; mediation
FIDES, CaMeL, FORGE (Costa et al., 2025; Debenedetti et al., 2025; Palumbo et al., 2026)	Agent information flow and reference monitors for actions	Narrow read-audience property at the memory boundary
IFC lineage (Bell & La Padula, 1976; Denning, 1976; Myers & Liskov, 1997; Sabelfeld & Sands, 2009)	Labels, lattices, declassification	Agent-memory lifecycle specialization only
Provenance policy (Bates et al., 2013; den Hartog & Zannone, 2016)	Provenance-based and derived-data access control	Specialized to relationship audiences and exact context
Zanzibar (Pang et al., 2019)	Relationship-based authorization	Userset containment only
Filtered ANN (Chronis et al., 2025)	Filters break index recall assumptions	Measured, not claimed

E Implementation and Evidence Details

This appendix adds implementation and evidence detail only; definitions, the threat model and non-goals are in §3.

E.1 Stored fields and provenance

When one artifact is stored as several physical copies (an origin copy and grant copies), the copies count as one item in every arm. Every recorded write persists canonical participant identifiers, the origin audience $\mathrm{Aud}(m)$ as a typed label, the transport and container identifiers, the identity-resolution confidence and evidence class, source identifiers, the write timestamp and policy version, and a stable fact-bearing object identity to which later grants link. Provenance is *code-bound*: a trusted route adapter mints each authoritative source identifier from an immutable input record, and derived source sets are constructed from the records a transformation actually consumed. Models may generate or transform content, but never generate, select, parse or rewrite an authoritative source, audience, grant or derivation identifier; output with missing or additional provenance is suppressed or quarantined. Shared audiences use one canonical digest over a typed, sorted member projection; a mismatch in kind, digest, member count or members is a security failure, not a lookup miss.

E.2 Revocation and narrowing

Three changes are simple under Eq. (1), though we do not evaluate them. Revoking a grant deletes one grant record, so later attempts no longer see it in $\mathrm{Grants}(m, t)$. Deleting or revoking an item makes it, and every item derived from it, non-current. Removing a member deletes that participant from every stored shared audience and grant that contains it; this can only shrink the set of admitting

containers, and no item needs to be deleted or merged, because every audience still contains o. Each change takes effect from the next attempt's snapshot.

E.3 NATIVE REALIZATION MECHANISMS

Three mechanisms required native-specific engineering: (i) *projection collapse*: origin and grant projections of one artifact share an *authorization subject identity*, consume at most one final result slot, and are refilled from the already-authorized candidate set with a bounded, deterministic schedule that returns fewer than k results rather than broaden the query; (ii) *identity-only entity reuse*: graph entity nodes may be reused across audiences for identity resolution but never carry factual authority, and only an explicit grant over an existing fact-bearing artifact creates a destination projection; and (iii) *attempt admission*: final source and authorization reads for each physical attempt use one private read-only snapshot, and response attribution after transport uses only immutable continuation checks, never a second currentness read.

E.4 ADDITIONAL EVIDENCE SOURCES

Independently authored external evidence. Retrieval preparation covers all 2,549 questions and 15,294 cells, and all 7,647 planned answer identities (1,500 LongMemEval, 5,943 LoCoMo, 204 EverMemBench across the three answer arms) have a recorded contract-valid or explicitly reused output. That is completion of generation, not a finding of accuracy: semantic grading has not been done. The stored retrieval scores of all 15,294 external cells show every arm retrieving identically: no arm, including unscoped retrieval, placed a forbidden item in context, and entitled Recall@5 is equal across all six arms (0.917 on 470 recall-eligible LongMemEval, 0.468 on 1,531 LoCoMo and 0.090 on 68 EverMemBench questions). The overlay leaves these corpora without competing cross-audience evidence, so they show recall parity on independently authored data rather than exclusion. The outputs come from three recorded generation models (Table 11); within each corpus the arms have similar model mixes, but 140 questions have different models across their arms, so their paired comparisons are model-confounded, and the corpora differ sharply in model mix. Accounting holds 8,015 records: 5,190 subscription physical attempts and 2,825 command-line reservations whose underlying send count is unknown; they are not all verified physical requests. Twelve historical answers were adopted without their earlier reviews. For eight valid responses recorded outside the run journal, the earliest valid completion differs from the recorded choice for seven identities, and the full response differs for five; both selections are retained and the scientific choice between them is unresolved.

Table 11: External answer identities by corpus and recorded generation model (three arms per question; generated from the pinned import). Mixed-model q.: questions whose three arms were answered by more than one model.

Corpus	Questions	`gpt-5.6-luna`	`gpt-6-luna`	`claude-opus-5-5`	Mixed-model q.
LongMemEval	500	0	86	1,414	81
LoCoMo	1,981	1,807	3,238	898	46
EverMemBench	68	0	15	189	13
Total	2,549	1,807	3,339	2,501	140

The dependence units are corpus-specific: the EverMemBench questions span five shared topics, not 68 independent histories. Each source is used under a deterministic, disclosed audience overlay that keeps gold evidence authorized for the reader; original evidence labels remain the recall ground truth. Thirty LongMemEval unanswerable controls and four LoCoMo questions are answer-eligible but have no positive-recall denominator (inapplicable, not zero). The full LongMemEval frame forms one connected session component, so it is reported descriptively. All 2,218 GateMem mapping checkpoints were joined; none carries authoritative audience and viewer labels, so no strict-audience cell is eligible. GroupMemBench has no locally validated eligible frame and an unresolved redistribution permission. Current coverage is in Table 12.

Native frame. B-NATIVE is evaluated on an outcome-blind frame of all representable controlled scenarios: 42 of 43 eligible scenarios were selected (one excluded before outcomes because a public

anonymous viewer has no exact native viewer state), each with five answers. The frame supports description of the evaluated implementation only.

Conformance suites. An executable reference oracle, independent of every query implementation, derives expected admission and reason codes. Randomized small universes of participants, items, audiences, viewers, transformations and grants are compared against every write, derivation, grant, retrieval, direct-lookup and context-assembly path of B-FLAT and B-GRAPH, and against each inventoried native path of B-NATIVE. Mutation operators (App. B.2) inject defects such as reversing the subset relation, treating unknown viewers as owner-private, unioning derived audiences or separately approved grants, admitting incomplete-provenance summaries, letting a grant authorize a new derivation, and authorizing after ranking.

E.5 ANSWER JUDGING AND PRE-CONFIRMATION REVIEW

Metadata. The judging routes and their amendment history are in App. B.3 (item 7). Session identifiers and exposed settings are retained privately; for the coding-assistant and `gpt-6-sol` routes, served model version, per-judgment token usage and provider response identifiers are unavailable and recorded as unknown; for Claude judgments the transport verified the served model from the response stream and retained usage. Recorded timestamps measure artifact creation, not model deliberation. Each judgment stores the rationale, exact evidence quotations and uncertainty, and passes the frozen review schema and citation validation; that validation binds what was judged but does not certify that the judgment is correct.

Qualification. For the first session, eighteen excluded, stipulated cases (negation, supersession, attribution, invariance, ambiguity and admission failure) all matched their expected labels; that session had seen the key, so this was non-blinded consistency qualification, and one case was a stipulated abstention rather than a newly generated response. Each later `gpt-6-sol` context was qualified individually on the same unchanged controls; fresh-session Claude judges share one procedure-level qualification reused unchanged across windows. No judge context saw keys, prior labels or other assignments; a failed qualification is not retried until it agrees.

Scope and completion. The review queue has 15,229 unique subjects: 15,193 original responses shared by both views, reviewed once, plus 36 recovered responses that need their own review. All 15,229 subjects carry a validated, durably saved judgment and none is pending: 170 parked subjects were admitted from existing saved judgments (86 held exact-projection records previously validated; 84 saved responses validated by the unchanged validator at admission), and 107 others (malformed, exposure-invalid, uncertain or non-verbatim-citation cases, and one held batch) received a single fresh judgment with no further holds. The judgments mix routes: an initial coding-assistant session, parallel `gpt-6-sol` sessions and fresh-session `claude-opus-5-5` judges, with the 84 admitted saved responses split between the latter two; each route is a recorded provenance stratum, and every record is marked as not independently validated. The last two judging windows were operated and checked by the executing AI coding session under study approval, neither independently nor by a human.

Audit. The approved audit re-assesses up to 500 operationally negative answers per target arm, drawn with a pre-recorded seed from frozen pools. Given evidence of rubric-verification noise and weak judge sensitivity (Peng et al., 2026; Bagaria et al., 2026; Chen et al., 2026a), judging and audit are AI consistency evidence, not independent validation or human ground truth. Pool membership and the seed are frozen before any draw; sampled identities and denominators are preserved, with no adaptive replacement. Any resulting bound is conditional on the AI label procedure and does not certify semantic accuracy. No audit sample has been drawn.

Pre-confirmation review. Before confirmation, review used study-approved AI-assisted judgments plus ten direct judgments by a study author presented without arm or system labels; the ten comparisons agree descriptively (10/10), but repeat concealment was not established and no independent inter-rater reliability was measured.

E.6 NATIVE SYSTEM VALIDATION (FULL)

All 42 selected native scenarios completed with five answers each (210 answers). Under executing-session AI review, 195 of 210 answers were marked correct, with 15 incorrect abstentions and 10 correct abstentions; retrieved gold credit was 180 of 200 eligible answer-level gold items. These counts are descriptive: the repeated answers are not 210 independent cases, the labels are AI reviews rather than human judgments, and the cases span earlier revisions of the native code rather than one homogeneous repaired candidate. Four unfavorable cases are retained: approved owner-private widening (five incorrect abstentions; context from another conversation), complete intersection (five incorrect abstentions), overlapping audiences (five incorrect abstentions; the requested group's evidence was ranked second), and summary consolidation (correct answers without full consolidated gold recall). Two coverage qualifications apply: the selected native program did not execute an independent direct-identifier probe for existence opacity, and the incomplete-provenance cases recorded non-emitting requests rather than runtime refusals.

Scoped native qualification. A separately accepted qualification of the selected native configuration passed all 39 phases (three controls, a collection phase and 35 selected modules; 377 unique selected tests, 1,048.6 seconds, no model calls). The acceptance checked 316 declared trace-file hashes, the exact collected and executed test union, boundary records and cleanup. This supports the selected configuration and checks only: not arbitrary native functionality, the reduced public artifact, or new semantic utility. Three historical native evidence references named by the integration record remain unauthenticated, and current success cannot authenticate them retrospectively.

E.7 SUPPORTING STUDIES AND NEGATIVE RESULTS (FULL)

Table 12: Required evidence tracks (exact joins). Reproduced is not scientifically accepted, an output is not a graded answer, and a missing receipt is not a zero.

Track	Expected frame	Retained coverage
Identity policies (Study A)	28 cells + 4 controls	32 executed; tables reproduced
Matched availability (H4b)	16 cells	16 executed; tables reproduced
Systems, original (Study B)	600 samples	600 retained; all 120 retrievals denied
Systems, companion	840 timed calls	840 executed; contended host
External retrieval	2,549 questions	2,549 prepared (15,294 cells)
External answers	7,647 identities	7,647 outputs; not graded
GateMem mapping	2,218 checkpoints	0 eligible (no audience labels)
Installation census	14 metrics	None; 7 controlled workloads instead
Native qualification	39 phases	39 passed (377 tests); scoped
Native evidence references	3 artifacts	Unauthenticated

Status. The identity-policy, matched-availability and systems studies below were executed locally without model calls. An independent review reproduced all four retained tables (Tables 13–15) from their unchanged inputs and accepted their packaging; that is technical reproduction, not independent scientific acceptance, so we report them as controlled, descriptive evidence within the scopes stated below, and no headline claim depends on them.

Identity evidence (Study A). Seven fault families (conflicting, extra, forged, missing, stale, unknown and unresolved identity evidence) were crossed with four declared policies, plus an exact one-to-one control through each (Table 13). Before reconciliation, fail-open exposed forbidden items in 11 of 13 opportunities; the three fail-closed alternatives exposed none but lost entitled evidence at different rates. After attested reconciliation, every origin and derived decision matched its known entitlement, all 16 superseded copies were denied, and all four controls kept owner and authorized access while denying other participants. This is a reference-policy experiment, not a change to B-NATIVE or the reference stores; "provisional" is one declared conservative operationalization (owner-only escrow until reconciliation).

Derivation availability (H4b). Across eight matched cases (four derivation families on both architectures), the intersection item was available in all eight and ranked first in four; blanket suppression made none available (Table 13). These are case counts, not population rates.

Table 13: Supporting studies (tables reproduced; not independently accepted). Top: origin-level probe opportunities across seven identity-fault families before reconciliation (controls excluded); denominators count controlled probes, not participants or histories. Bottom: matched derivation cases over both reference architectures; availability is direct lookup of the eligible derived item, reported separately from top-1 retrieval inclusion. Generated from pinned exports.

Identity policy (before reconciliation)	Forbidden exposed	Entitled lost
Fail open	11/13	1/43
Public-only fallback	0/13	15/43
Quarantine (delayed binding)	0/13	27/43
Provisional, later reconciled	0/13	11/43
Derivation treatment (8 cases)	**Derived item available**	**Ranked first**
Intersection	8/8	4/8
Blanket suppression	0/8	0/8

Systems realization (Study B). Matched measurements cover two architectures, two representations (audience labels and principal postings), five operations and 30 repetitions (600 samples) on warm-cache SQLite. All 120 of their retrieval samples returned empty results (denial), so they do not describe successful retrieval latency. The companion studies add 840 timed owner calls (120 positive retrieval and 720 lifecycle and materialization; 30 per cell over 28 cells) after 140 excluded warm-ups (Table 14). Timings use warm SQLite and operating-system caches on one laptop-class machine (10 cores, solid-state storage, SQLite 3.51.3), with unrelated background activity of about half a core recorded, so they are descriptive, not uncontended or production latency, and none concerns native latency. The timer covers the durable owner operation including its authorization index; the source-missing operation measures owner-visible state, not physical purge, and reopen with an idempotent write is not crash recovery. We also count objects, audiences and grants in seven controlled workload prestates, each realized in four physical configurations (Table 15); this is not an installation census, which needs a production deployment that does not exist.

Table 14: Companion systems timings, nearest-rank median / 95th percentile in milliseconds over 30 repetitions per cell (warm cache; contended host; generated from the pinned table).

	B-FLAT		B-GRAPH	
Operation (ms, p50 / p95)	Audience labels	Postings	Audience labels	Postings
Positive retrieval	575 / 597	585 / 604	815 / 855	812 / 837
Derived materialization	12.0 / 12.5	11.6 / 12.1	12.1 / 12.4	11.9 / 12.8
Grant materialization	10.6 / 11.3	10.5 / 11.3	10.6 / 11.3	10.2 / 10.6
Grant removal	0.56 / 0.87	0.68 / 1.48	0.50 / 0.60	0.46 / 0.59
Artifact retirement	0.71 / 0.98	0.65 / 0.84	0.55 / 0.70	0.45 / 0.55
Source-missing state	0.60 / 0.83	0.53 / 0.88	0.52 / 0.69	0.47 / 0.61
Reopen + idempotent replay	11.6 / 12.3	11.7 / 12.6	11.8 / 12.2	11.4 / 12.0

Table 15: Controlled workload prestates used by the systems study (logical counts; not deployment data).

Controlled workload	Objects	Sources	Audiences	Grants
Positive retrieval	134	134	2	0
Derived materialization	66	66	3	0
Grant materialization	66	66	3	0
Artifact retirement	67	66	3	0
Source-missing state	67	66	3	0
Grant removal	66	68	3	1
Reopen + idempotent replay	67	66	3	0

Held-out pilot. The corrected historical pilot report covers draws 8–39: 32 paired histories, 1,920 retrieval cells and 479 valid answer records from 480 logical answers and 483 physical attempts, with one retained terminal failure. Retained AI labels mark 160/160 B-FLAT, 157/159 B-GRAPH and 120/160 unscoped answers correct (unscoped: 25 wrong non-abstaining answers and 15 abstentions),

with no wrong-principal labels. These are descriptive labels, not verified correctness; the primary unit is the 32 paired histories, not 479 observations, and zero wrong-principal labels neither show the planned reduction nor prove a population zero. The report binds its inputs but leaves the exact executed producer revision of the older pilot reports unresolved.

Retrieval budget. Every confirmation cell stores its 20 ranked candidates, so Recall@k follows for $k = 1, \ldots, 20$ at the frozen fetch budget without new runs (recomputed $k = 5$ equals the stored score in all 360,000 dose-8 cells). At every k the architectures equal the policy-equivalent post-filter and postings exactly and are never below unscoped or silo retrieval; they reach full recall at $k = 10$, unscoped retrieval only at $k = 18$, and the silo plateaus at 0.750 (Figure 6).

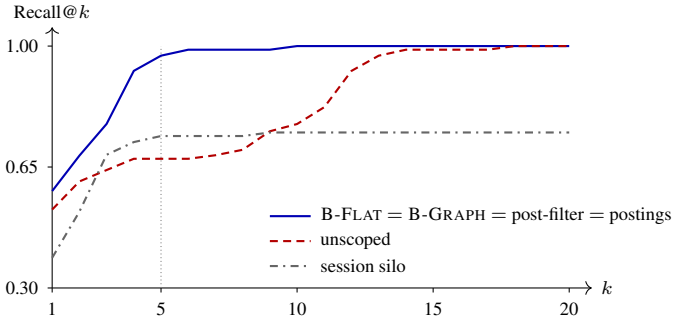


Figure 6: Mean entitled Recall@k over 10,000 histories (six questions, $d = 8$, fetch budget 20); the dotted line marks the frozen $k = 5$.

Negative, null and weak results. We retain these results as reported: the frozen H3a/b test could not detect a substitution reduction at the observed base rate, so the joint decision is not established (§6.3); the held-out pilot of 32 histories did not demonstrate a wrong-principal substitution reduction, so the planned effect sizes remained planning targets; four unfavorable native cases (three with incorrect abstentions, one with incomplete gold recall; §6.4); 85 original terminal citation-contract failures, 18 of which remain unresolved under recovery; and, in a 75-answer supporting slice (39 controlled and 36 external answers; descriptive AI review), 47 correct answers, 27 valid misses and one terminal failure, where twelve LoCoMo misses concern questions that lack upstream evidence annotations.