Beyond Retrieval: Selective Reconsolidation in Persistent Agent Memory

Anonymous Author(s)

***Abstract*—Persistent agent memory prevents loss of prior work, yet it can preserve conclusions after their evidence has been invalidated. Retrieving a correction does not determine which conclusions remain supported. We study *selective reconsolidation*, a memory operation that withdraws unsupported consequences while retaining conclusions with independent justification. The mechanism records AND/OR justification structure, re-evaluates it after review, and compiles bounded working context from the revised state. We prove that influence-only reachability permits exact revision for every audit if and only if each claim has a unique minimal justification. In SciMapRepair, selective reconsolidation grounds 139/140 delayed answers, compared with 117/140 when the same memory records a correction without propagating it. In 80 paired worlds with identical text, timestamps, embeddings, and influence edges, justification-aware repair is exact while no propagation under-revises and influence propagation over-revises. After matched rendering, this separation replicates across three models from two providers: justification-aware state yields 75–80/80 grounded decisions, compared with 34–40/80 for either baseline. Controlled diagnosis episodes and heterogeneous support structures further connect the mechanism to delayed action and test it beyond the paired templates. The results identify recoverable justification structure as a requirement for selectively revising persistent beliefs.**

***Index Terms*—agent memory, memory reconsolidation, belief revision, cognitive architectures, provenance, long-horizon agents**

I. Introduction

Persistence prevents forgetting while allowing contamination to endure. A research agent may retain an observation, its interpretation, a mechanistic conclusion, and the experiment planned from that conclusion. If the observation is later invalidated, storing the correction does not automatically stop the agent from retrieving and citing the derived conclusion.

Most agent-memory work asks what should be stored and retrieved. Long context retains history but remains sensitive to position and observation budget [1]. Retrieval-augmented generation (RAG), graph retrieval, and hierarchical maps improve access under bounded context [2]–[5]. PEEK stores an orientation cache for recurring situations [6]. These systems address an access question:

$$\text{stored state} \longrightarrow \text{relevant past information}. \tag{1}$$

Evidence change poses a different question:

$$\text{new evidence} \longrightarrow \text{revised active belief state}. \tag{2}$$

We call the second operation *selective reconsolidation*. Retrieval relevance and revision dependence are different relations. Similarity can retrieve a correction; it cannot by itself establish which consequences lose support.

Revision must also preserve unaffected beliefs. Propagating every correction through every connected node removes stale consequences, but it can also erase a claim that retains an independent justification. The memory must distinguish jointly required evidence from alternative support. A graph that records only influence edges does not record that distinction.

We model revision as a third cognitive-memory function alongside retention and retrieval. Minimal justification families specify the evidence combinations that license each claim. This representation yields a boundary theorem: influence-only reachability permits exact revision only when every claim has a unique minimal justification. An information corollary shows that systems supporting broader audits must store, encode, or reconstruct the missing distinctions.

We implement justification-aware selective reconsolidation in a persistent research agent. Four controlled studies test under-revision when corrections remain local, over-revision when influence graphs conflate conjunction with alternative support, the diagnosis–repair–action sequence, and heterogeneous scientific support structures. Together, the formal and empirical results connect a representational requirement to revised state and subsequent model behavior.

II. Memory Beyond Retrieval

A. Three functions of external memory

External representations reduce working-memory demand and change how a cognitive task is performed [7], [8]. Language-agent architectures now externalize episodic records, summaries, plans, and learned skills [9]–[11]. This motivates separating three memory functions.

Retention keeps information available across sessions. *Retrieval* selects past information relevant to the current task. *Revision* changes the active consequences of stored information after its epistemic status changes. A representation can perform the first two functions well and still fail the third. Persistence then gives an early error a longer lifetime.

Reconsolidation provides a functional analogy. Reactivated biological memories can become available for updating [12]. The analogy concerns function rather than neural mechanism. In our system, a review event reactivates the affected justification region, the controller revises validity while retaining immutable history, and a compiler constructs the next bounded context from the revised state.

B. Relation to agent-memory repair

Recent benchmarks extend memory evaluation beyond recall to temporal updates, invalidated records, self-evolution,

poisoning, and compositional consequences [13]–[18]. TrustMem verifies proposed consolidation transitions for coverage, preservation, and faithfulness [19]. MemoRepair withdraws descendants before validated successors are republished under complete influence provenance [20]. Kumiho gives versioned graph memory belief-revision semantics [21]. Work on exact deletion likewise separates addressable contributions, which admit subtraction, from entangled writes, which require replay [22].

These systems establish memory updating and descendant repair as explicit operations. The remaining representational question is whether influence edges contain enough information to preserve independently supported conclusions. Assumption-based truth-maintenance systems (ATMS) and provenance semirings distinguish conjunction from alternative derivations [23], [24]. We apply that distinction to persistent agent memory, connect repair to delayed behavior, and hold rendering fixed to avoid the presentation confound identified by Evidence-State Revision [25].

III. Beliefs Need Justifications

Let $E = \{e_1, \ldots, e_m\}$ be primitive evidence and let c be an initially justified derived claim. Its minimal justification family is a nonempty antichain

$$\mathcal{S}(c) = \{S_1, \ldots, S_k\}, \qquad S_i \subseteq E, \tag{3}$$

of nonempty subsets of E. Every item inside a support is required, and any one intact support is sufficient. If an audit invalidates $R \subseteq E$, then

$$c \text{ remains active} \iff \exists S \in \mathcal{S}(c) : S \cap R = \varnothing, \tag{4}$$

and

$$c \text{ is retired} \iff \forall S \in \mathcal{S}(c) : S \cap R \neq \varnothing. \tag{5}$$

We model evidential support rather than truth. A claim may remain true after its evidence fails, but the agent should no longer treat the broken derivation as support. For the analysis, $\mathcal{S}(c)$ denotes minimal supports after recursively expanding derived dependencies to primitive evidence. The implementation evaluates the equivalent structure incrementally over derived nodes.

Equation (4) gives the selective-repair rule. It has three immediate properties. *Propagation*: a claim with no intact support is withdrawn. *Preservation*: an independently supported claim remains active. *Attribution*: each active claim has at least one intact support certificate.

A. When influence-only reachability is insufficient

An influence-only graph records the influence set

$$U(c) = \bigcup_{S \in \mathcal{S}(c)} S. \tag{6}$$

A reachability cascade retires c whenever $R \cap U(c) \neq \varnothing$. Correct repair instead uses the universal condition in (5).

Theorem 1 (Justification-structure boundary)**.** *Influence-only reachability is correct for every audit $R \subseteq E$ if and only if c has a unique minimal justification.*

Proof. If $\mathcal{S}(c) = \{S\}$, then $U(c) = S$ and both rules retire the claim exactly when $R \cap S \neq \varnothing$. If two distinct minimal supports S_1, S_2 exist, antichain minimality gives an item $e \in S_1 \setminus S_2$. For $R = \{e\}$, graph reachability retires the claim while S_2 remains intact. □

Figure 1 is the smallest witness. The observable text and influence graph are identical. Only the grouping of edges into justifications differs, yet the correct revision reverses.

B. Information required for revision

Let $\mathcal{R}$ be the audits a memory promises to answer and let

$$\Sigma_{\mathcal{R}}(\theta) = \big(D_\theta(R)\big)_{R \in \mathcal{R}} \tag{7}$$

be the revision signature, where $D_\theta(R)$ is the correct retirement set in world θ.

Corollary 1 (Revision sufficiency)**.** *If the possible worlds induce M distinct revision signatures, every fixed-length exact memory code needs at least $\lceil \log_2 M \rceil$ bits that distinguish them.*

The proof is pigeonhole counting: two worlds that require different repairs cannot share a codeword. The system may retain the distinction as explicit provenance, encode it another way, or reconstruct it from logs and tools after the audit. It cannot answer exactly if the distinction is absent from both stored state and repair-time observations.

A simple aggregation example shows how large the gap can be. Let independent $U_i \sim \mathrm{Bernoulli}(1/2)$ be summarized only by $T = \sum_i U_i$. The total needs $\lceil \log_2(m+1) \rceil$ bits. If a later request names record i and requires the exact deleted total $T - U_i$, auxiliary state Z must satisfy

$$H(Z \mid T) \geq H(U_{1:m} \mid T) = m - H(T). \tag{8}$$

Thus a summary sufficient for the current total can require almost m additional bits to become record-retraction ready. Appendix A gives the proof and the audit-order result.

IV. Justification-Aware Selective Reconsolidation

Figure 2 shows the implemented cognitive loop. A background reviewer compares a claim with executed work and identifies a discrepancy. The controller treats the failed evidence as a reactivation root, evaluates the support families of its dependents, writes invalidated revisions only for claims with no intact support, and propagates from those newly unsupported claims. The next context excludes invalidated latest revisions and renders each active derived claim with its valid supports.

A. Reconsolidation as a cognitive control loop

The loop separates metacognitive diagnosis from memory-state revision. A discrepancy first signals that the active account no longer agrees with an executed calculation, retrieved source, or other observable result. The reviewer then performs source

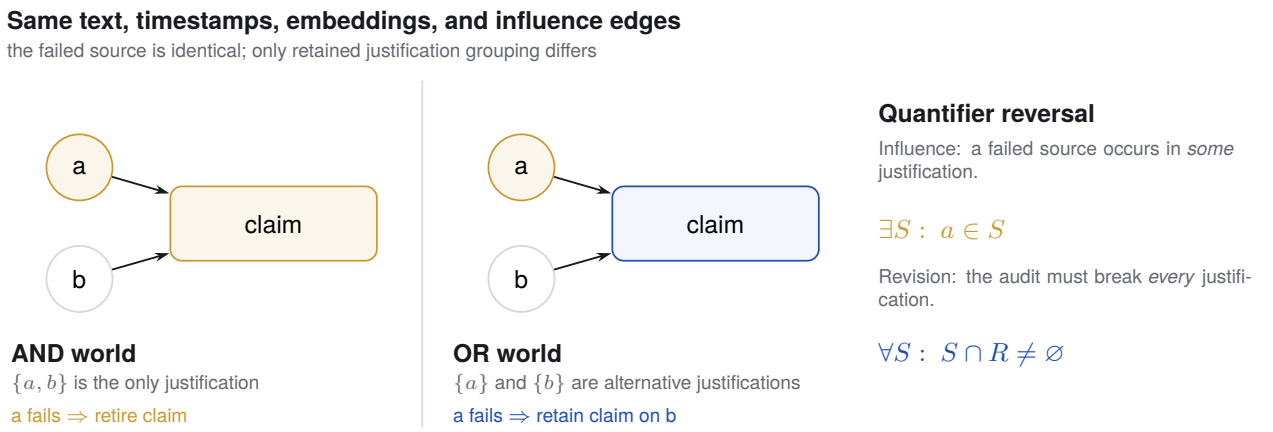

Fig. 1. Identical influence topology can require opposite revisions after source a is invalidated. The influence graph records that a and b affected the claim, but omits whether they are jointly required or independently sufficient.

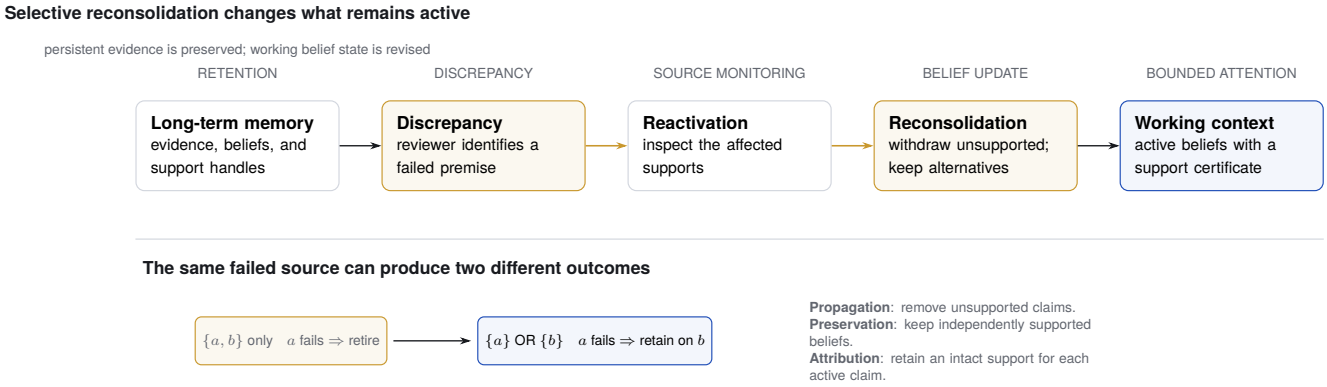

Fig. 2. Selective reconsolidation is a state-changing memory operation between review and bounded working context. Immutable history remains available, but unsupported beliefs no longer enter the active context.

monitoring by comparing the claim with the work that produced it and nominating the premise whose status changed. This diagnosis is fallible and is evaluated separately from repair.

Once a root is nominated, *reactivation* bounds the update. The controller traverses only the downstream justification region instead of rebuilding the entire memory. Within that region, support validation requires each retained claim to have at least one intact evidential route. *Reconsolidation* commits new active-state revisions for claims that have lost every route while preserving their immutable pre-audit versions. Finally, *working-memory compilation* renders the revised latest state under a fixed context budget. The stages solve different computational problems:

$$\begin{aligned} &\text{discrepancy} \rightarrow \text{diagnosis} \rightarrow \text{reactivation} \\ &\rightarrow \text{source monitoring} \rightarrow \text{reconsolidation} \\ &\quad \rightarrow \text{working-memory compilation}. \end{aligned} \tag{9}$$

Repair accuracy therefore depends on both root diagnosis and state transition. A correct repair rule cannot compensate for a misdiagnosed root, and a correct diagnosis leaves behavior unchanged when the correction is appended without revising active state.

B. Persistent representation

Mantis Forge stores immutable resource revisions, version-specific handles, derivation dependencies, active status, and audit history. A derived variable may additionally carry

$$\mathcal{S}(c) = \{S_1, \ldots, S_k\}, \tag{10}$$

represented as alternative lists of conjunctive handles. The repository rejects empty or redundant support families and includes every support handle in dependency traversal. The evaluated implementation uses a transparent depth-two AND/OR representation; shared circuits could compress repeated sub-derivations.

The repairer starts with prior invalidations and the current review finding. A dependent variable is retired only when every recorded support contains a broken handle. Variables without explicit alternatives retain unique-justification behavior. User-owned evidence and frozen resources remain unchanged;

controller-owned derived state receives an immutable invalidated revision. This separation prevents the repair operation from rewriting source evidence.

C. Maximal preservation

For failed evidence R, let $\mathcal{C}$ be the set of claims under consideration and define

$$A_R = \{c \in \mathcal{C} : \exists S \in \mathcal{S}(c),\ S \cap R = \varnothing\}. \tag{11}$$

Call an active set *support-sound* after R when each retained claim has an intact minimal support. Then A_R is the unique largest support-sound active set, and $\mathcal{C} \setminus A_R$ is the smallest retirement set whose complement is support-sound. This follows directly from the membership condition defining A_R. Under fixed support semantics, repair is idempotent, monotone in R, and independent of the order in which audits are accumulated.

Incomplete provenance requires a third outcome. A bounded representation may certify retention, certify retirement, or abstain. The compressed representation evaluated below abstains when its recorded semantics do not support a binary decision.

D. Predictions at the cognitive level

The architecture yields four observable predictions. First, appending a correction without changing active descendants should produce *under-revision*: a model may choose the updated answer while still grounding it in a withdrawn intermediate claim. Second, an influence cascade should remove that contamination in unique-justification chains but produce *over-revision* when a claim has independent support. These two errors should remain visible after the revised state is compiled into matched bounded context; otherwise presentation could account for the apparent memory effect.

Third, evidence redundancy should invert the behavior of influence-only reachability. As the number k of independent supports increases, true fragility falls because more routes must be broken. Influence reachability instead becomes more likely to encounter at least one failed source and retire the claim. Fourth, diagnosis and state transition should contribute separately to end-to-end performance. A reviewer that nominates the wrong root and a repairer that mishandles the right root are cognitively distinct failures, even when both produce the same final mistake. Studies 1–4 test these predictions in that order.

V. Evaluation

The evaluation asks four questions. Study 1 tests whether changing active memory matters beyond storing and retrieving a correction. Study 2 isolates justification semantics from graph connectivity. Study 3 measures the diagnosis–repair–behavior loop without an oracle root. Study 4 tests the mechanism on heterogeneous scientific support structures. All exact-repair endpoints are computed from program state; no model judges a reported result.

A. Study 1: storing versus reconsolidating

SciMapRepair contains 20 controlled scientific workspaces. Each has a failed measurement, three downstream derived nodes, an independent sibling branch, a later audit and replacement result, and 28 intervening administrative sessions. Seven delayed questions per workspace probe direct invalidation, one- and two-hop consequences, action repair, replacement evidence, calibrated insufficiency, and preservation. The 140 questions use fixed choices and exact evidence identifiers.

Six conditions use GPT-5.6 Sol at low reasoning, one generation per cell, and a 3,200-character evidence budget: task-blind long window; MiniLM vector RAG; temporal RAG; graph RAG with derivation metadata but no validity operation; the active map with direct invalidation only; and selective reconsolidation. The two map conditions use identical ordering and rendering. The primary endpoint requires the correct choice, an admissible gold citation, and no contaminated citation.

The benchmark conditions on correct failed-root identification so that it isolates repair after discrepancy detection. Twenty repository checks verify that the sanitizer withdraws all three descendants and preserves the independent branch before questions are issued. Inference uses workspace means. Exact paired sign tests discard ties, Holm correction covers five contrasts, and bootstrap intervals resample workspaces.

B. Study 2: paired justification worlds

RevisionPair holds the visible memory fixed while changing only justification semantics. In the one-source regime, each pair has edges $a \to h$, $b \to h$, and $h \to u$. One world has the single support $\{a, b\}$; the other has alternatives $\{a\}$ and $\{b\}$. Source a fails. Claim h should retire in the first world and remain active in the second.

The two-source regime tests a higher-order cutset. One world has supports $\{a, c\}$ and $\{b, c\}$, so $\{a, b\}$ breaks every justification. The matched world has three singleton alternatives, so source c keeps claim h active. Across 20 scientific domains, every pair has identical source text, claim text, timestamps, audit, embedding inputs, and influence edges. Forty SHA-256 checks confirm pre-repair identity; the correct active sets differ in all 40 pairs.

The core comparison is no repair, influence reachability, and justification-aware reconsolidation. We also diagnose singleton criticality, reconstruction from retained derivation logs, and a compressed one-source representation that abstains beyond its contract. Eighty worlds produce 480 deterministic repair decisions. The justification-aware condition invokes the active repository and reviewer sanitizer; 80/80 product-state checks match the formal repair set.

We then test whether those state differences reach agent behavior. Each core mechanism's repaired state is compiled into the same four fields: status, claim identifier, claim text, and current valid support. The rendering contains no mechanism name, justification topology, or raw graph. GPT-5.6 Sol receives that bounded snapshot and an identical delayed action question, with one low-reasoning generation for each mechanism–world cell. The 240 generations are scored by exact choice and

citation to the rendered claim-state record. Domain means are the inferential units.

After the primary result, worlds, mechanism states, prompts, and gold answers were frozen; we then repeated all 240 cells with GPT-5 Mini and Claude Sonnet 4.6. This post hoc robustness check uses the same scoring and one generation per cell. SHA-256 checks confirm that all 240 rendered task prompts are byte-identical across model runs.

For each deterministic mechanism we also record the serialized structure retained before audit and any derivation log read during repair. These byte counts characterize the benchmark representations only. For the active implementation, we time the sanitizer call and count immutable invalidated revisions. Repair requires no model call once a root is supplied.

C. Study 3: end-to-end diagnosis and repair

The first two studies condition on the failed root to isolate revision. Study 3 removes that condition in 20 diagnosis episodes derived from the SciMapRepair scenarios. The reviewer receives four rotated candidate evidence records, the stored derivation, the new audit, and replacement evidence. It must identify the affected premise by exact identifier. Its nomination is then passed to the same persistent repository and sanitizer used in Study 2. A later action question is answered from the revised compiled snapshot.

We report three endpoints separately: root-detection accuracy, repair accuracy conditional on a correct root, and end-to-end grounded action accuracy. The decomposition prevents a successful deterministic repair from hiding diagnosis errors. A changed repository state, rather than the reviewer's explanation, is the evidence of repair. The study uses 40 GPT-5.6 Sol low-reasoning calls: one diagnosis and one delayed decision per episode.

D. Study 4: heterogeneous revision cases

The Heterogeneous Revision set contains 20 manually authored cases spanning biology, machine learning, and quantitative science. Unlike RevisionPair, the cases do not share one text template. Their structures include independent replication, jointly required evidence, shared preprocessing bottlenecks, partial rescue, code-version and configuration failures, invalid statistical assumptions, superseded measurements, independent cohorts and instruments, negative controls, and mixed alternative/conjunctive support.

Each case specifies primitive evidence, minimal support families, an audit set, and a delayed action. Cases, support families, audits, and gold delayed actions were frozen in source before any compared mechanism was executed or any model response was generated. The same three core mechanisms produce a revised state, which is rendered through the same bounded schema before one model decision. We report exact state repair and grounded behavior descriptively. The set is a deliberately heterogeneous stress test; population inference over real research histories is outside its scope.

TABLE I
SCIMAPREPAIR OVER 140 DELAYED QUESTIONS (PERCENT).

Condition	Grounded	Invalid	Repair	Preserve
Long window	34.3	7.9	60.0	0.0
Vector RAG	37.9	47.9	2.5	25.0
Temporal RAG	12.9	45.0	5.0	5.0
Graph RAG	43.6	42.9	17.5	50.0
Map–direct only	83.6	15.7	45.0	100.0
Selective reconsolidation	**99.3**	**0.0**	**100.0**	**100.0**

VI. RESULTS

A. Reconsolidation prevents under-revision

Table I reports SciMapRepair. Selective reconsolidation grounds 139/140 answers (99.3%), compared with 117/140 (83.6%) for the identical map without dependent propagation. Both select the correct option on 139/140 questions. Support validity accounts for the difference: direct-only invalidation cites withdrawn descendants 22 times and repairs 18/40 propagated consequences; reconsolidation emits no contaminated citation and repairs 40/40. Both preserve all 20 independent branches.

Reconsolidation improves 16 workspace means, ties four, and loses none against direct-only invalidation. The exact two-sided sign test is $p = 3.05 \times 10^{-5}$, unchanged by Holm adjustment. Workspace-bootstrap 95% intervals are 97.9–100.0% for reconsolidation and 78.6–88.6% without propagation. Each retrieval contrast favors reconsolidation in all 20 workspaces (Holm-adjusted $p = 9.54 \times 10^{-6}$). These intervals describe variation within the fixed suite; population inference over scientific failures is outside the design.

B. Justification semantics prevent over-revision

Figure 3 reports RevisionPair. No repair is exact in the 40 worlds where the challenged source is dispensable and leaves 40 unsupported derived nodes active in conjunctive worlds. Influence reachability is exact in the corresponding 40 conjunctive worlds and collaterally retires 40 nodes with independent support. Justification-aware reconsolidation is exact in all 80 worlds. The same pattern appears downstream: justification-aware rendering yields 80/80 grounded model decisions, while no propagation and influence reachability each yield 40/80. There are no invalid responses or missing claim citations. Justification-aware behavior exceeds each baseline in all 20 domain means (two-sided exact sign $p = 1.91 \times 10^{-6}$ for each comparison).

The behavioral separation replicates across models (Table II). Claude Sonnet 4.6 reproduces the primary 40/80, 40/80, and 80/80 pattern. GPT-5 Mini is noisier but retains the ordering: 37/80, 34/80, and 75/80. Justification-aware state exceeds each baseline in 19/20 GPT-5 Mini domain means, with one tie and no loss (exact sign $p = 3.81 \times 10^{-6}$), and in all 20 Claude domain means ($p = 1.91 \times 10^{-6}$). GPT-5 Mini produces 3, 8, and 5 invalid responses in the three conditions; invalid responses count as failures. The other two models produce none.

TABLE II
GROUNDED REVISIONPAIR DECISIONS ACROSS MODELS.

Model	No propagation	Influence	Justification-aware
GPT-5.6 Sol	40/80	40/80	**80/80**
GPT-5 Mini	37/80	34/80	**75/80**
Claude Sonnet 4.6	40/80	40/80	**80/80**

Singleton criticality solves all 40 one-source worlds but only 20/40 two-source worlds, where it unsafely retains 40 derived nodes. The compressed one-source representation is exact within its contract and abstains in all 40 two-source worlds. Reconstruction is also exact, replacing persistent justification state with a mean of 31 or 39 serialized log bytes read during repair.

The redundancy check evaluates $k \in \{1, 2, 4, 8\}$ disjoint supports, depths $d \in \{1, 2, 4\}$, and independent $p = .05$ failures with 100,000 seeded trials in each of 12 cells. For $d = 2$, increasing k from one to eight reduces true retirement from 9.75% to 8.17×10^{-7}%, while influence-reachability retirement rises from 9.75% to 55.99%. Exact enumeration of all 189 nonempty antichain support families over up to four evidence items verifies the reachability boundary.

C. Diagnosis connects review to later action

In Study 3, the reviewer identifies the affected premise in 20/20 episodes. The observed conditional-repair rate and delayed grounded-action rate are also 20/20. These descriptive rates characterize the controlled 20-episode set; open-world diagnostic reliability remains unmeasured.

The median repair latency after diagnosis is 51.3 ms. The nominated root is applied to the repository, unsupported descendants receive immutable invalidated revisions, and the later decision reads the resulting active snapshot. The experiment therefore measures a durable state change and its effect on subsequent behavior.

D. Heterogeneous cases reproduce both errors

Study 4 reproduces both failure directions without the paired templates. No propagation repairs 9/20 cases exactly and under-revises the other 11. Influence reachability repairs 11/20 and over-revises the other nine. Justification-aware reconsolidation repairs all 20/20. Bounded compilation yields the same grounded-decision counts: 9/20, 11/20, and 20/20, respectively, with no invalid responses. All 20 active product-state checks agree with the declared support semantics.

The cases require different visible outcomes under one support rule. A code-version defect retires outputs while retaining raw observations; a shared calibration failure defeats apparently separate analyses; an orthogonal assay can retain a conclusion after one conjunctive route breaks. These outcomes follow from the recorded justifications without a domain-specific repair policy.

E. Support structure, audit order, and cost

We next enumerate the scaling surface rather than sampling failures. Let $k \in \{1, 2, 4, 8\}$ be the number of disjoint alternative supports, $d \in \{1, 2, 4\}$ the number of jointly required items per support, and $q \in \{1, 2, 3, 4\}$ the number of audited items. Exact fixed-cardinality enumeration yields 41 valid (k, d, q) cells. Figure 4 shows the $d = 2$ slice; the supplement reports all cells.

The enumerated surface exposes the audit-order hierarchy. When $k > 1$, retiring the claim requires a cutset that intersects every support. Singleton criticality can represent the $q = 1$ boundary but fails when a higher-order combination is decisive. Influence reachability exhibits the opposite error: increasing independent support makes the true claim more robust while making graph deletion more likely. Justification-aware repair has zero error in all 41 cells.

Table III shows the corresponding representation–repair tradeoff in RevisionPair. Exactness requires either persistent structure or repair-time reconstruction. The measured sanitizer makes no model call: over 80 repairs its median latency is 31.2 ms (95th percentile 55.8 ms) and it writes a mean 2.5 invalidated revisions.

VII. DISCUSSION

A. A revision operation for persistent memory

The studies expose symmetric failures. Without propagation, the agent under-revises because invalid consequences remain active. Without justification semantics, it over-revises because valid conclusions lose independent support. Selective reconsolidation withdraws precisely the claims whose recorded supports are broken:

$$\boxed{\begin{array}{c}\text{no propagation} \longrightarrow \text{under-revision},\\ \text{influence only} \longrightarrow \text{over-revision},\\ \text{justification-aware} \longrightarrow \text{selective revision}.\end{array}} \tag{12}$$

A revisable belief record needs both content and a recoverable account of its justification. Retrieval keys describe when the content may be useful, while revision requires an account of why the belief remains active. That account may be a justification list, a circuit, a replayable derivation log, a typed provenance structure, or an explicit decision to abstain. The functional requirement is that the distinction remain recoverable when evidence changes.

The distinction also explains why more connected evidence can hurt an influence-only cascade. Independent corroboration should make a claim more robust. The influence graph sees additional incident edges and increases the chance that some failed source reaches the claim. Justification semantics restores the intended direction of evidence redundancy.

The end-to-end result locates the operation within a cognitive architecture. The reviewer acts as a metacognitive monitor by detecting conflict and attributing it to a stored premise. The repository applies the resulting belief-state transition, and the compiler renders the revised long-term state as the limited evidence available for action. Separating these roles makes a wrong diagnosis, an unsound repair, and stale rendering observable as different errors.

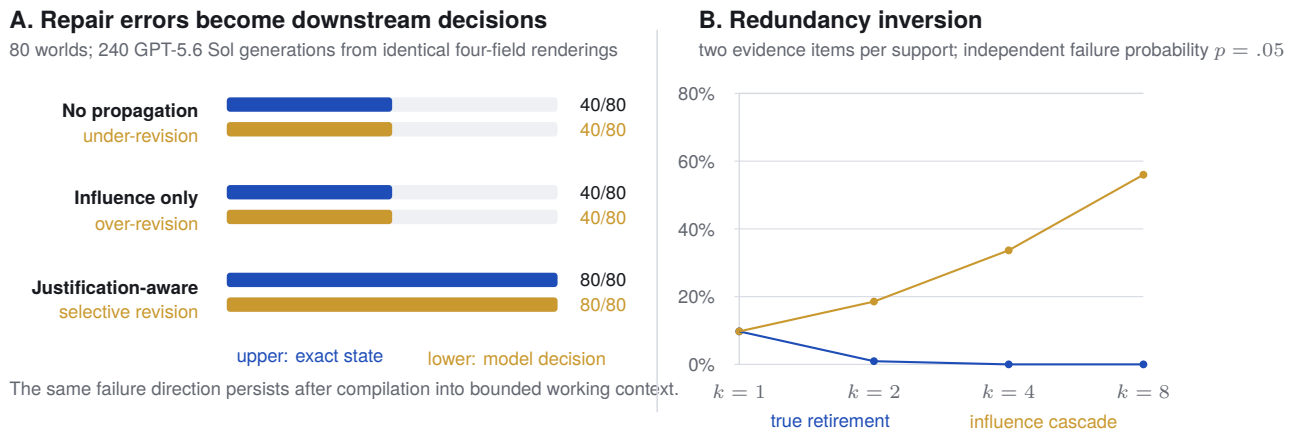

Fig. 3. RevisionPair results under matched rendering. Panel A shows that under- and over-revision persist in downstream model decisions; justification-aware repair is exact at both the state and behavior levels. Panel B shows redundancy inversion: independent corroboration reduces true retirement while increasing retirement under influence reachability.

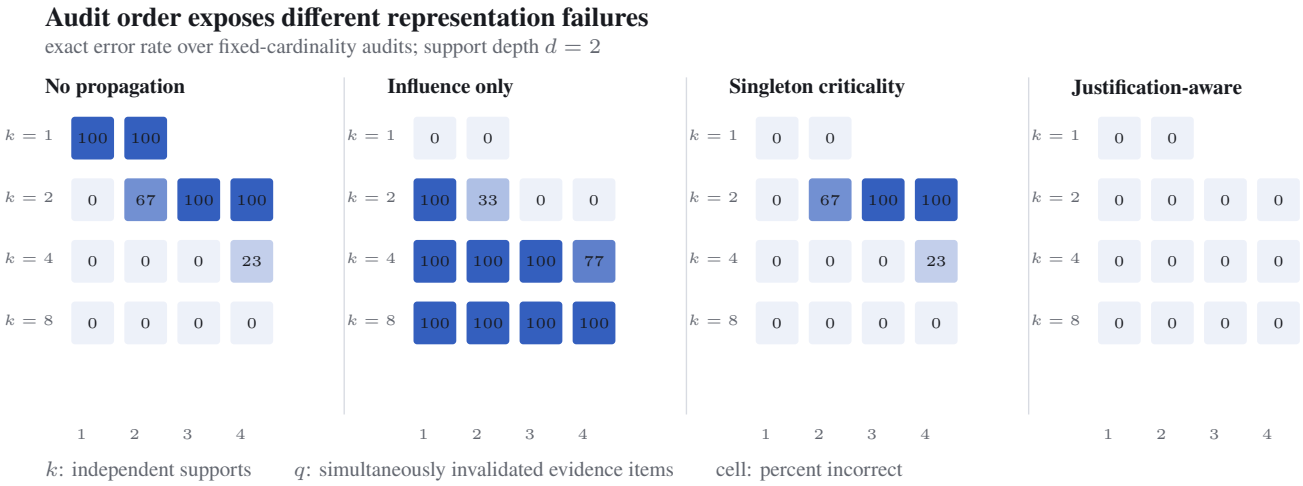

Fig. 4. Exact repair-error surfaces at support depth $d = 2$. No propagation misses audits that hit every justification. Influence reachability deletes a claim after any hit and therefore worsens as independent support grows. Singleton criticality handles first-order failures but misses higher-order cutsets. Justification-aware repair is exact in every enumerated cell. Blank cells have $q > kd$.

Immutable history and revised active state serve different purposes. The former preserves the episode for later inspection, learning, and possible reversal. The latter controls what the agent may currently treat as established. Selective reconsolidation withdraws a mistaken conclusion from active use while preserving the record of how the system reached it.

B. Architectural implications

The architecture separates three memory surfaces. An *episodic record* preserves observations, actions, and prior revisions. An *active belief state* records which claims are currently licensed and why. A *working context* is a budgeted rendering of that active state for the next decision. Collapsing these surfaces creates familiar pathologies: treating the append-only episode as current belief preserves stale claims, while treating the prompt as the memory makes revision disappear when the session ends.

The reviewer belongs at the boundary between evidence and active belief. It can inspect calculations, citations, and generated artifacts to propose a discrepancy root; the proposal remains observable and revisable, and history remains unchanged. Once a root is accepted, support evaluation is a state operation with a deterministic contract. This division uses model intelligence for interpretation while keeping the resulting state transition inspectable. It also permits delayed or asynchronous review: a finding can arrive after the original task and still alter what future sessions receive.

Support structure can be captured while a research agent executes tools. A result records the dataset, code version, assumptions, and intermediate claims that licensed it; alternative experiments create alternative supports. Any representation is

TABLE III
REVISIONPAIR COST AND ACCURACY. BYTES ARE MEAN SERIALIZED BENCHMARK STATE BEFORE AUDIT OR LOGS OBSERVED DURING REPAIR; Q1/Q2 DENOTE ONE- AND TWO-SOURCE AUDITS.

Memory rule	Persistent bytes q1/q2	Repair-read bytes q1/q2	Exact q1	Exact q2	Abstains q2
No propagation	0 / 0	0 / 0	20/40	20/40	0
Influence reachability	48 / 62	0 / 0	20/40	20/40	0
Singleton criticality	5.5 / 3.5	0 / 0	40/40	20/40	0
Justification-aware	12 / 20	0 / 0	40/40	40/40	0
Reconstruct on demand	0 / 0	31 / 39	40/40	40/40	0
q=1 compressed	5.5 / 3.5	0 / 0	40/40	0/40	40

adequate if it recovers these distinctions at revision time. When the recorder cannot establish whether two routes are independent, abstention avoids both silent retention and destructive cascade. Table III makes the design choice explicit: retain persistent structure, reconstruct from later observations, or restrict the audits that admit exact answers.

C. Matched controls

Each study addresses a different alternative explanation. SciMapRepair holds rendering and ordering fixed between direct invalidation and reconsolidation, which separates the 22 contaminated citations from retrieval presentation. RevisionPair holds influence-graph topology and all visible text fixed while reversing the correct repair, so influence connectivity alone cannot encode the decision. Its downstream layer renders the same four fields for every mechanism, showing that deterministic state errors persist through a common prompt interface. Study 3 requires the reviewer to name the premise and applies that name to the repository before the later question; only a repository state change counts as repair. The heterogeneous cases extend the test beyond the two paired templates.

These controls establish a mechanism effect under the evaluated suites; ecological prevalence remains unmeasured. The observed advantage follows from changing active support state before bounded context is compiled. Retrieval window and mechanism labels are held fixed in the matched comparisons.

D. Scope and limitations

All four studies are controlled. SciMapRepair and Studies 3–4 use one model and one sample per cell. RevisionPair behavior is replicated across three models from two providers, with one sample per model–cell. RevisionPair repeats two logical templates across scientific domains, and its behavioral layer tests whether a state error persists after context rendering. It does not test model reconstruction of hidden justification topology. The heterogeneous set broadens the logical forms but remains manually authored, descriptive, and separate from human-adjudicated natural research histories.

Study 3 tests diagnosis among four candidate records with explicit identifiers under audits written to distinguish the affected source. Open-world scientific fault localization remains outside the experiment. The 20/20 result connects diagnosis to actual repair and downstream behavior within that setting; it does not establish a general diagnosis rate. Cross-model RevisionPair replication provides a broader behavioral check while leaving model and task coverage limited. Reconstruction reads structured derivation logs rather than asking a model to infer missing provenance. The results establish mechanism behavior and a representation boundary within controlled suites. They do not estimate ecological prevalence or general scientific competence.

The current semantics are monotone. Defeaters, uncertain causal links, inference-rule changes, cyclic dependencies, and probabilistic evidence require richer models. Flat minimal-support sets can also grow exponentially; shared AND/OR circuits or bounded cutsets are natural extensions. Root diagnosis and audit selection are separate problems not addressed here. Appendix A quantifies how the state required for revision grows with record-level retraction and audit order.

VIII. CONCLUSION

Long-lived agents must revise active beliefs when corrections arrive. Retrieval can surface a discrepancy, while selective revision depends on which derivations the discrepancy breaks. Untyped influence reachability is sufficient in the unique-justification case. Once beliefs have independent support, correct reconsolidation requires justification structure or an equivalent reconstruction process.

Persistent memory should retain enough derivational information to revise beliefs selectively when their evidence changes.

REPRODUCIBILITY AND AI DISCLOSURE

The anonymous supplement provides the study definitions, 1,660 model outputs, deterministic mechanism outcomes, product repair checks, statistical analyses, theory enumerations, figure sources, and a SHA-256 manifest for the reported studies. It includes the 20 SciMapRepair workspaces, 80 RevisionPair worlds, 20 diagnosis episodes, and 20 heterogeneous cases. The experiments use no human participants or private research data.

AI Disclosure: OpenAI GPT-5.6 Sol was used through Mantis Forge for literature retrieval, implementation assistance, manuscript drafting, SciMapRepair generations, RevisionPair downstream decisions, Study 3 diagnosis and downstream decisions, and Study 4 downstream decisions. OpenAI GPT-5 Mini and Anthropic Claude Sonnet 4.6 were used for the RevisionPair cross-model replication. The authors checked the

proofs, inspected generated artifacts, and verified references and calculations; they take responsibility for the theory, experimental design, and paper. Primary scoring and statistical analysis are programmatic, and no model judge determines a reported endpoint.

APPENDIX A
RETRACTION AND AUDIT-ORDER BOUNDS

A. Retraction gap

Let $U_{1:m}$ take values in an additive group and let $T = \sum_i U_i$. Suppose Z is auxiliary state from which a decoder can return $T_{-i} = T - U_i$ for every named record i. Then $U_i = T - T_{-i}$, so the full vector $U_{1:m}$ is recoverable from (T, Z). Therefore

$$H(Z \mid T) \geq I(U_{1:m}; Z \mid T) = H(U_{1:m} \mid T). \tag{13}$$

For independent fair Bernoulli variables, $H(U_{1:m}) = m$ and T is a deterministic function of the vector, giving (8). The binomial entropy satisfies

$$H(T) = \tfrac{1}{2} \log_2(\pi e m/2) + o(1). \tag{14}$$

B. Audit-order hierarchy

Let $\mathcal{R}_q = \{R \subseteq E : |R| \leq q\}$ and let $\mathrm{RC}_q(m, n)$ be the worst-case number of exact revision bits for n initially justified monotone claims. With $r_q = \min(q, \lfloor m/2 \rfloor)$,

$$n\binom{m}{r_q} \leq \mathrm{RC}_q(m, n) \leq n \sum_{r=1}^{q} \binom{m}{r}. \tag{15}$$

For fixed q, this is $(1 + o(1))n\binom{m}{q}$. The upper bound stores one answer bit for every claim–audit pair. For the lower bound, choose an arbitrary family of r_q-subsets as minimal invalidating cutsets. Every such family is an antichain, and distinct families disagree on an allowed audit. Under unrestricted audits there are $M_m - 1$ initially justified monotone signatures per claim, where M_m is the Dedekind number [26].

REFERENCES

[1] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, 2024.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," 2020.

[3] P. Sarthi, S. Abdullah, A. Tuli, S. Khanna, A. Goldie, and C. D. Manning, "RAPTOR: Recursive abstractive processing for tree-organized retrieval," 2024.

[4] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, "From local to global: A graph RAG approach to query-focused summarization," 2024.

[5] B. J. Gutiérrez, Y. Shu, Y. Gu, M. Yasunaga, and Y. Su, "HippoRAG: Neurobiologically inspired long-term memory for large language models," 2024.

[6] Z. Gu, Q. Zhang, O. Khattab, and S. Madden, "PEEK: Context map as an orientation cache for long-context LLM agents," 2026.

[7] J. Zhang and D. A. Norman, "Representations in distributed cognitive tasks," *Cognitive Science*, 1994.

[8] E. F. Risko and S. J. Gilbert, "Cognitive offloading," *Trends in Cognitive Sciences*, 2016.

[9] T. R. Sumers, S. Yao, K. Narasimhan, and T. L. Griffiths, "Cognitive architectures for language agents," 2023.

[10] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez, "MemGPT: Towards LLMs as operating systems," 2023.

[11] W. Xu, Z. Liang, K. Mei, H. Gao, J. Tan, and Y. Zhang, "A-MEM: Agentic memory for LLM agents," 2025.

[12] J. L. C. Lee, K. Nader, and D. Schiller, "An update on memory reconsolidation updating," *Trends in Cognitive Sciences*, vol. 21, no. 7, pp. 531–545, 2017.

[13] D. Wu, H. Wang, W. Yu, Y. Zhang, K.-W. Chang, and D. Yu, "LongMemEval: Benchmarking chat assistants on long-term interactive memory," 2024.

[14] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang, "Evaluating very long-term conversational memory of LLM agents," *ACL 2024*, 2024.

[15] M. N. Uddin, K. Shubham, E. Blanco, C. Baral, and G. Wang, "From recall to forgetting: Benchmarking long-term memory for personalized agents," 2026, accepted to ACL 2026 Findings.

[16] Y. Wang, Z. Zhang, M. Chi, K. Yu, Y. Li, M. Peng, B. Tong, C. Zhang, Y. Zhou, and J. Li, "EvoMemBench: Benchmarking agent memory from a self-evolving perspective," 2026.

[17] M. S. Arya, "RECON: Benchmarking agent memory for compositional reasoning over long contexts," 2026.

[18] X. Chen, X. Xie, W. Fu, J. Zhou, S. Yu, and Q. Xuan, "MemSecBench: Tracking agent memory poisoning from persistence to consequence and repair," 2026.

[19] T. Yang, S. Paul, V. Srinivasan, V. Kulkarni, and S. Chappidi, "TrustMem: Learning trustworthy memory consolidation for LLM agents with long-term memory," 2026.

[20] Y. Zhao, C. Dai, M. Kou, and Y. Xiu, "MEMOREPAIR: Barrier-first cascade repair in agentic memory," 2026.

[21] Y. B. Park, "Graph-native cognitive memory for AI agents: Formal belief revision semantics for versioned memory architectures," 2026.

[22] V. Ramesh, "Subtract or replay? exact deletion from language-model memory," 2026.

[23] J. de Kleer, "An assumption-based TMS," *Artificial Intelligence*, vol. 28, no. 2, pp. 127–162, 1986.

[24] T. J. Green, G. Karvounarakis, and V. Tannen, "Provenance semirings," in *Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 2007, pp. 31–40.

[25] Z. Jiang, Z. Fu, Z. Li, Y. Kim, J. Mi, X. Peng, F. Teng, and H. Wu, "Presentation, not mechanism: A render confound in deprecation-aware memory evaluation," 2026.

[26] A. D. Korshunov, "Solution of Dedekind's problem on the number of monotonic boolean functions," *Doklady Akademii Nauk SSSR*, vol. 233, no. 4, pp. 543–546, 1977.